Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE
Corso di Laurea in Matematica

MODELLI REGRESSIVI E METODI DI CLASSIFICAZIONE

Tesi di Laurea in Probabilità e Statistica Matematica

Relatore:
Chiar.mo Prof.
Andrea Pascucci

Presentata da:
Alessia Sitta

Sessione Unica
Anno Accademico 2017/2018

Dedico questo lavoro alla mia famiglia...

Introduzione

L'obiettivo di questo elaborato è quello di studiare i modelli regressivi, lineari e non, al fine di applicare la teoria della regressione al problema della classificazione degli individui.
Nel primo capitolo principalmente vengono introdotti i concetti base della regressione lineare e multipla.
Il secondo capitolo è totalmente incentrato sull'Analisi delle Componenti Principali (PCA), uno dei metodi più famosi e utilizzati per risolvere il problema della riduzione di dimensione dei modelli regressivi.
Nell'ultimo capitolo, successivamente a una breve introduzione del problema della classificazione, viene mostrato come esso può essere risolto sfruttando tutti i concetti e gli strumenti analizzati nel primo capitolo, uniti a un nuovo tipo di regressione: la regressione logistica.
Mentre il primo capitolo è puramente teorico, il secondo e il terzo sono completati da esempi pratici che, a mio avviso, permettono una comprensione più completa dell'elaborato.

Indice

Capitolo 1

Correlazione e regressione

1.1 Covarianza e coefficiente di correlazione

Definizione 1.1. *Siano X e Y variabili aleatorie. Si chiama* ***covarianza tra X e Y*** *il numero reale*

$$Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$$

dove μ_X e μ_Y sono rispettivamente le medie di X e Y.

Questa definizione ha senso se μ_X e μ_Y sono finite e il valor medio complessivo è finito, cosa che avviene se ad esempio X, Y $\in L^2(\Omega, P)$, ovvero $E[X^2] < \infty$ e $E[Y^2] < \infty$.
Si noti che la definizione è analoga a quella di varianza $Var[X] = E[(X-\mu_X)^2]$, risulta infatti:

$$Var[X] = Cov(X,X)$$

$$Cov(X,Y) = E[XY] - \mu_X\mu_Y$$

Definizione 1.2. *Chiamiamo* ***coefficiente di correlazione lineare tra X e Y*** *il numero definito da*

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X\sigma_Y}$$

con σ_X e σ_Y deviazioni standard di X e Y.
Si scrive anche ρ_{XY} al posto di $\rho(X,Y)$.

Proposizione 1.1. *Vale $-1 \leq \rho(X,Y) \leq 1$.*

Dimostrazione. Si utilizza la disuguaglianza di Cauchy-Schwarz:

$$|E[XY]| \leq \|X\|_2\|Y\|_2.$$

$|Cov(X,Y)| \leq \sqrt{E[(X-\mu_X)^2]E[(Y-\mu_Y)^2]} = \sigma_X\sigma_Y$.
Allora

$$|\rho_{XY}| = \frac{|Cov(X,Y)|}{\sigma_X\sigma_Y} \leq \frac{\sigma_X\sigma_Y}{\sigma_X\sigma_Y} = 1.$$

□

Teorema 1.1. *Se X e Y sono variabili aleatorie indipendenti, allora si ha che $Cov(X,Y) = 0$ e $\rho(X,Y) = 0$.*

Dimostrazione. Se X e Y sono indipendenti allora $E[XY] = E[X]E[Y]$. Infatti:

$$E[X]E[Y] = \int_{\mathbb{R}} x\mu_X dx \int_{\mathbb{R}} y\mu_Y dy = \int_{\mathbb{R}^2} xy(\mu_X \otimes \mu_Y)d(x,y) =$$
$$= \int_{\mathbb{R}^2} xy\mu_{(X,Y)}d(x,y) = E[XY].$$

Se poniamo $\hat{X} = X - \mu_X$ e $\hat{Y} = Y - \mu_Y$, questi sono indipendenti perchè X e Y lo sono. Allora $Cov(X,Y) = E[X-\mu_X]E[Y-\mu_Y] = (E[X]-\mu_X)(E[Y]-\mu_Y) = 0$. Dalla definizione di coefficiente di correlazione segue immediatamente che se X e Y sono indipendenti allora $\rho(X,Y) = 0$. □

Il viceversa non è vero: $Cov(X,Y) = 0$ non implica necessariamente che X e Y sono indipendenti (tranne nel caso di (X,Y) gaussiano: questo fatto contribuisce alla tendenza pratica di ritenere la condizione $Cov(X,Y) = 0$ un notevole sintomo di indipendenza).

Definizione 1.3. *Due variabili aleatorie X e Y si dicono **scorrelate** se hanno correlazione nulla, $\rho(X,Y) = 0$, o equivalentemente se $Cov(X,Y) = 0$.*

Quindi indipendenza ⇒ scorrelazione.

A livello numerico su dati sperimentali, se la correlazione è molto vicino a zero, questo è un buon indicatore di indipendenza, o più precisamente di scorrelazione.
Analizziamo meglio cosa si intende con correlazione dei dati sperimentali.
Supponiamo di avere a disposizione n coppie di valori sperimentali (x_i, y_i) riassunti nella seguente tabella nella quale le righe corrispondono agli individui e le colonne alle variabili:

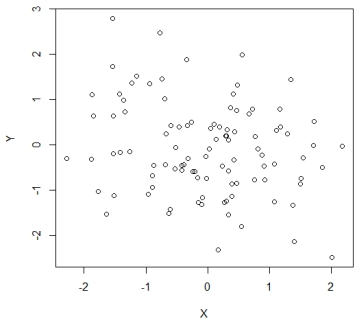
	X	Y
1	x_1	y_1
...	...	...
n	x_n	y_n

Di questi dati sperimentali possiamo calcolare la covarianza empirica e il coeff. di correlazione empirica definiti da:

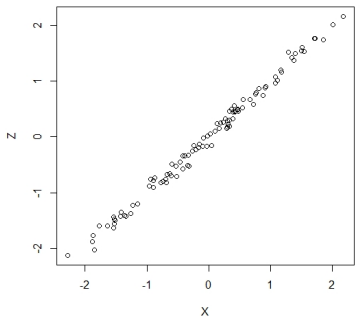
$$\hat{Cov} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

$$\hat{\rho} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

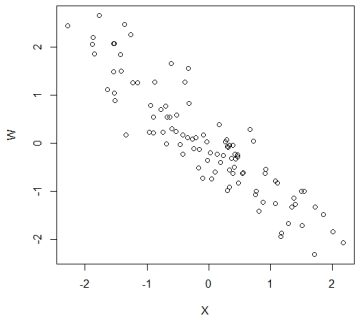
dove $\bar{x}$ e $\bar{y}$ sono rispettivamente le medie empiriche di X e Y, ovvero $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ e $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$.
La vicinanza a zero di $\hat{\rho}$ si interpreta come sintomo di indipendenza o comunque bassa dipendeza di X e Y, mentre la vicinanza a 1 come elevato legame positivo, a -1 come elevato legame negativo.
Chiariamo meglio questo concetto attraverso l'uso del software R.
Partiamo generando due campioni X e Y con cardinalità 100, indipendenti e con distribuzione gaussiana, mostriamo le coppie (x_i, y_i) nel piano cartesiano e calcoliamo la correlazione empirica.
X e Y sono indipendenti quindi $\rho(X,Y) = Cov(X,Y) = 0$; questa è appunto una situazione a correlazione nulla:



Se invece costruiamo un campione Z simile a X ma leggermente perturbato, ovvero $Z = X + \frac{1}{10}Y$, la situazione che si presenta è a elevatissima correlazione positiva. Infatti:
$Var[Z] = Var[X + \frac{1}{10}Y] = Var[X] + \frac{1}{100}Var[Y] + 2Cov(X, \frac{1}{10}Y) = 1 + \frac{1}{100}$.
$Cov(X,Z) = Cov(X, X + \frac{1}{10}Y) = Cov(X,X) + \frac{1}{10}Cov(X,Y) = 1$.
Allora $\rho(X,Z) = \frac{1}{\sqrt{1+\frac{1}{100}}} = 0.99496280$.



Se ancora costruiamo W simile a -X e lo perturbiamo, ovvero $W = -X + \frac{1}{2}Y$, otteniamo una situazione a moderata/elevata correlazione negativa. Infatti:
$Var[W] = Var[-X + \frac{1}{2}Y] = Var[X] + \frac{1}{4}Var[Y] + 2Cov(X, \frac{1}{2}Y) = 1 + \frac{1}{4}$.
$Cov(X, W) = Cov(X, -X + \frac{1}{2}Y) = -Cov(X, X) + \frac{1}{2}Cov(X, Y) = -1$.
Allora $\rho(X, W) = \frac{-1}{\sqrt{1+\frac{1}{4}}} = -0.8987381$.



Come si vede da questi esempi, il segno di $\hat{Cov}$ corrisponde all'inclinazione positiva o negativa della nuvola di punti; questi però sono solo alcuni dei possibili esempi quindi è utile un ragionamento più generale basato sulla formula della covarianza empirica $\hat{Cov} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$.
Il punto $(\bar{x}, \bar{y})$ è collocato pressochè al centro della nuvola di punti, comunque essa sia orientata.
Analizziamo il caso in cui sia $\hat{Cov} > 0$.
Questo significa che nella somma $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ predominano addendi positivi: per semplicità di ragionamento supponiamo siano tutti positivi.
Se $(x_i - \bar{x})(y_i - \bar{y}) > 0$ allora i fattori hanno lo stesso segno: se sono positivi significa che il punto (x_i, y_i) si trova a nord-est di $(\bar{x}, \bar{y})$; se sono negativi significa che il punto

si trova a sud-ovest di $(\bar{x}, \bar{y})$. In ogni caso, i punti si trovano come nella seconda figura presentata sopra.
Per semplicità abbiamo ipotizzato che tutti i termini $(x_i - \bar{x})(y_i - \bar{y})$ fossero positivi, in generale ovviamente non è così: ci saranno anche termini negativi ma comunque predominano quelli positivi, quindi predomina di fatto la struttura grafica del tipo detto sopra (anche se alcuni punti si troveranno a nord-ovest o sud-est di $(\bar{x}, \bar{y})$).
Il ragionamento nel caso $\hat{Cov} < 0$ è analogo.

1.2 Regressione lineare semplice

Ipotizziamo che tre v.a. X, Y e ϵ siano legate dalla seguente funzione lineare

$$Y = aX + b + \epsilon$$

dove a e b sono numeri reali.
Interpretiamo questa struttura pensando che X e Y siano legate da una relazione lineare (graficamente una retta di equazione $y = ax+b$, per cui a si dirà **coefficiente angolare** e b **intercetta**), perturbata però da un errore casuale ϵ. La v.a. X è detta **predittore** e la v.a. Y è detta **output**.
Supporremo sempre che ϵ sia standardizzata, cioè $E[\epsilon] = 0$, e inoltre supporremo X e ϵ scorrelate.

