Alma Mater Studiorum · Università di Bologna

scuola di scienze

Corso di Laurea in Matematica

APPRENDIMENTO AUTOMATICO CON METODO KERNEL

—

Tesi di Laurea in
Analisi Matematica

Relatore:
Chiar.mo Prof.
NICOLA ARCOZZI

Presentata da:
ALESSANDRA TULLO

VI Sessione
Anno Accademico 2019/2020

"Insegnami la dolcezza ispirandomi la carità,
insegnami la disciplina dandomi la pazienza
e insegnami la scienza illuminandomi la mente."
-Sant'Agostino

Indice

Introduzione

L'apprendimento automatico è una branca dell'informatica che può essere considerata parente stretta dell'intelligenza artificiale. Quando si parla di apprendimento automatico si ha a che fare con differenti meccanismi che permettono ad una macchina intelligente di migliorare le proprie capacità e prestazioni nel tempo. Esistono diverse modalità di apprendimento automatico e queste sono divise in apprendimento supervisionato, non supervisionato e per rinforzo. La differenza tra le tre è data dagli algoritmi utilizzati e dallo scopo per cui le macchine dell'intelligenza artificiale vengono realizzate.

Molti problemi nelle applicazioni del mondo reale dell'apprendimento automatico possono essere formalizzati come problemi statistici classici: problemi di riconoscimento di modelli, di regressione o di riduzione delle dimensioni.

I metodi kernel sono una classe di algoritmi adatti a risolvere tali problemi. Infatti estendono l'applicabilità di molti metodi statistici inizialmente progettati per i vettori praticamente a qualsiasi tipo di dati, senza la necessità di una vettorizzazione esplicita degli stessi. Per estendere questi algoritmi ai non vettori bisogna definire un particolare tipo di funzioni, dette funzioni kernel definite positive.

Il presente lavoro di tesi parte dalla volontà di approfondire lo studio degli spazi di Hilbert a nucleo riproducente e le loro proprietà principali. Tali spazi sono gli spazi di funzioni in cui si lavora con le funzioni kernel definite positive. La definizione di spazio di Hilbert a nucleo riproducente è data in più modi, tutti tra loro equivalenti, i quali forniscono diversi punti di vista sull'argomento. Vengono mostrati alcuni esempi di funzioni kernel definite positive: lineare, polinomiale e gaussiano. Per ogni tipo di funzione kernel è possibile costruire lo spazio di Hilbert corrispondente a quest'ultima. Viene mostrato il procedimento per la costruzione di tale spazio nel caso particolare di kernel polinomiale di grado 2.

Si procede poi ad analizzare i metodi kernel. Tali metodi approcciano al problema dell'apprendimento automatico mappando i dati del problema in uno spazio "comodo", detto spazio delle caratteristiche (future space). In esso i dati vengono trasformati in vettori dello spazio euclideo multidimensionale e ogni coordinata corrisponde a una caratteristica degli stessi. Vengono analizzati nel dettaglio il kernel trick e il representer theorem. Il primo è una proposizione matematicamente banale ma con risvolti pratici molto interessanti: utilizza i kernel definiti positivi come prodotti interni e per que-

sto rende lo spazio delle caratteristiche dotato di proprietà geometriche molto comode. Il secondo è un teorema che dà una soluzione concreta al problema di ottimizzazione attraverso le proprietà dello spazio di Hilbert a nucleo riproducente.

A conclusione di questo lavoro di tesi si esaminano i metodi kernel applicandoli in particolare all'apprendimento supervisionato. Tali problemi possono essere di regressione o di classificazione. Viene altrsì analizzato nel dettaglio il problema della regressione lineare del kernel calcolandone la soluzione nello spazio di Hilbert a nucleo riproducente utilizzando il representer theorem e studiandone l'unicità.

I metodi kernel vedono alcune possibili applicazioni nella geostatistica, nella ricostruzione 3D, nella chemioinformatica e nella bioinformatica.