Definizione 1.4. *La relazione $Y = aX + b + \epsilon$ è detta* ***modello di regressione lineare semplice*** *e chiameremo* ***retta di regressione*** *la retta $y = ax + b$.*

Ci poniamo ora due obiettivi:

1. Trovare formule che permettano di calcolare approssimativamente a, b e la deviazione standard σ_ϵ dell'errore ϵ a partire da dati sperimentali, quando si ipotizza il modello lineare ma non si conoscono i coefficienti;

2. Interpretare il concetto di coefficiente di correlazione nell'ambito del modello lineare.

Per risolvere il primo punto, partiamo dal seguente

Teorema 1.2. *Se tre v.a. X, Y e ϵ sono legate dalla relazione lineare $Y = aX+b+\epsilon$, ϵ e X sono scorrelate, $\mu_\epsilon = 0$ e $\sigma_X > 0$ allora i coefficienti a e b sono univocamente determinati:*

$$a = \frac{Cov(X, Y)}{\sigma_X^2}$$
$$b = \mu_Y - a\mu_X.$$

Inoltre

$$\sigma_\epsilon^2 = \sigma_Y^2 - a^2\sigma_X^2.$$

Dimostrazione. Per linearità del valore atteso e per l'ipotesi $E[\epsilon] = 0$ vale

$$E[Y] = aE[X] + b.$$

Vale, inoltre, per le regole sulla varianza ed usando $Cov(X, \epsilon) = 0$:

$$Var[Y] = a^2 Var[X] + Var[\epsilon]$$

$$Cov(X, Y) = Cov(X, aX + b + \epsilon) = aCov(X, X) + Cov(X, \epsilon) = aVar[X].$$

Da queste relazioni si ricavano immediatamente le tre formule. □

Ora supponiamo di avere n dati sperimentali, ovvero n coppie (x_i, y_i) che rappresentano i valori di due grandezze X e Y trovati esaminando n individui.
Possiamo calcolare i numeri

$$\bar{x} = \frac{\sum_i x_i}{n}, \bar{y} = \frac{\sum_i y_i}{n}$$

$$\frac{\sum_i (x_i - \bar{x})^2}{n}, \frac{\sum_i (y_i - \bar{y})^2}{n}$$

$$\frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n}$$

e considerarli come approssimazioni rispettivamente di

$$E[X], E[Y]$$

$$Var[X], Var[Y]$$

$$Cov(X, Y).$$

Tramite queste approssimazioni è possibile stimare a, b e σ_ϵ grazie al teorema 1.2.

1.2.1 Interpretazione del coefficiente di correlazione

Per quanto riguarda l'interpretazione di ρ_{XY}, il teorema 1.2 stabilisce una relazione tra coefficiente di correlazione e coefficiente angolare della retta di regressione:

Corollario 1.1. $\rho_{XY} = \frac{\sigma_X}{\sigma_Y} a$.

Dimostrazione. Dal teorema 1.2 e dalla definizione di coefficiente di correlazione si ha

$$a = \frac{Cov(X, Y)}{\sigma_X^2} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} \frac{\sigma_Y}{\sigma_X} = \rho_{XY} \frac{\sigma_Y}{\sigma_X}.$$

□

Innanzitutto questo chiarisce che a non è il coeff. di correlazione, come invece per una sorta di gioco di parole si può essere portati a pensare: del resto ρ_{XY} può variare solo tra -1 e 1, mentre la pendenza di una retta può benissimo essere maggiore di quella delle bisettrici.
Vale però la regola: $a > 0 \Leftrightarrow \rho_{XY} > 0$ (ed analogamente per valori negativi). Quindi, ρ_{XY} è indice di legame lineare diretto tra le variabili X e Y (ovvero una variabile cresce se l'altra cresce), mentre $\rho_{XY} < 0$ è indice di legame lineare inverso.
Quindi almeno il segno di ρ_{XY} è facilmente interpretabile.
Un'interpretazione più precisa deriva dallo studio dell'errore:
dal teorema 1.2 si ha che

$$\sigma_\epsilon^2 = \sigma_Y^2 - a^2\sigma_X^2.$$

Sostituendo $a = \rho_{XY}\frac{\sigma_Y}{\sigma_X}$ si trova

$$\sigma_\epsilon^2 = \sigma_Y^2(1 - \rho_{XY}^2).$$

Questo risultato fa capire che la varianza dell'errore, cioè la grandezza che misura quanto preciso sia il legame lineare tra X e Y, è tanto maggiore quanto più vicino a zero è ρ_{XY}: i valori vicini a zero di ρ_{XY} implicano un pessimo legame lineare (errore elevato); viceversa, valori di ρ_{XY} vicini a ± 1 implicano un legame lineare stretto.
Quindi ρ_{XY} non è legato tanto all'inclinazione della retta di regressione quanto piuttosto alla precisione con cui essa descrive il legame tra le variabili.

Infine, introduciamo alcuni nuovi termini che si rifanno all'idea che con un modello lineare di fatto stiamo cercando di dare una spiegazione della variabilità della grandezza Y.
Supponiamo di avere una v.a. Y che varia in modo imprevedibile.
Noi vorremmo capire se queste variazioni siano almeno in parte spiegabili tramite un legame lineare con un predittore X: quando osserviamo ad esempio valori di Y più grandi della media, questo non è dovuto semplicemente al caso, ma al fatto che il predittore X ha assunto valori più grandi del solito (se $a > 0$).
Tutto questo, però, è pur sempre corrotto dall'errore, per cui la spiegazione della variabilità di Y offerta dalla retta di regressione non è mai una spiegazione completa.
In quest'ottica, Y ha una sua varianza, una sua variabilità.
L'espressione $aX + b$ riesce a spiegarne una parte.
La parte non spiegata di Y è la differenza tra Y e la parte spiegata, cioè $aX + b$.
Quindi la parte non spiegata di Y è proprio l'errore ϵ.
Con questo nuovo linguaggio:

Definizione 1.5. *La **varianza spiegata** è la percentuale della varianza che è stata spiegata da $aX + b$, mentre la **varianza non spiegata** è la percentuale complementare.*

In termini matematici la varianza non spiegata è

$$\frac{\sigma_\epsilon^2}{\sigma_Y^2}$$

mentre la varianza spiegata è

$$1 - \frac{\sigma_\epsilon^2}{\sigma_Y^2}.$$

Teorema 1.3. *Il coefficiente di correlazione al quadrato corrisponde alla varianza spiegata* $1 - \frac{\sigma_\epsilon^2}{\sigma_Y^2}$.

Dimostrazione. Sempre sfruttando il teorema 1.2 si ha

$$\rho_{XY}^2 = \frac{Cov(X,Y)^2}{\sigma_X^2 \sigma_Y^2} = \frac{a^2 \sigma_X^2}{\sigma_Y^2} = 1 - \frac{\sigma_\epsilon^2}{\sigma_Y^2}$$

□

Più ρ_{XY}^2 è alto (vicino a 1) più la relazione lineare riesce a spiegare la variabilità di Y.

1.2.2 Matrice di covarianza e correlazione empirica

Definizione 1.6. *La **matrice di covarianza** Q (rispettivamente **matrice di correlazione** ρ) di un vettore $X = (X_1, ..., X_n)$ è la matrice di ordine n definita da $Q_{ij} = Cov(X_i, X_j)$ (rispettivamente $\rho_{ij} = \rho(X_i, X_j)$) per $i, j = 1, ..., n$.*

Osservazione 1.1. *Dalle proprietà della covarianza risulta banalmente che Q è simmetrica e definita positiva.*

Supponiamo di avere una tabella di dati del tipo

	X_1	$\cdots$	X_p
1	$x_{1,1}$	$\cdots$	$x_{1,p}$
2	$x_{2,1}$	$\cdots$	$x_{2,p}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$
n	$x_{n,1}$	$\cdots$	$x_{n,p}$

dove le colonne rappresentano diverse variabili (ad es. $X_1 = PIL$, ..., $X_{p-1} =$ spese per istruzione, $X_p =$ spese per sanità), le righe rappresentano diversi individui (ad esempio le nazioni europee) e i valori numerici sono noti (sono stati misurati). Possiamo calcolare la matrice di covarianza empirica $\hat{Q}$ e la matrice di correlazione empirica $\hat{\rho}$ associate alla tabella dei dati:

$$\hat{Q}_{ij} = \frac{\sum_{k=1}^{n} (x_{k,i} - \bar{x}_i)(x_{k,j} - \bar{x}_j)}{n}$$

$$\hat{\rho}_{ij} = \frac{\sum_{k=1}^{n}(x_{k,i} - \bar{x}_i)(x_{k,j} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{k,i} - \bar{x}_i)^2 \sum_{k=1}^{n}(x_{k,j} - \bar{x}_j)^2}}$$

dove $\bar{x}_i = \frac{\sum_{k=1}^{n} x_{k,i}}{n}$.
Il significato applicativo di queste matrici è la conseguenza del significato di covarianza e correlazione tra stringhe di dati; nel caso della correlazione, il numero $\hat{\rho}_{ij}$ misura il legame lineare tra le colonne i e j, con le relative interpretazioni del segno descritte al paragrafo precedente.
E' importante notare che i legami così catturati sono a due a due, non globali tra tutte le variabili nel loro complesso.

1.3 Regressione lineare multipla

Supponiamo di avere sempre una tabella del tipo:

	X_1	$\cdots$	X_p	Y
1	$x_{1,1}$	$\cdots$	$x_{1,p}$	y_1
2	$x_{2,1}$	$\cdots$	$x_{2,p}$	y_2
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
n	$x_{n,1}$	$\cdots$	$x_{n,p}$	y_n

dove le colonne rappresentano diverse variabili (ad esempio reddito, anni di istruzione, spese per mostre) e le righe i diversi individui.
Il problema che ci poniamo è se le variabili Y e $X_1, ..., X_p$ siano legate da una relazione, a meno di errore, della forma

$$Y = f(X_1, ..., X_p, \epsilon).$$

Le grandezze $X_1, ..., X_p$ sono come sempre chiamate **predittori** mentre Y è detto **output**.
E' interessante capire il perchè del termine predittore; supponiamo appunto che Y e $X_1, ..., X_p$ siano legate da una relazione funzionale: calcolata in nuovi valori delle $X_1, ..., X_p$ (non già sperimentati), essa produce dei corrispondenti valori Y, ovvero permette di predire il valore di Y in corrispondenza di input mai sperimentati.
Per capire meglio: sia Y il volume di vendita di un prodotto, X_i grandezze socio-economiche del territorio; una volta costruito il modello sulla base di dati sperimentali in una regione, è possibile esportare il modello in regioni meno note prevedendo le vendite, sulla base dei dati socio-economici di quella nuova regione.

Nel seguito ci concentreremo sullo studio di relazioni funzionali di tipo lineare:

$$Y = a_1 X_1 + \cdots + a_p X_p + b + \epsilon$$

con a_i e b parametri del modello; b è detta **intercetta**. Questa relazione è detta **regressione lineare multipla**.
Come per la regressione lineare semplice, ci poniamo l'obiettivo di trovare i parametri del modello partendo da una tabella di dati sperimentali.
A partire dai dati specifici, esaminiamo una relazione lineare tra essi della forma

$$y_i = a_1 x_{i,1} + \cdots + a_p x_{i,p} + b + \epsilon_i \quad \text{con } i = 1, ..., n \tag{1.1}$$

dove abbiamo introdotto i numeri ϵ_i, detti **residui**, definiti dalla relazione stessa, cioè da

$$\epsilon_i = y_i - (a_1 x_{i,1} + \cdots + a_p x_{i,p} + b).$$

Così facendo, non stiamo ipotizzando una relazione lineare tra le variabili: l'identità 1.1 vale sempre, con la definizione data dei residui.
Possiamo poi calcolare

Definizione 1.7. *Lo* ***scarto quadratico medio*** *dei residui è*

$$SQM(a_1, ..., a_p, b) = \frac{\sum_{i=1}^n \epsilon_i^2}{n} = \frac{\sum_{i=1}^n (y_i - (a_1 x_{i,1} + \cdots + a_p x_{i,p} + b))^2}{n}.$$

Questa grandezza misura la validità del modello lineare: se piccola, il modello è buono.
Pertanto la strategia sarà quella di cercare i parametri $\hat{a_1}, ..., \hat{a_p}, \hat{b}$ che rendono minimo $SQM(a_1, ..., a_p, b)$. Tali parametri forniscono il migliore tra i modelli lineari.

1.3.1 Calcolo dei coefficienti: metodo dei minimi quadrati

Il metodo dei minimi quadrati è un metodo universale per la ricerca di parametri ottimali (nel senso che con opportune modifiche lo si applica ai più svariati contesti).
Procediamo dunque al calcolo dei parametri.
Introduciamo innanzitutto la matrice degli input:

$$X = \begin{pmatrix} x_{1,1} & \cdots & x_{1,p} & 1 \\ \cdots & \cdots & \cdots & \cdots \\ x_{n,1} & \cdots & x_{n,p} & 1 \end{pmatrix}$$

(l'ultima colonna serve fittiziamente per manipolare l'intercetta); i vettori dei parametri, degli output e dei residui:

$$a = \begin{pmatrix} a_1 \\ \vdots \\ a_p \\ b \end{pmatrix}, y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

ottenendo la scrittura vettoriale dell'equazione 1.1:

$$y = aX + \epsilon.$$

Allora si può riscrivere anche l'errore quadratico medio

$$SQM(a) = \frac{\sum_{i=1}^{n} \epsilon_i^2}{n} = \frac{|\epsilon|^2}{n} = \frac{|y - Xa|^2}{n}.$$

Vogliamo dunque minimizzarlo.
Nello spazio $\mathbb{R}^n$ abbiamo un punto dato y ed il sottospazio ImX, l'immagine di $\mathbb{R}^{p+1}$ attraverso X, cioè $ImX = \{z \in \mathbb{R}^n / z = Xa, a \in \mathbb{R}^{p+1}\}$.
Cercare il punto $\hat{a} \in \mathbb{R}^{p+1}$ che minimizza $|y - Xa|$ equivale allora a cercare il punto $\hat{z} \in ImX$ che minimizza $|y - z|$, dovendo poi come passo finale trovare $\hat{a} \in \mathbb{R}^{p+1}$ tale che $X\hat{a} = \hat{z}$.

Lemma 1.1. *Dati un punto $y \in \mathbb{R}^n$ ed un sottospazio $V \subset \mathbb{R}^n$, esiste uno e un solo punto $\hat{z} \in V$ tale che*

$$|y - \hat{z}| \leq |y - z| \quad \forall z \in V.$$

Inoltre, il vettore $y - \hat{z}$ è perpendicolare a V.

Dimostrazione. Sia d la dimensione di V, $d \leq n$.
Consideriamo una base ortonormale $e_1, ..., e_n$ di $\mathbb{R}^n$ tale che $e_1, ..., e_d$ sia base di V.
Allora i vettori $e_{d+1}, ..., e_n$ sono ortogonali a V.
Il punto y si può scrivere come

$$\begin{aligned} y &= a_1 e_1 + \cdots + a_n e_n \\ &= a_1 e_1 + \cdots + a_d e_d + a_{d+1} e_{d+1} + \cdots + a_n e_n \\ &= \hat{z} + w \end{aligned}$$

dove $\hat{z} = a_1 e_1 + \cdots + a_d e_d \in V$, $w = y - \hat{z}$ perpendicolare a V.
Se $z \in V$, $z = b_1 e_1 + \cdots + b_d e_d$, allora

$$y - z = (a_1 - b_1)e_1 + \cdots + (a_d - b_d)e_d + a_{d+1} e_{d+1} + \cdots + a_n e_n.$$

Quindi

$$|y - \hat{z}| = |w| \leq |y - z| \quad \forall z \in V.$$

□

Lemma 1.2. *Supponiamo che $KerX = \{0\}$. Allora, dato $\hat{z} \in ImX$, esiste uno e un solo punto $\hat{a} \in \mathbb{R}^{p+1}$ tale che $X\hat{a} = \hat{z}$.*

Dimostrazione. L'esistenza è insita nell'ipotesi $\hat{z} \in ImX$:

$$\hat{z} \in ImX \Rightarrow \exists \hat{a} \in \mathbb{R}^{p+1} \quad \text{tale che} \quad X\hat{a} = \hat{z}.$$

L'unicità discende dall'ipotesi $KerX = \{0\}$:
se $\tilde{a}, \hat{a}$ fossero due soluzioni allora $X(\hat{a}-\tilde{a}) = 0$, cioè $\hat{a}-\tilde{a} \in KerX$, cioè $\hat{a}-\tilde{a} = 0$. □

Lemma 1.3. *Se* $KerX = \{0\}$ *allora* $det(X^TX) \neq 0$.

Dimostrazione. Se $KerX = \{0\}$ allora $\forall v \neq 0$ si ha $Xv \neq 0$, quindi $|Xv|^2 \neq 0$, cioè $< X^TXv, v > \neq 0$. Ma allora $X^TXv \neq 0$.
Abbiamo mostrato che $v \neq 0$ implica $X^TXv \neq 0$ quindi $KerX^TX = \{0\}$; ma X^TX è una matrice quadrata quindi è invertibile, ovvero $det(X^TX) \neq 0$. □

Il seguente risultato fornisce una formula per il calcolo dei coefficienti.

Teorema 1.4. *Supponiamo valga* $KerX = \{0\}$. *Allora esiste uno e un solo vettore* $\hat{a}$ *che minimizza la funzione*

$$f(a) = |y - Xa|^2.$$

Esso è dato da

$$\hat{a} = (X^TX)^{-1}X^Ty.$$

Se inoltre $p + 1 = n$ *allora il minimo è nullo e vale* $\hat{a} = X^{-1}y$.

Dimostrazione. Dato $y \in \mathbb{R}^n$, siano $\hat{z} \in ImX$ e $\hat{a} \in \mathbb{R}^{p+1}$ tali che $\hat{z}$ minimizza $|y - z|$ (esistenza assicurata dal lemma 1.1), $\hat{a}$ risolve $X\hat{a} = \hat{z}$ (lemma 1.2).
Preso $a \neq \hat{a}$ e posto $z = Xa$, vale $z \neq \hat{z}$ e $z \in ImX$.
Quindi

$$|y - \hat{z}| < |y - z|$$

ovvero

$$|y - X\hat{a}| < |y - Xa|.$$

Pertanto $f(\hat{a}) < f(a) \quad \forall a \neq \hat{a}$, quindi $\hat{a}$ è punto di minimo ed è l'unico.
Siccome $\hat{a}$ soddisfa $X\hat{a} = \hat{z}$, allora (moltiplicando a sinistra ambo i membri per X^T) soddisfa anche $X^TX\hat{a} = X^T\hat{z}$.
Per ipotesi e per il lemma 1.3, X^TX è invertibile quindi $\hat{a} = (X^TX)^{-1}X^T\hat{z}$; ma $< X^T(\hat{z} - y), a >=< \hat{z} - y, Xa >= 0 \quad \forall a$, perchè $\hat{z} - y$ è perpendicolare a ImX (dal lemma 1.1).
Quindi $X^T(\hat{z} - y) = 0$, perciò $X^T\hat{z} = X^Ty$.
Allora

$$\hat{a} = (X^TX)^{-1}X^Ty.$$

Infine, se $p + 1 = n$ (quindi $detX \neq 0$ perchè $KerX = \{0\}$) le matrici X e X^T sono invertibili, allora

$$\hat{a} = (X^TX)^{-1}X^Ty = X^{-1}(X^T)^{-1}X^Ty = X^{-1}y.$$

□

1.3.2 Significato del valore numerico dei parametri

Supponiamo di aver eseguito la regressione e di aver quindi trovato i valori numerici dei parametri $\hat{a}_1, ..., \hat{a}_p, \hat{b}$.
Supponiamo ad esempio che $\hat{a}_1 = 3.5$; ciò significa che se aumentiamo il valore di X_1 di un'unità, il valore di Y aumenta di 3.5 unità: se $y = \hat{a}_1 x_1 + \cdots + \hat{a}_p x_p + \hat{b} + \epsilon$ allora $y + \hat{a}_1 = (x_1 + 1)\hat{a}_1 + \cdots + \hat{a}_p x_p + \hat{b} + \epsilon$.
Quindi il valore numerico dei parametri fornisce la misura della variazione di Y in corrispondenza di variazioni unitarie dei fattori.
Oltre a questo, la cosa basilare è il segno dei parametri: se $\hat{a}_1 \geq 0$, allora Y cresce al crescere di X_1; se $\hat{a}_1 < 0$ allora Y cresce al decrescere di X_1.
Questo permette ragionamenti interpretativi sul fatto che certe grandezze varino in modo concorde o discorde.