Ci sono molte possibilità di futuro sviluppo dell'apprendimento automatico perchè molti settori potrebbero avvantaggiarsi dall'uso di macchine in grado di fare scelte intelligenti. Probabilmente l'unico fattore limitante al pieno utilizzo di strumenti in grado di imparare da soli è il timore dell'uomo che le macchine possano diventare troppo intelligenti, togliendogli la facoltà di scelta e libertà. Un timore che, come afferma il professor Pedro Domingos, esperto di apprendimento automatico e data mining, non esiste visto che "la gente ha paura che i computer diventino troppo intelligenti e dominino il mondo, ma il vero problema è che pur essendo ancora troppo stupidi lo hanno già conquistato".

Capitolo 1

Spazi di Hilbert e spazi di Hilbert a nucleo riproducente

1.1 Prime definizioni e proprietá

Definizione 1.1. Sia V uno spazio vettoriale su $\mathbb{R}$ o su $\mathbb{C}$ sul quale è definito un prodotto scalare $<,>$. Lo spazio $H := (V, <, >)$ si definisce spazio di Hilbert se la distanza d indotta dal prodotto scalare rende (V, d) uno spazio metrico completo.

Definizione 1.2. Sia V un $\mathbb{C}$-spazio vettoriale. Lo spazio duale di V, indicato con V^*, è formato da tutti gli operatori lineari $f : V \to \mathbb{C}$. L'insieme V^* assume la struttura algebrica di spazio vettoriale con le operazioni somma $(f + g)(w) := f(w) + g(w)$ e prodotto per scalari $(\alpha f)(w) := \alpha f(w)$, con $f, g \in V^*$ e $\alpha \in \mathbb{C}$.

Per gli spazi di Hilbert vale il Teorema di rappresentazione di Riesz:

Teorema 1.1. Sia H uno spazio di Hilbert e T un operatore lineare continuo su H, cioè $T \in H^*$. Allora esiste un unico $x \in H$ che rappresenta T; nel senso che $Ty =< y, x >$ $\forall y \in H$. Inoltre $\|T\|_{H^*} = \|x\|_H$. Si dice perciò che il duale H^* di uno spazio di Hilbert H si può identificare con H stesso.

Definizione 1.3. Sia H_X uno spazio di Hilbert di funzioni $f : X \to \mathbb{C}$, con X insieme qualsiasi. Diciamo che H_X ha valutazione nei punti limitata (Bounded Point Evaluation: BPE) se $\forall x \in X$ la funzione di valutazione

$$\begin{aligned} \eta_x \colon H_X &\to \mathbb{C} \\ f &\mapsto f(x) \end{aligned}$$

é limitata.

1.2 Spazi di Hilbert a nucleo riproducente

Una prima motivazione per introdurre gli spazi di Hilbert a nucleo riproducente è che in molti spazi di funzioni il valore di una singola funzione in un singolo punto non può essere calcolato, perlomeno non in modo quantitativo. Ovviamente ci sono anche altri motivi per studiare gli spazi di Hilbert a nucleo riproducente ed è per questo che vi sono diverse definizioni equivalenti di tali spazi. Queste definizioni forniscono punti di vista diversi sull'argomento.

Definizione 1.4. Sia H_X uno spazio di Hilbert di funzioni $f : X \to \mathbb{R}$. H_X è uno spazio di Hilbert a nucleo riproducente se $\forall x \in X$ esiste $K_x \in H_X$ funzione riproducente, cioé tale che vale la seguente proprietá: $f(x) =< f, K_x > \ \forall f \in H_X$.

Teorema 1.2. Sia H_X uno spazio di Hilbert di funzioni su X. H_X è uno spazio di Hilbert a nucleo riproducente se e solo se $\forall x \in X$ la funzione di valutazione in x

$$\begin{aligned} \eta_x \colon H_X &\to \mathbb{C} \\ f &\mapsto f(x) \end{aligned}$$

é un funzionale limitato su H_X.