Capitolo 2

Riduzione di un modello: il metodo PCA

2.1 Metodo della varianza spiegata

Come per la regressione lineare semplice, anche nel caso di più predittori si può parlare di varianza spiegata (o indice R^2) definita da

$$R^2 = 1 - (\sigma_\epsilon^2/\sigma_Y^2)$$

dove σ_Y^2 è la varianza dei dati $y_1, ..., y_n$ e $\sigma_\epsilon^2 = SQM(\hat{a}_1, ..., \hat{a}_p, \hat{b})$, ovvero $\sigma_\epsilon^2 = \frac{\sum_{i=1}^{n} (y_i - (\hat{a}_1 x_{i,1} + \cdots + \hat{a}_p x_{i,p} + \hat{b}))^2}{n}$.
Ora ci poniamo il seguente problema: in genere si tende a preferire un modello matematico con pochi fattori, ma nei casi applicativi "fortunati" si posseggono molti dati e molti fattori potenzialmente utili. Quali utilizzare?
Per esempio, si pensi a una situazione con 20 fattori (variabili economiche, sociali, etc.) che potrebbero essere utili per spiegare il problema della disoccupazione. Gli individui potrebbero essere le regioni italiane, le nazioni europee o mondiali.
Quale gruppo di variabili può spiegare al meglio la disoccupazione? E tramite quale modello regressivo?
Un metodo ragionevole e abbastanza intuitivo consiste nel partire dal modello completo, con tutti i fattori, ed eliminarne uno alla volta fino al raggiungimento del modello ridotto ottimale. Ma quando si arriva effettivamente al modello finale? Ovvero quando ci si deve fermare? Qui entra in gioco la varianza spiegata R^2 e la sua utilità è spiegata dalla seguente

Proposizione 2.1. *Siano*

$$Y = a_1 X_1 + \cdots + a_p X_p + b + \epsilon$$
$$Y = a_1 X_1 + \cdots + a_p X_p + a_{p+1} X_{p+1} + b + \epsilon$$

due modelli relativi alla stessa tabella di dati.
Allora il modello con più variabili ha R^2 maggiore.

Dimostrazione. $R^2 = 1 - (\sigma_\epsilon^2/\sigma_Y^2)$. Il numero σ_Y^2 è lo stesso per entrambi i modelli, mentre σ_ϵ^2 cambia.
Indichiamo con R_p^2 e R_{p+1}^2 i valori R^2 relativi al modello con p e p+1 fattori rispettivamente. In modo simile definiamo $\sigma_{\epsilon,p}^2$ e $\sigma_{\epsilon,p+1}^2$.
Per dimostrare la tesi, è sufficiente dimostrare che $\sigma_{\epsilon,p}^2 \geq \sigma_{\epsilon,p+1}^2$.
Infatti, se questo vale, allora
$\frac{\sigma_{\epsilon,p}^2}{\sigma_Y^2} \geq \frac{\sigma_{\epsilon,p+1}^2}{\sigma_Y^2}$, perciò $1 - \frac{\sigma_{\epsilon,p}^2}{\sigma_Y^2} \leq 1 - \frac{\sigma_{\epsilon,p+1}^2}{\sigma_Y^2}$, quindi $R_p^2 \leq R_{p+1}^2$.
Dunque,
$\sigma_{\epsilon,p}^2 = (SQM)_p(\hat{a}_1, ..., \hat{a}_p, \hat{b}) = \frac{\sum_{i=1}^n (y_i - (\hat{a}_1 x_{i,1} + \cdots + \hat{a}_p x_{i,p} + \hat{b}))^2}{n}$.
$\sigma_{\epsilon,p+1}^2 = (SQM)_{p+1}(\hat{a}_1, ..., \hat{a}_p, \hat{a}_{p+1}, \hat{b}) = \frac{\sum_{i=1}^n (y_i - (\hat{a}_1 x_{i,1} + \cdots + \hat{a}_p x_{i,p} + \hat{a}_{p+1} x_{i,p+1} + \hat{b}))^2}{n}$.
Pertanto $(SQM)_{p+1}(\hat{a}_1, ..., \hat{a}_p, 0, \hat{b}) = (SQM)_p(\hat{a}_1, ..., \hat{a}_p, \hat{b})$.
Allora

$$\begin{aligned}(SQM)_{p+1}(\hat{a}_1, ..., \hat{a}_p, \hat{a}_{p+1}, \hat{b}) &\leq (SQM)_{p+1}(\hat{a}_1, ..., \hat{a}_p, 0, \hat{b}) \\ &= (SQM)_p(\hat{a}_1, ..., \hat{a}_p, \hat{b}).\end{aligned}$$

Quindi
$\sigma_{\epsilon,p}^2 \geq \sigma_{\epsilon,p+1}^2$. □

Pertanto R^2 non può essere usato direttamente per scegliere il modello migliore perchè prediligerebbe sempre quello con più fattori, ma può risultare utile per analizzare il grado di peggioramento del modello al momento dell'eliminazione dei fattori. Dunque, eliminado fattori R^2 decresce (peggiora), ma accadrà che eliminando certi fattori esso non cambierà molto: in questo caso, l'eliminazione di quei fattori non peggiora granchè la spiegazione che il modello dà della variabilità di Y; mentre ad un certo punto, eliminando un ulteriore fattore, R^2 calerà in maniera rilevante, segno che con tale eliminazione il modello è peggiorato e si sono perse informazioni importanti sulla variabilità di Y.
Così si può decidere di terminare il processo di eliminazione, prima di eliminare tale fattore determinante.
Ovviamente le variazioni numeriche di R^2 possono essere sfumate, quindi molte decisioni restano soggettive. E' consigliabile avere sotto controllo anche altri parametri, al fine di evitare decisioni superficiali: uno di questi può essere la matrice di covarianza ottenuta dalla tabella di dati con anche la variabile Y, la quale aiuta a capire quali siano le variabili più legate ad essa.

Un metodo alternativo e sicuramente più rigoroso è dato dal metodo PCA, il quale tratteremo in seguito all'introduzione di alcune nozioni utili.

2.2 Vettori gaussiani

Ricordiamo che una variabile aleatoria gaussiana o normale $\mathbb{N}(\mu, \sigma^2)$ è una v.a. con densità di probabilità

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-|x-\mu|^2}{2\sigma^2}\right)$$

Ricordiamo, inoltre, il seguente fatto importante: se Z è una v.a. di tipo normale standard, allora $\mu + \sigma$Z è $\mathbb{N}(\mu, \sigma^2)$, ed ogni gaussiana $\mathbb{N}(\mu, \sigma^2)$ si può scrivere nella forma $\mu + \sigma$Z con Z una normale standard.

Definizione 2.1. *Un **vettore normale (o gaussiano) standard in dimensione d** è un vettore aleatorio $Z = (Z_1, ..., Z_d)$ con densità congiunta*

$$f(z_1, ..., z_d) = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z_i^2}{2}\right) = \frac{1}{\sqrt{(2\pi)^d}} \exp\left(-\frac{z_1^2 + ... + z_d^2}{2}\right).$$

$X = (X_1, ..., X_n)$ si dice vettore gaussiano se può essere rappresentato nella forma

$$X = AZ + b$$

dove $Z = (Z_1, ..., Z_d)$ è un vettore normale standard, $A : \mathbb{R}^d \longrightarrow \mathbb{R}^n$ è una matrice e b un vettore di $\mathbb{R}^n$.

Ora analizziamo alcuni risultati che ci serviranno in seguito.

Proposizione 2.2. *Siano $X = (X_1, ..., X_n)$ un vettore casuale, $A : \mathbb{R}^n \longrightarrow \mathbb{R}^d$ una matrice, b un vettore di $\mathbb{R}^d$, $Y = (Y_1, ..., Y_d)$ un vettore casuale definito da $Y = AX + b$.*
Sia $\mu^X = (\mu_1^X, ..., \mu_n^X)$ il vettore dei valori medi di X, ovvero $\mu_i^X = E[X_i]$.
Sia μ^Y il vettore dei valori medi di Y.
Allora

$$\mu_Y = A\mu_X + b.$$

Dimostrazione. L'identità Y=AX+b scritta per componenti diventa

$$Y_i = \sum_{j=1}^{n} A_{ij} X_j + b_i.$$

Pertanto, per linearità del valor medio,

$$E[Y_i] = E[\sum_{j=1}^{n} A_{ij} X_j + b_i] = \sum_{j=1}^{n} A_{ij} E[X_j] + b_i.$$

Questa è esattamente la versione per componenti di $\mu_Y = A\mu_X + b$. □

Proposizione 2.3. *Siano $X = (X_1, ..., X_n)$ un vettore casuale, $A : \mathbb{R}^n \longrightarrow \mathbb{R}^d$ una matrice, b un vettore di $\mathbb{R}^d$, $Y = (Y_1, ..., Y_d)$ un vettore casuale definito da Y=AX+b. Se Q^X e Q^Y sono le matrici di covarianza di X e Y rispettivamente, allora*

$$Q^Y = AQ^X A^T.$$

Dimostrazione. Usando sempre l'identità per componenti $Y_i = \sum_{j=1}^n A_{ij}X_j + b_i$, risulta

$$\begin{aligned} Q^Y_{ij} = Cov(Y_i, Y_j) = Cov(\sum_{i'=1}^n A_{ii'}X_{i'} + b_i, \sum_{j'=1}^n A_{jj'}X_{j'} + b_j) = \\ = \sum_{i'=1}^n A_{ii'} Cov(X_{i'}, \sum_{j'}^n A_{jj'}X_{j'} + b_j) = \\ = \sum_{i'=1}^n A_{ii'} \sum_{j'=1}^n A_{jj'} Cov(X_{i'}, X_{j'}) = \sum_{i'=1}^n \sum_{j'=1}^n A_{ii'} Q^X_{i'j'} A_{jj'}. \end{aligned}$$

In generale, date due matrici A e B, vale

$$(AB)_{ij} = \sum_k A_{ik}B_{kj}.$$

Allora

$$\begin{aligned} \sum_{i'=1}^n \sum_{j'=1}^n A_{ii'} Q^X_{i'j'} A_{jj'} = \sum_{j'=1}^n (AQ^X)_{ij'} A_{jj'} = \\ = \sum_{j'=1}^n (AQ^X)_{ij'} A^T_{j'j} = (AQ^X A^T)_{ij}. \end{aligned}$$

□

Ora siamo pronti per enunciare una particolare proprietà dei vettori gaussiani.

Proposizione 2.4. *Se X è un vettore gaussiano della forma X=AZ+b (con le notazioni della definizione), allora il vettore μ dei valori medi e la matrice di covarianza Q di X sono dati da*

$$\mu = b$$
$$Q = AA^T$$

Dimostrazione. Dalla definizione di vettore gaussiano deriva direttamente che $\mu^Z = 0$ e $Q^Z = Id$. Pertanto applicando le proposizioni precedenti si ottiene banalmente la tesi. □

La matrice di covarianza sappiamo essere diagonalizzabile (in quanto simmetrica), quindi $Q = UQ_eU^T$ dove U è ortogonale e Q_e è diagonale con gli autovalori di Q sulla diagonale, e definita non negativa (quindi i suoi autovalori sono non negativi). Dalla diagonalizzazione discende che Q è invertibile se e solo se i suoi autovalori sono strettamente positivi se e solo se Q è definita positiva.
Sappiamo inoltre che se Q è invertibile, allora $X = (X_1, ..., X_n)$ possiede densità congiunta data da

$$f(x) = \frac{1}{\sqrt{(2\pi)^n det(Q)}} \exp\left(-\frac{(x-\mu)^T Q^{-1}(x-\mu)}{2}\right)$$

dove $x = (x_1, ..., x_n) \in \mathbb{R}^n$.

Teorema 2.1. *Un vettore gaussiano $X = (X_1, ..., X_n)$ centrato ($\mu_X = 0$) è standard se e solo se ha matrice di covarianza Q uguale alla matrice identità.*

Dimostrazione. Se X è standard allora $Q_{ii} = Var[X_i] = 1$ per $i = 1, ..., n$, e $Q_{ij} = Cov(X_i, X_j) = 0$ per $i \neq j$ in quanto le componenti di X sono gaussiane standard indipendenti. Quindi $Q = Id$.
Viceversa, se Q è la matrice identità allora $X = (X_1, ..., X_n)$ possiede densità congiunta data da

$$f(x) = \frac{1}{\sqrt{(2\pi)^n}} \exp\left(-\frac{x^T x}{2}\right)$$

e quindi X è una gaussiana standard. □

2.2.1 Raffigurazioni

Il grafico della funzione di densità gaussiana può essere rappresentato attraverso curve di livello.

Definizione 2.2. *Dati $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ e $a \in \mathbb{R}$, l'**insieme di livello a** è il luogo dei punti $x \in \mathbb{R}^n$ tali che $f(x) = a$.*

Nel caso di una densità *f*, essendo positiva, ha senso esaminare solo il caso $a > 0$. Nel caso della gaussiana $N(\mu, Q)$, dobbiamo analizzare l'equazione

$$\frac{1}{\sqrt{(2\pi)^n det(Q)}} \exp\left(-\frac{(x-\mu)^T Q^{-1}(x-\mu)}{2}\right) = a.$$

Per semplicità algebrica restringiamoci al caso $\mu = 0$, altrimenti basterà traslare di μ il risultato finale. Poniamo $C = \sqrt{(2\pi)^n det(Q)}$. Allora vale

$$\exp\left(-\frac{1}{2}x^T Q^{-1}x\right) = aC$$

e posto $a^{'} = -2\log(aC)$, l'equazione diventa

$$x^T Q^{-1}x = a^{'}.$$

Questa, per $a^{'} \geq 0$, è l'equazione di un'elissoide.
Infatti, usando la scomposizione $Q = UQ_eU^T$ e ricordando che

$$Q^{-1} = (U^T)^{-1}Q_e^{-1}U^{-1} = UQ_e^{-1}U^T$$

l'equazione diventa

$$x^T UQ_e^{-1}U^T x = a^{'}.$$

Posto $y = U^T x$ troviamo

$$y^T Q_e^{-1} y = a^{'}$$

che in coordinate si legge

$$\frac{y_1^2}{\lambda_1} + \cdots + \frac{y_n^2}{\lambda_n} = a^{'} \quad \text{con } \lambda_i \text{ autovalori di Q}$$

ovvero un'elissoide.
Più precisamente, le coordinate y sono quelle nella base $e_1, ..., e_n$ di autovettori di Q, quindi si tratta di un'elissoide rispetto a tale base, cioè avente $e_1, ..., e_n$ come assi. Inoltre, la lunghezza dei semiassi è $\sqrt{\lambda_1}, ..., \sqrt{\lambda_n}$.
In conclusione (torniamo a μ qualsiasi):

Proposizione 2.5. *Le ipersuperfici (curve per $n = 2$) di livello di un vettore gaussiano $N(\mu, Q)$ in $\mathbb{R}^n$ sono elissoidi di centro μ ed assi dati dagli autovettori di Q, con lunghezze degli assi pari alle radici degli autovalori.*

2.3 Il metodo delle componenti principali

Una tabella della forma

	X_1	$\cdots$	X_p
1	$x_{1,1}$	$\cdots$	$x_{1,p}$
2	$x_{2,1}$	$\cdots$	$x_{2,p}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$
n	$x_{n,1}$	$\cdots$	$x_{n,p}$

può essere pensata come un insime di n punti di $\mathbb{R}^p$: i punti $(x_{i,1}, ..., x_{i,p})$ al variare di $i = 1, ..., n$.
Ora, da un lato, se $p > 2$ la raffigurazione è difficile ($p = 3$) o impossibile ($p > 3$), quindi vogliamo effettuare proiezioni 2-dimensionali dell'insieme dei punti. Dall'altro lato, nel momento in cui eseguiamo tali proiezioni, queste potrebbero essere poco leggibili se i punti sono troppo sovrapposti. Perciò bisogna trovare la visuale 2-dimensionale più conveniente.
Questo è il **problema di PCA**: *rappresentare un insieme di punti di* $\mathbb{R}^p$ *nel modo bidimensionale più sparpagliato possibile.*
La soluzione è immediata se si utilizzano le idee della teoria dei vettori gaussiani.
Si immagini che i punti suddetti di $\mathbb{R}^p$ siano realizzazioni sperimentali di un vettore gaussiano $X = (X_1, ..., X_p)$; sotto tale ipotesi, essi si dispongono in maniera pressappoco elissoidale.
Come abbiamo visto, per conoscere gli elissoidi (insiemi di livello della densità) serve la matrice di covarianza Q, la quale viene calcolata utilizzando i dati in tabella. Successivamente possiamo trovare autovalori e autovettori che indicheremo con $e_1, ..., e_p$ e $\lambda_1, ..., \lambda_p$.
Gli elissoidi hanno quindi $e_1, ..., e_p$ come assi e $\sqrt{\lambda_1}, ..., \sqrt{\lambda_p}$ come lunghezze.
Gli assi con le lunghezze maggiori sono da preferire nell'ottica detta sopra di avere una visione più sparpagliata possibile dei punti sperimentali.
Volendo appunto una visione bidimensionale, fissiamo gli assi e_1, e_2 (adottando la convenzione che sia $\lambda_1 \geq ... \geq \lambda_p$).
Il piano individuato da e_1 e e_2 è detto **piano principale** ed è quello rispetto a cui è maggiore lo sparpagliamento dei dati.
Volendo proiettare i punti su tale piano, si calcolano i punti di coordinate $< x, e_1 >$ e $< x, e_2 >$ al variare di x nell'insieme dei punti sperimentali considerato.
Lo scopo, con metodo PCA e la sua visualizzazione finale, è di vedere come gli individui (nazioni, regioni ecc...) si pongono gli uni rispetto agli altri: se sono raggruppati, come si suddividono medianti gli assi, ecc.

2.3.1 Varianza lungo le componenti principali

I punti proiettati lungo l'asse principale sono dati dai numeri $< x, e_1 >$.
Chiamiamo e_1 **componente principale** (o **prima componente principale**) ed usiamo il linguaggio di componenti principali anche per e_2, e_3 e così via (es. e_2 è la seconda componente principale).
In maniera simile al caso della prima, consideriamo anche i punti proiettati lungo la i-esima componente principale, ovvero $< x, e_i >$ al variare di x.
Qual è la varianza di questi numeri?
Supponiamo di avere il vettore aleatorio $X = (X_1, ..., X_p)$, la sua matrice di cova-

rianza Q e di averla diagonalizzata ($Q = UQ_eU^T$, $\lambda_1, ..., \lambda p$ autovalori).
Sia $V_i =< X, e_i >$ la proiezione del vettore aleatorio sulla componente principale e_i.

Teorema 2.2. *$Var[V_i] = \lambda_i$. Inoltre, $Cov(V_i, V_j) = 0$, quindi V_i e V_j sono indipendenti se X è gaussiano.*

Dimostrazione. Sia $V = U^T X$, allora $V = (V_1, \cdots, V_p)$ e $Q_V = U^T QU$ dove $Q = Q_X$.
$Q = UQ_eU^T \Rightarrow Q_e = U^T QU$. Quindi

$$Q_V = Q_e = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix}$$

Ma $V = (V_1, ..., V_p)$, quindi $Var[V_i] = \lambda_i$ e $Cov(V_i, V_j) = 0$.
Se X è gaussiano anche V lo è, quindi l'ultima affermazione del teorema è verificata. □

Osservazione 2.1. *Le proiezioni V_i sono le coordinate di X rispetto alla base $e_1, ..., e_p$:*

$$X =< X, e_1 > e_1 + \cdots + < X, e_p > e_p = V_1 e_1 + \cdots + V_p e_p.$$

*Quindi la **varianza totale del vettore $\boldsymbol{X}$** è*

$$\lambda_1 + \cdots + \lambda_p$$

pari a $Var[V_1] + \cdots + Var[V_p]$.
*Pertanto si può vedere il numero $\frac{\lambda_1}{\lambda_1 + \cdots + \lambda_p}$ come la **proporzione di varianza spiegata dall'asse principale** e il numero $\frac{\lambda_1 + \lambda_2}{\lambda_1 + \cdots + \lambda_p}$ come la **proporzione di varianza spiegata dal piano principale** (detta anche varianza spiegata cumulativa).*

2.3.2 Il metodo PCA per capire la dimensione di un problema

Con dimensione di un problema intuitivamente si intende un indicatore della sua complessità.
La varianza lungo le componenti principali è utile per farsi un'idea della dimensione dei dati, cioè quante componenti sono necessarie per analizzare i dati in maniera esaustiva.
La varianza spiegata cumulativa è il principale parametro dell'efficacia del metodo PCA, dato che quantifica l'accuratezza della visualizzazione dei dati data dal piano principale.