Dimostrazione. Dato che H_X è uno spazio di Hilbert a nucleo rirpoducente, allora $f(x) =< f, K_x > \ \forall f \in H_X$, quindi

$$|f(x)| = |< f, K_x >| \leq \|f\| \, \|K_x\|$$

per la disuguaglianza di Cauchy-Schwarz. Allora

$$\begin{aligned} \eta_x \colon H_X &\to \mathbb{C} \\ f &\mapsto f(x) \end{aligned}$$

che è il funzionale di valutazione in x, è limitato su H_X. Viceversa, dato che il funzionale di valutazione η_X è limitato, allora vale il teorema di Riesz-Fischer (1.1) e quindi $\eta_X(f) =< f, K_x >$ per un dato K_x. □

Teorema 1.3. Sia H uno spazio di Hilbert di funzioni $f : X \to \mathbb{C}$. H è uno spazio di Hilbert a nucleo riproducente se e solo se è tale che esiste una famiglia $\{K_x\}_{x \in X}$ di funzioni $K_x : X \to \mathbb{C}$ con la proprietà che $span\,\{K_x\}$ è denso in H.

Dimostrazione. $Span\,\{K_x; x \in X\}$ è denso in H perchè $\forall \varphi \in H$, φ è ortogonale a $K_x \ \ \forall x \in X \Leftrightarrow 0 =< \varphi, K_x >= \varphi(x) \ \forall x \in X \Leftrightarrow \varphi = 0$ punto per punto, cioè $\varphi = 0$ in H. Viceversa dato che $span\,\{K_x\}$ è denso in H, per il teorema delle proiezioni avremo che $\forall f \in H$ si ha $f(x) =< f, K_x >$. □

Proposizione 1.1. Sia $K : X \times X \to \mathbb{C}$ un prodotto hermitiano semidefinito positivo. $\forall x \in X$ poniamo $K_x(y) = K(y,x)$, $K_x : X \to \mathbb{C}$. Su $H_0 = span\{K_x; x \in X\}$ definiamo la forma bilineare $< \sum_{i=1}^m a_i K_{x_i}, \sum_{j=1}^n b_j K_{x_j} >= \sum_{i=1}^m \sum_{j=1}^n a_i \overline{b_j} K(y_j, x_i)$. Allora $<,>$ definisce un prodotto interno su H_0. Il completamento H di H_0 può essere identificato con uno spazio di Hilbert a nucleo riproducente su X che ha nucleo $\{K_x\}_{x\in X}$.

Dimostrazione. L'unica proprietà del prodotto interno che richiede sforzo per essere dimostrata è la seguente: se $h = \sum_{i=1}^n c_i K_{x_i}$ e $\|h\| = 0$ allora h è la funzione nulla:

$$0 = \sum_{i,j=1}^n c_i \overline{c_j} K(x_j, x_i) \Rightarrow \sum_{i,j=1}^n c_i K(y, x_i) = 0 \ \forall y \in X. \tag{1.1}$$

Supponiamo $x_{n+1} = y$ e $c_{n+1} = 0$ vediamo (sostituendo $n+1$ con n) che è sufficiente mostrare l'implicazione per $y = x_n$. Cioè: la matrice $\mathbf{K} := [K(x_j, x_i)]_{i,j=1}^n$ è hermitiana e definita positiva, quindi può essere scritta come $\mathbf{K} = \mathbf{R}^* \begin{bmatrix} 0 & 0 \\ 0 & \Delta \end{bmatrix} \mathbf{R}$ con $\mathbf{R} \in SU(n)$, dove $SU(n)$ è il gruppo unitario speciale e Δ è una matrice diagonale con diagonale positiva $\Delta = diag(\lambda_1, ..., \lambda_m)$. Scrivendo il vettore $d := \mathbf{R}^* c = \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} \in \mathbb{C}^{(n-m)+m}$ abbiamo che

$$0 = c^* \mathbf{K} c = c^* \mathbf{R}^* \begin{bmatrix} 0 & 0 \\ 0 & \Delta \end{bmatrix} \mathbf{R} c = d^* \begin{bmatrix} 0 & 0 \\ 0 & \Delta \end{bmatrix} d = d_2^* \Delta d_2 \tag{1.2}$$

questo implica $d_2 = 0$. Quindi $\mathbf{K}c = \mathbf{R}^* \begin{bmatrix} 0 & 0 \\ 0 & \Delta \end{bmatrix} \begin{pmatrix} d_1 \\ 0 \end{pmatrix} = 0$. In particolare, con $\delta_n(j) = 0$ per $j \neq n$ e $\delta_n(n) = 1$ si ha che $\sum_{i,j=1}^n c_i \mathbf{K}(x_n, x_i) = \delta_n^* \mathbf{K} c = 0$. □

Teorema 1.4. Definire uno spazio di Hilbert a nucleo riproducente equivale a dare una funzione $K : X \times X \to \mathbb{C}$ semidefinita positiva.

Dimostrazione. Definisco $K(x,y) =< K_y, K_x >$. In questo modo K è definita positiva perchè $\sum_{i,j} a_i \overline{a_j} K(x_j, x_i) = \sum_{i,j} a_i \overline{a_j} < K_{x_j}, K_{x_i} > =< \sum_i a_i K_{x_i}, \sum_j \overline{a_j} K_{x_j} >= \|\sum_i a_i K_{x_i}\|^2 \geq 0$. Il viceversa è la proposizione precedente (1.1). □

Teorema 1.5. Sia H uno spazio di Hilbert di funzioni su X. H è uno spazio di Hilbert a nucleo riproducente se e solo se $\forall \{e_n\}_{n=1}^\infty$ base ortonormale di H si ha che $\sum_{n=1}^\infty \overline{e_n(x)} e_n(\cdot)$ converge in H.

Dimostrazione. Se H è uno spazio di Hilbert a nucleo riproducente con kernel $\{K_x\}_{x\in X}$ e scriviamo K_y rispetto alla base ortonormale: $K_y = \sum_n \hat{K}_y(n) e_n$ dove $\hat{K}_y(n) =< K_y, e_n >= \overline{< e_n, K_y >} = \overline{e_n(y)}$ dalla proprietà riproducente di K_y. Quindi $\sum_n \overline{e_n(y)} e_n = \sum_n \hat{K}_y(n) e_n = K_y$ converge. Viceversa presa $f \in H$ e definita $K_x = \sum_{n=1}^\infty \overline{e_n(x)} e_n$,

si ha che $< f, K_x(\cdot) >=< f, \sum_{n=1}^{\infty} e_n(x)\, e_n > \underbrace{=}_{perlaconvergenza} \sum_{n=1}^{\infty} e_n(x) < f, e_n >=$ $[\sum_{n=1}^{\infty} < f, e_n > e_n](x) = f(x)$ perchè $\sum_{n=1}^{\infty} < f, e_n > e_n$ esprime f in un dato sistema ortonormale. □

Osservazioni

- $K(x,y) = \sum_{n=1}^{\infty} e_n(x)\, e_n(y)$ tipicamente in uno spazio di Hilbert qualsiasi questa diverge, ma se X è finito dimensionale allora converge.
- Se uno spazio di Hilbert a nucleo riproducente ha dimensione finita N, allora ogni famiglia di vettori di funzioni riproducenti $\{K_{x_1}, ..., K_{x_M}\}$ con $M > N$ sono linearmente dipendenti e relazioni come (1.1) sono comuni.

Vediamo ora alcune proprietà di base degli spazi di Hilbert a nucleo riproducente

Proposizione 1.2. 1. $K_x(y) =< K_x, K_y >$;

2. $\overline{K_x(y)} = K_y(x)$;
3. $K_x(x) = \|K_x\|_H^2$;
4. $K_x = 0 \Leftrightarrow \forall f \in H$ si ha $f(x) = 0$.