In genere, si considera il metodo PCA efficiente quando il piano principale rappresenta l'80% o 90% della varianza totale dei dati, cioè quando la parte di informazione persa attraverso la proiezione bidimensionale si aggira attorno al 10-20% del totale. Tuttavia, anche quando la rappresentazione bidimensionale data dal piano principale è insufficiente, il metodo PCA rimane comunque un buon metodo per comprendere meglio i dati analizzati, in particolare indicando quante variabili al minimo bastano per rappresentarli efficacemente.
Vediamo un esempio.
Esaminiamo cinque potenziali indicatori di benessere nelle diverse regioni italiane:

X_1 = PLIC (posti letto in istituti di cura)
X_2 = SC (spese complessive per famiglia)
X_3 = SA.SC (proporzione di SC dedicata agli alimentari)
X_4 = TD (tasso di disoccupazione)
X_5 = TMI (tasso di mortalità infantile)

e consideriamo la seguente tabella di dati:

	PLIC	SC	SA.SC	TD	TMI
Sicilia	0.08833073	0.471218868	-0.70738393	-0.6079983	-0.395594374
Valle d'Aosta	-1.54531281	0.348570963	-0.64222892	-0.8134750	1.578973068
Basilicata	0.20230586	1.397587799	-0.83615834	-0.7908017	-0.538580292
Campania	0.67720223	0.435951016	-1.26986734	-0.9665197	-0.075578271
Puglia	0.08833073	1.334386404	-1.21054741	-0.8489020	-0.497727172
Sardegna	0.63921052	-0.005441075	-1.02808828	-0.8049725	-1.301171855
Liguria	1.19009032	-0.247332925	0.47073536	-0.4294462	-0.354741254
Calabria	0.65820638	1.177628694	-1.31590440	-0.8630728	-0.347932401
Abruzzo	0.12632244	1.092670016	-0.79594148	-0.6448424	-1.355642681
Umbria	-1.43133768	0.675982968	-0.14084928	-0.5243906	-1.287554149
Marche	0.27828928	1.090985581	-0.26509116	-0.7029427	-0.000680885
Molise	2.32984163	0.546807878	-0.08030122	-0.1134373	-0.014298592
Toscana	0.33527684	-0.373279515	0.40242546	-0.4563707	0.040172234
Lazio	0.65820638	-1.289120729	0.06583846	0.4519778	-1.151377084
Trentino Alto Adige	-1.81125478	-1.314422342	2.03132778	1.6649986	0.414659163
Veneto	-0.76648275	-0.926405778	1.03856609	0.6489520	1.109162194
Lombardia	-0.74748690	-1.154085209	0.66193679	0.8445091	2.001121969
Emilia Romagna	-0.50054078	-1.727319412	1.57182583	2.1538913	0.632542467
Piemonte	-0.91844959	-1.130924231	1.33235861	1.5176222	1.783238665
Friuli-Venezia Giulia	0.44925197	-0.403458971	0.71734736	1.285221	-0.238990749

Ci troviamo quindi con 20 punti disposti su una nuvoletta simile ad un'elissoide in 5 dimensioni. E' abbastanza difficile visualizzare una tale figura; l'idea di base del metodo PCA è quella di operare un cambio di variabili, cioè un cambio di base dello spazio vettoriale di dimensione 5 in cui stiamo lavorando, in modo da avere una proiezione bidimensionale dove i dati sono, come al solito, il più distinti tra loro.
Per fare ciò, è necessario innanzitutto calcolare la matrice di covarianza Q relativa alla tabella dei dati

$$Q = \begin{pmatrix} 1 & 0.32 & -0.41 & -0.37 & -0.44 \\ 0.32 & 1 & -0.84 & -0.85 & -0.48 \\ -0.41 & -0.84 & 1 & 0.90 & 0.51 \\ -0.37 & -0.85 & 0.90 & 1 & 0.49 \\ -0.44 & -0.48 & 0.51 & 0.49 & 1 \end{pmatrix}$$

e gli autovettori relativi agli autovalori maggiori; in questo caso risulta:

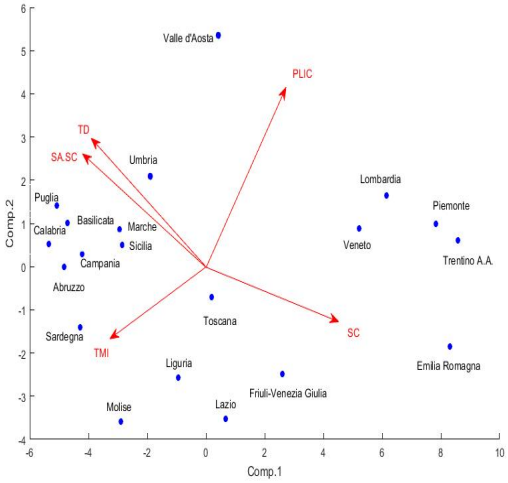
$$e_1 = (-0.817, -1.302, 1.356, 1.338, 1)$$

$$e_2 = (-1.749, 0.7, -0.484, -0.644, 1)$$

Successivamente si proiettano le realizzazioni sperimentali riassunte nella prima tabella (viste come punti di $\mathbb{R}^5$) sul piano principale:

	Comp.1	Comp.2
Sicilia	-2.85	0.51
Valle D'Aosta	0.43	5.36
Basilicata	-4.72	1.01
Campania	-4.21	0.28
Puglia	-5.08	1.41
Sardegna	-4.28	-1.41
Liguria	-0.94	-2.56
Calabria	-5.36	0.52
Abruzzo	-4.82	-0.01
Umbria	-1.90	2.09
Marche	-2.95	0.86
Molise	-2.90	-3.59
Toscana	0.19	-0.71
Lazio	0.68	-3.53
Trentino Alto Adige	8.59	0.61
Veneto	5.22	0.88
Lombardia	6.14	1.64
Emilia Romagna	8.30	-1.83
Piemonte	7.84	0.98
Friuli-Venezia Giulia	2.61	-2.48

Per una visione grafica di tale risultato, si può utilizzare il software R che, inserendo la tabella dei dati, con il comando *princomp* calcola tutti i valori sopra riportati e li rappresenta attraverso il seguente grafico, che è appunto la proiezione bidimensionale che stiamo cercando.



Una prima analisi qualitativa può essere svolta in base ai rapporti tra i vettori che rappresentano gli indicatori (ortogonalità, parallelismo con versi concordi o discordi, ecc...), e ai raggruppamenti e alle posizioni dei dati:

- SC, TD e SA.SC sono quasi tutti paralleli, quindi vi è una forte correlazione tra di loro: si può pensare, ad esempio, alla direzione comune come a un indicatore complessivo di benessere economico.
- Il verso di SC è opposto a quelli di TD e SA.SC, segno che questi indicatori sono correlati negativamente: una maggiore disoccupazione media si riflette su una minore spessa complessiva media, mentre se la spesa complessiva media è molto bassa sarà in gran parte dedicata agli alimentari. Allo stesso modo, la correlazione positiva tra TD e SA.SC indica che nelle zone a più alta disoccupazione le (poche) spese sono destinate per lo più a generi alimentari.
- PLIC e TMI sono abbastanza paralleli tra loro: come sopra, si può pensare alla direzione comune come a un indicatore complessivo di salute.
- PLIC e TMI sono abbastanza perpendicolari agli altri indicatori, indice che le due "direzioni"', indicatore complessivo di salute e indicatore complessivo di benessere economico, sono abbastanza scorrelati tra loro.

- L'area di maggiore benessere è quella nella direzione positiva di SC, e in tale zona si trovano raggruppate le regioni Lombardia, Veneto, Friuli-Venezia Giulia, Emilia Romagna, Piemonte e Trentino Alto Adige.
- La Valle d'Aosta eccelle in PLIC, a indicare una buona cura sanitaria.
- Particolarmente negativo, sia rispetto all'asse del benessere economico che a quello della salute, risulta il raggruppamento composto da Puglia, Calabria, Basilicata, Marche, Sicilia, Campania e Abruzzo.

Una possibile interpretazione delle prime due componenti principali, cioè delle nuove variabili aleatorie, potrebbe essere quindi quella dove la prima descrive il benessere di tipo economico e la seconda quello sanitario.

Pertanto, una buona idea quando si vuole creare un modello regressivo (o in particolare quando si vuole appunto ridurne la dimensione), è quella di affiancare l'analisi della matrice di correlazione e di R^2 all'esplorazione visiva offerta da PCA.

Capitolo 3

La classificazione degli individui

3.1 Problema della classificazione

In generale, nei problemi di classificazione si hanno due classi C1 e C2 (o più di due) ed un individuo (o più di uno) da classificare, cioè da assegnare a una delle classi.
Per fare questo, si cerca di stabilire una regola che, dietro osservazione di alcune caratteristiche dell'individuo in esame, lo assegni ad una delle classi.
Ovviamente se le caratteristiche osservabili includono anche la classe stessa, non vi è nessun problema da risolvere (ad esempio, se si deve decidere se un individuo è biondo o moro e si possono osservare i suoi capelli, l'assegnazione alla classe è ovvia).
Ci occuperemo, quindi, del caso in cui le caratteristiche osservabili siano in qualche modo indirette rispetto alle classi. Per capire meglio, si pensi ad un medico che deve stabilire se un paziente è affetto da una certa malattia, la quale è direttamente osservabile solo attraverso un'operazione chirurgica. Prima di sottoporre il paziente a tale intervento, il medico cercherà di capire attraverso esami non invasivi (esami del sangue, visite esterne,...) a quale classe appartiene il paziente, $C1 = sano$ o $C2 = malato$; è importante ricordare che gli esami non invasivi non danno la certezza di aver eseguito la classificazione corretta, la quale è assicurata unicamente dall'operazione chirurgica.
Supporremo che le caratteristiche osservabili siano descritte da variabili X_1, ..., X_m, l'analogo dei fattori della regressione, mentre la classe viene descritta dall'output Y.
Quindi, per riassumere, dobbiamo trovare delle regole che permettano di assegnare un individuo ad una delle due classi, C1 o C2, sulla base dei valori x_1, ..., x_m che misuriamo per quell'individuo.
Come si può costruire una tale regola?
L'idea è quella di usare n altri individui di cui si conoscano sia i valori x_{i_1}, ..., x_{i_m}, $i = 1, ..., n$, sia la classe C1 o C2. Questo campione di individui viene chiamato

training set e può essere riassunto in una tabella del tipo

	X_1	$\cdots$	X_m	Classe
1	$x_{1,1}$	$\cdots$	$x_{1,m}$	C1
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
n_1	$\cdots$	$\cdots$	$\cdots$	C1
$n_1 + 1$	$\cdots$	$\cdots$	$\cdots$	C2
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
n	$x_{n,1}$	$\cdots$	$x_{n,m}$	C2

3.2 La regressione logistica

Come si è studiato nel capitolo 1, la regressione è un buon strumento per riprodurre i valori di una variabile Y (l'output) in funzione dei valori assunti da una o più variabili (i predittori).
È proprio grazie ai modelli regressivi che troviamo una soluzione al problema della classificazione: la regressione logistica.

Perchè non si possono utilizzare i modelli della regressione lineare?
Principalmente per due motivi:

1. Essendo che la v.a. Y rappresenta la classe, essa può essere vista come una v.a. di Bernoulli. Nella regressione lineare la relazione tra l'output Y e i predittori è appunto di tipo lineare additivo, $Y = a_1X_1 + \cdots + a_mX_m + b + \epsilon$, quindi il modello assume che i valori della variabile Y possano variare tra $-\infty$ e $+\infty$. Nel nostro caso, essendo Y una bernoulliana, è espressa in termini di probabilità, quindi valori inferiori a 0 o superiori a 1 sono del tutto privi di senso;

2. La linearità degli effetti esercitati su Y è uno dei presupposti della regressione lineare. Peraltro, dato che la probabilità ha un campo di variazione limitato abbiamo buone ragioni per pensare che l'effetto sulla variazione esercitato da un certo predittore non è costante per tutti i livelli di probabilità su cui agisce.

In virtù di queste considerazioni, non solo il modello di regressione più adatto a studiare la probabilità come output dovrebbe tenere conto del fatto che si distribuisce tra 0 e 1, ma anche che la relazione tra i predittori e l'output non è lineare.
La regressione logistica soddisfa entrambi i requisiti. Per capire in che modo essa rappresenta gli effetti delle variabili $X_1, ..., X_m$ su Y, vediamone un esempio.
Si consideri il comportamento elettorale come variabile Y e il titolo di studio come

variabile X (predittore). Se il fatto di essere laureati aumenta la probabilità di andare a votare, l'effetto di tale caratteristica è necessariamente più contenuto quando agisce su una probabilità già elevata. Ad esempio, immaginiamo che da un'analisi dei dati risulti che gli iscritti ad un partitio abbiamo un'elevata probabilità di recarsi a votare (diciamo l'80%) indipendentemente dal proprio titolo di studio; è evidente che il fatto di essere laureati può aumentare la probabilità di andare a votare degli iscritti a un partito solo nella misura del 20% che manca per raggiungere il 100%. Invece, se un soggetto ha, indipendentemente dal titolo di studio, una probabilità di andare a votare del 50%, il fatto di essere laureato può aumentare la sua probabilità in maniera maggiore.
In ultimo possiamo immaginare una terza situazione: un soggetto che, indipendemente dal titolo di studio, è caratterizzato da una probabilità di andare a votare molto bassa (diciamo del 30%); in tal caso possiamo immaginare che il fatto di possedere una laurea comunque influenzi la probabilità di andare a votare, ma in misura ridotta, proprio perchè la probabilità di partenza è molto bassa.
Applicare un modello di regressione logistica significa ipotizzare questa particolare non linearità degli effetti dei predittori sulla probabilità $p = P(Y = 1)$: più intensi quando agiscono su livelli di probabilità intermedi; meno intensi quando agiscono su livelli di probabilità alti o bassi.

Nella regressione logistica, il parametro p è univocamente determinato dai predittori e dipende da loro attraverso una combinazione affine, detta **predittore lineare**:

$$\eta = a_1 X_1 + \cdots + a_m X_m + b.$$

Una volta noti i coefficienti $a_1, ..., a_m$ e b, dati i valori dei predittori $x_1, ..., x_m$, l'output è una v.a. di Bernoulli Y, di parametro p che dipende da η attraverso una certa funzione.
Essendo p una probabilità, non possiamo pensare che la relazione tra p e η sia del tipo $p = \eta$, cioè

$$p = a_1 x_1 + \cdots + a_m x_m + b$$

altrimenti otterremmo per p valori anche esterni a [0,1].
Si deve perciò adottare un modello del tipo

$$g(p) = a_1 x_1 + \cdots + a_m x_m + b$$

dove g è una funzione definita in [0,1], a valori reali e invertibile, in modo che sia

$$p = g^{-1}(a_1 x_1 + \cdots + a_m x_m + b).$$

La funzione g viene detta **mean function** e g^{-1} viene detta **link function**.
Una scelta molto comune è la funzione **logit**:

$$g(p) = \log(\frac{p}{1-p}).$$

Per $p \to 0$ tende a $-\infty$ e per $p \to 1$ tende a $+\infty$, ed è strettamente crescente.
La funzione inversa è

$$p = g^{-1}(p) = \frac{\exp(\eta)}{1 + \exp(\eta)}$$

Verifichiamo che siano effettivamente una l'inversa dell'altra:

$$\log(\frac{p}{1-p}) = \eta \Rightarrow \frac{p}{1-p} = \exp(\eta) \Rightarrow p = (1-p)\exp(\eta)$$

quindi

$$p(1 + \exp(\eta)) = \exp(\eta) \Rightarrow p = \frac{\exp(\eta)}{1 + \exp(\eta)}.$$

Di seguito è riportato il grafico di g^{-1}:

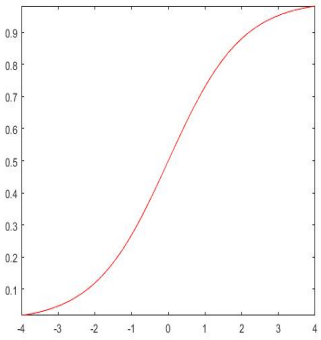


Figura 3.1: Funzione logistica $\frac{\exp(\eta)}{1+\exp(\eta)}$

In definitiva il modello di regressione logistica è

$$Y \sim Be(p) \quad \text{con} \quad p = \frac{\exp(\eta)}{1 + \exp(\eta)}, \quad \text{dove} \quad \eta = a_1 x_1 + \cdots + a_m x_m + b.$$

Quando i coefficienti $a_1, ..., a_m$ e b sono divenuti noti, preso un nuovo individuo e calcolati i valori $x_1, ..., x_m$ dei predittori, si può calcolare la probabilità p relativa a quell'individuo.
Se p è molto elevata, siamo abbastanza sicuri che per quell'individuo sarà $Y = 1$ (cioè appartiene alla classe C1), mentre se è molto bassa sarà $Y = 0$; per valori intermedi di p ovviamente c'è molta indecisione sul valore di Y, pur valendo comunque che se $p > \frac{1}{2}$ è più probabile $Y = 1$ e viceversa.

3.2.1 Calcolo dei coefficienti del modello

Il problema di trovare i coefficienti del modello logistico si risolve con il **metodo della massima verosimiglianza**.
In generale, questo metodo consiste nell'estrarre, da una famiglia di variabili aleatorie, un campione costituito da n variabili casuali X_i i.i.d. con funzione di distribuzione $f(x, \theta)$, e costruire la funzione di verosimiglianza che rappresenta la funzione di distribuzione del campione stesso: in questo ambito si suppone che essa sia funzione del vettore dei parametri θ, mentre le realizzazioni campionarie x_i sono fisse. Analiticamente si ha:

$$L(x_1, ..., x_n; \theta) = \prod_{i=1}^{n} f(x_i, \theta) \tag{3.1}$$

Definizione 3.1. *La funzione $\hat{\theta} = t(x_1, ..., x_n)$ è detta **stimatore di massima verosimiglianza** se, in ciascun campione estratto, assegna un valore al vettore θ che massimizza la funzione di verosimiglianza. In simboli:*

$$max L(x, \theta) = L(x, \hat{\theta}) \quad con\ x = (x_1, ..., x_n)\ fissato.$$

Ovviamente la stima di massima verosimiglianza è definita così:

$$\hat{\theta} = argmax\ L(x,\theta).$$

Per poter calcolare lo stimatore di massima verosimiglianza si ricorre alla funzione **log-verosimiglianza** ottenuta attraverso l'applicazione del logaritmo naturale, quindi risulta:

$$l(x, \theta) = \ln L(x, \theta).$$

Dato che la funzione logaritmica è una trasformazione monotòna crescente, con il passaggio alla log-verosimiglianza non si perdono le caratteristiche della funzione $L(x, \theta)$ in termini di crescenza e decrescenza e soprattutto si ottiene una forma analitica più semplice da trattare.
Nel caso di variabili casuali i.i.d. questo è particolarmente vero perchè la funzione di distribuzione congiunta del campione può essere espressa come produttoria delle marginali; per le proprietà del logaritmo, dalla 3.1 si ricava la log-verosimiglianza come sommatoria, infatti:

$$l(x, \theta) = \ln \left[\prod_{i=1}^{n} f(x_i, \theta) \right] = \sum_{i=1}^{n} \ln f(x_i, \theta).$$

Tornando ora al nostro modello di regressione logistica, abbiamo detto che una volta noti i valori $x_1, ..., x_m$ dei predittori di un individuo, Y é Be(p) con $p = g^{-1}(\eta)$, $\eta = a_1 x_1 + \cdots + a_m x_m + b$.

Quindi $P(Y = 1) = p$ e $P(Y = 0) = 1 - p$. Se indichiamo uno dei due numeri 0 o 1 con y, si può scrivere in una sola formula

$$P(Y = y) = p^y(1 - p)^{1-y}.$$

Per applicare il metodo della massima verosimiglianza dobbiamo avere innanzitutto n individui dei quali conosciamo i valori dei predittori e la classe, ovvero dobbiamo avere il training set.
Per ogni $i = 1, ..., n$ indichiamo con y^i il valore di Y relativo all'i-esimo individuo, con $x_1^i, ..., x_m^i$ i suoi valori noti e $p^i = g^{-1}(\eta^i)$, $\eta^i = a_1x_1^i + \cdots + a_mx_m^i + b$. Il numero $(p^i)^{y^i}(1 - p^i)^{1-y^i}$ è la verosimiglianza relativa a quell'individuo; in generale essa è funzione di molte grandezze, ovvero di $x_1^i, ..., x_m^i, a_1, ..., a_m$ e b, ma trattandosi di un individuo del training set, quindi con $x_1^i, ..., x_m^i$ noti, la verosimiglianza è funzione solo di $a_1, ..., a_m$ e b.
Se poi consideriamo tutti gli n individui del training set, la funzione di verosimiglianza del campione totale risulta essere

$$P(Y^1 = y^1, ..., Y^n = y^n) = \prod_{i=1}^{n} (p^i)^{y^i}(1 - p^i)^{1-y^i}$$

dove $p^i = g^{-1}(\eta^i)$, $\eta^i = a_1x_1^i + \cdots + a_mx_m^i + b$.
Essa è funzione di $a_1, ..., a_m$ e b, ovvero con le notazioni delle definizione 3.1 è funzione di $\theta = (a_1, ..., a_m, b)$.
Come già osservato precedentemente, per calcolare i valori di $a_1, ..., a_m, b$ che massimizzano la verosimiglianza, conviene massimizzare il logaritmo della verosimiglianza, cioè

$$\sum_{i=1}^{n} (y^i \ln p^i + (1 - y^i) \ln(1 - p^i)).$$

3.2.2 Esempio: classificazione con regressione logistica

Supponiamo di voler valutare se l'insorgenza di tumori dipende dall'assunzione di alcool e fumo.
Indichiamo le variabili aleatorie con la seguente notazione: Y = presenza di tumore, X_1 = fumatore e X_2 = bevitore.
Il training set, formato da un campione di 100 persone, è riassunto nella seguente tabella:

Fumatore	Bevitore	Tumore	Totale (n. persone)
SI	SI	SI	35
SI	SI	NO	5
SI	NO	SI	8
SI	NO	NO	10
NO	SI	SI	3
NO	SI	NO	12
NO	NO	SI	10
NO	NO	NO	17

La formula del modello di regressione logistica che stiamo cercando è

$$P(Y=1) = \frac{\exp(a_1X_1 + a_2X_2 + b)}{1 + \exp(a_1X_1 + a_2X_2 + b)}$$

ovvero la regola che ci permetta di calcolare la probabilità che Y assuma valore 1 (presenza di tumore), in funzione dei predittori X_1 e X_2.
Utilizzando il software R si calcolano velocemente i coefficienti del modello:

$$\begin{aligned} a_1 &= 1.0774 \\ a_2 &= 1.6472 \\ b &= -1.6088 \end{aligned}$$

Allora

$$P(Y=1) = \frac{\exp(1.08X_1 + 1.65X_2 - 1.61)}{1 + \exp(1.08X_1 + 1.65X_2 - 1.61)}.$$

Ora possiamo calcolare diversi valori:

- La probabilità di avere tumore per una persona che non fuma e non beve è

$$\frac{\exp(-1.61)}{1 + \exp(-1.61)} = 0.17$$

- La probabilità di avere tumore per un fumatore non bevitore è

$$\frac{\exp(-1.61 + 1.08)}{1 + \exp(-1.61 + 1.08)} = 0.37$$

- La probabilità di avere tumore di un bevitore non fumatore è

$$\frac{\exp(-1.61 + 1.65)}{1 + \exp(-1.61 + 1.65)} = 0.51$$

- La probabilità di avere tumore per una persona che fuma e beve è

$$\frac{\exp(-1.61 + 1.08 + 1.65)}{1 + \exp(-1.61 + 1.08 + 1.65)} = 0.75$$

3.3 Punto di vista geometrico della teoria della classificazione

Riprendendo la tabella del training set

	X_1	$\cdots$	X_m	Classe
1	$x_{1,1}$	$\cdots$	$x_{1,m}$	C1
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
n_1	$\cdots$	$\cdots$	$\cdots$	C1
n_1+1	$\cdots$	$\cdots$	$\cdots$	C2
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
n	$x_{n,1}$	$\cdots$	$x_{n,m}$	C2

ogni individuo lo possiamo considerare come un punto nello spazio $\mathbb{R}^m$, come in PCA, solo che ora esso ha un ulteriore attributo: la classe. Abbiamo quindi, in $\mathbb{R}^m$, due nuvole di punti (una formata dai punti della classe C1 e una dai punti della classe C2), sperabilmente un pò separate, non completamente sovrapposte.
Basandoci su queste due nuvole, vogliamo dividere lo spazio $\mathbb{R}^m$ in due regioni A_1 e A_2 (circa uguale a A_1^c) in modo tale da poterle usare per classificare nuovi individui di cui si conoscano solo i valori $x_1, ..., x_m$. In generale, tutti i nuovi individui la cui stringa $(x_1, ..., x_m)$ cadrà in A_1 verranno classificati C1, gli altri C2:

$$(x_1, ..., x_m) \in A_1 \longrightarrow \text{classe C1}$$
$$(x_1, ..., x_m) \in A_2 \longrightarrow \text{classe C2}$$

Formalizzando, in $\mathbb{R}^m$ abbiamo due insiemi di punti relativi a individui del training set: i punti $P_1, ..., P_{n_1}$ degli individui di classe C1 e quelli $P_{n_1+1}, ..., P_n$ degli individui di classe C2.

Come effettuare la suddivisione di $\mathbb{R}^m$ in due regioni? Conviene fare in modo che A_1 contenga tutti i punti $P_1, ..., P_{n_1}$ ed A_2 tutti i punti $P_{n_1+1}, ..., P_n$? Non necessariamente.
Infatti, questa strategia ha diversi difetti:

- Non è univoca in quanto infinite regioni soddisfano la suddetta condizione ed è del tutto arbitrario sceglierne due;
- Non tiene conto del fatto che **le sole variabili $X_1, ..., X_m$ non dovrebbero permettere una classificazione sicura**, quindi deve essere possibile che un individuo di classe C1 stia nella regione A_2 e viceversa;

- Per dividere i punti come detto sopra, è possibile trovarsi nella situazione di dover immaginare regioni molto contorte; se pensiamo che dietro il nostro tentativo di classificazione c'è una realtà fisica, un legame reale tra le variabili $X_1, ..., X_m$ e la classe (a meno di errore e variabili non considerate), è insolito che questo legame passi attraverso complicate formule matematiche (quelle necessarie a descrivere una regione molto contorta): di solito i legami fisici tra grandezze hanno natura polinomiale o comunque abbastanza semplice.

Quindi, si rinuncia al requisito che A_1 contenga tutti i punti $P_1, ..., P_{n_1}$ ed A_2 tutti i punti $P_{n_1+1}, ..., P_n$ ma si vuole comunque che ciò avvenga per la maggior parte dei punti.
Una scelta molto comune è quella di dividere $\mathbb{R}^m$ con un iperpiano, ovvero scegliendo come regioni due semispazi.

3.3.1 Esempio: suddivisione tramite regressione lineare multipla

Consideriamo sempre la tabella del training set, in una forma leggermente diversa

	X_1	$\cdots$	X_m	Y
1	$x_{1,1}$	$\cdots$	$x_{1,m}$	-1
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
n_1	$\cdots$	$\cdots$	$\cdots$	-1
$n_1 + 1$	$\cdots$	$\cdots$	$\cdots$	1
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
n	$x_{n,1}$	$\cdots$	$x_{n,m}$	1

e applichiamo la regressione lineare classica, cioè cerchiamo di spiegare il valore ± 1 (che è una semplice codifica della classe) tramite i predittori $X_1, ..., X_m$. Supponiamo di aver trovato i coefficienti $\hat{a}_1, ..., \hat{a}_m, \hat{b}$ del modello regressivo. Se ora esaminiamo un nuovo individuo, di cui si conoscono i valori $x_1, ..., x_m$, possiamo calcolare il suo valore y:

$$y = \hat{a}_1 x_1 + \cdots + \hat{a}_p x_m + \hat{b}.$$

Ovviamente questo valore non sarà uguale a ± 1, però adottiamo la regola: se $y < 0$ allora l'individuo è di classe C1; se invece $y > 0$ allora è di classe C2.
Questo metodo è quasi uguale alla regressione logistica: anche in essa si calcola, per un nuovo individuo, il predittore lineare

$$\eta = \hat{a}_1 x_1 + \cdots + \hat{a}_m x_m + \hat{b}$$

con cui poi si calcola $p = g^{-1}(\eta)$ e si classifica l'individuo in base a se $p < \frac{1}{2}$ o $p > \frac{1}{2}$; ma questo corrisponde esattamente a dire che lo si classifica C1 se $\eta < 0$ (se g è la funzione logit) e C2 se $\eta > 0$.
Come si interpreta questo metodo in termini di regioni A_1 e A_2? Semplicemente:

$$A_1 = \left\{ x \in \mathbb{R}^m : \hat{a}_1 x_1 + \cdots + \hat{a}_m x_m + \hat{b} < 0 \right\}$$

$$A_2 = \left\{ x \in \mathbb{R}^m : \hat{a}_1 x_1 + \cdots + \hat{a}_m x_m + \hat{b} > 0 \right\}.$$

A_1 e A_2 sono quindi due semispazi e la separazione è fatta tramite l'iperpiano $\hat{a}_1 x_1 + \cdots + \hat{a}_m x_m + \hat{b} = 0$.
Concludiamo con un esempio concreto. Supponiamo di voler capire, in base al peso e all'altezza di una persona, se questa è un maschio o una femmina; consideriamo la seguente tabella del training set:

Peso (Kg)	Altezza (cm)	Sesso
50	158	F
62	150	F
49	165	F
45	163	F
58	168	F
54	175	F
70	175	F
53	164	F
61	160	F
77	178	F
80	180	M
71	175	M
81	179	M
85	181	M
87	185	M
89	190	M
93	185	M
75	175	M
67	168	M

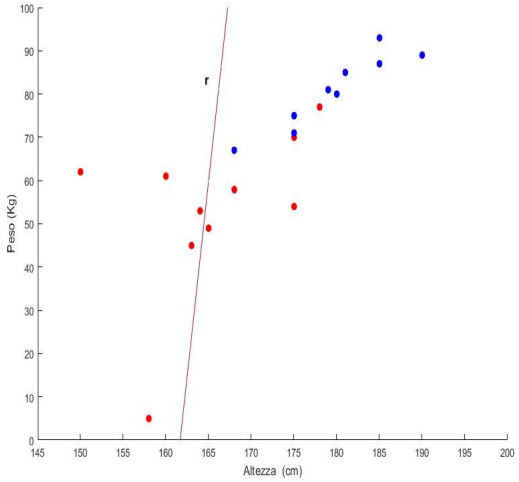
Sempre utilizzando il software R si trovano i coefficienti del modello logistico:

$$a_1 = 0.002$$
$$a_2 = 0.036$$
$$b = -5.82$$

Quindi la retta che divide il piano nelle due regioni cercate è

$$r : 0.002x_1 + 0.036x_2 - 5.82 = 0$$

dove x_1 indica il peso e x_2 l'altezza.
Graficamente risulta una situazione del tipo seguente, dove i punti in rosso indicano le femmine e quelli in blu i maschi:



Bibliografia

[1] Martire Fabrizio - La regressione logistica e i modelli log-lineari nella ricerca sociale, **FrancoAngeli**.

[2] Flandoli Franco - Dispense di Statistica II, A.A. 2013/2014, Scuola Normale Superiore di Pisa.

-

Ringraziamenti

In primis vorrei ringraziare il prof. Pascucci per avermi aiutato nella realizzazione di questa tesi, per essere stato presente e disponibile, per avermi guidato e consigliato tenendo sempre in considerazione le mie idee.

Un enorme grazie va alla mia famiglia, che ha saputo sostenermi in questi anni, soprattutto nei momenti più difficili nei quali cominciavo a pensare che questo giorno non sarebbe mai arrivato.

Ringrazio le mie amiche di sempre, Sofia, Sara e Stella, per essermi state vicine e avermi incoraggiato a non mollare mai.

Un grazie speciale lo dedico a Morgan. Tutti dicono che mi "faranno santa" per la pazienza che devo avere con te, ma in verità solo io so la pazienza che hai dimostrato in questi anni di università: tra pianti, crisi isteriche e mille dubbi hai sempre trovato le parole giuste per farmi superare qualsiasi ostacolo.

Infine, un pensiero va anche alla mia "famiglia universitaria": Martina, Daniele, Simone e tutti gli altri amici che mi hanno accompagnato in questo percorso; un percorso si difficile ma che grazie a voi è stato straordinario e indimenticabile, ricco di giornate intensissime di studio e ansie pre-esame ma anche di risate e grandi soddisfazioni.

Questo mio traguardo è anche un pò vostro, GRAZIE.