Dimostrazione. 1. $K_x(y) = K(x,y) =< K_x, K_y >_H$;

2. $\overline{K_x(y)} = \overline{K(x,y)} \underbrace{=}_{Khermitiana} K(y,x) = K_y(x)$;
3. $K_x(x) = K(x,x) =< K_x, K_x >_H = \|K_x\|_H^2$;
4. Se $K_x = 0 \Rightarrow f(x) =< f, K_x >= 0\ \forall f \in H$.
   Se $f(x) = 0\ \forall f \in X \Rightarrow < f, K_x >= f(x) = 0 \Rightarrow < f, K_x >= 0\ \forall f \in H \Rightarrow K_x = 0$.

□

Vediamo un esempio semplice, ma non banale, di spazio di Hilbert a nucleo riproducente.

Esempio 1.1. Sia $X = [0,1]$ e sia $H = \{p : [0,1] \to \mathbb{R}; deg(p) \leq 2\}$ con $\|p\|_H = \sqrt{\int_0^1 p(x)^2 dx}$. Esiste $K_x \in H$ tale che $p(x) =< p, K_x >= \int_0^1 p(y) K_x(y) dy$?

La risposta é sí: se $p(x) = ay^2 + by + c$ e se $K_x(y) = Ay^2 + By + C$ con a, b, c, A, B, C che variano, si ha che:

$$\begin{aligned} ax^2 + bx + c &= \int_0^1 (ay^2 + by + c)(Ay^2 + By + C)dy = \\ &= \int_0^1 \left[aAy^4 + (aB + bA)y^3 + (aC + bB + cA)y^2 + (cB + bC)y + cC\right] dy = \\ &= \frac{aA}{5} + \frac{aB + bA}{4} + \frac{aC + bB + cA}{3} + \frac{cB + bC}{2} + cC = \\ &= a\left(\frac{A}{5} + \frac{B}{4} + \frac{C}{3}\right) + b\left(\frac{A}{4} + \frac{B}{3} + \frac{C}{2}\right) + c\left(\frac{A}{3} + \frac{B}{2} + C\right) \end{aligned}$$

$$\Rightarrow \begin{cases} \frac{A}{5} + \frac{B}{4} + \frac{C}{3} = x^2 \\ \frac{A}{4} + \frac{B}{3} + \frac{C}{2} = x \\ \frac{A}{3} + \frac{B}{2} + C = 1 \end{cases}$$

Risolvendo il sistema otteniamo che

$$K_x(y) = \left(180x^2 - 180x + 30\right) y^2 + \left(-180x^2 + 192x - 36\right) y + 30x^2 - 36x + 9.$$

1.3 Esempi di funzioni kernel definite positive

Proponiamo ora esempi di funzioni kernel definite positive. Tra le più note ci sono le funzioni kernel lineare, polinomiale e gaussiano. Analizziamole nel dettaglio:

Esempio 1.2 (Kernel lineare)**.**

Definizione 1.5. Sia $X \subseteq \mathbb{C}^d$ un insieme. La funzione $K : X \times X \to \mathbb{C}$ data da $\forall (x, x') \in X \times X, K(x, x') =< x, x' >_{\mathbb{C}^d}$ è definita positiva. Tale K si definisce kernel lineare.

Osservazione Ricordando la definizione di kernel definito positivo, dimostriamo che il kernel lineare che abbiamo appena definito è davvero definito positivo:

1. K è hermitiana: $K(x, x') =< x, x' >_{\mathbb{C}^d} = \overline{x'^T}x = \overline{< x', x >_{\mathbb{C}^d}} = \overline{K(x', x)}$;
2. dato $(x_1, ..., x_N) \in X^N$ e $(a_1, ..., a_N) \in \mathbb{C}^N$ si ha:
   $\sum_{i=1}^N \sum_{j=1}^N a_i \overline{a_j} < x_i, x_j >_{\mathbb{C}^d} = \left\| \sum_{i=1}^N a_i x_i \right\|^2 \geq 0$

Per le funzioni kernel lineari vale il seguente

Teorema 1.6. Lo spazio di Hilbert a nucleo riproducente delle funzioni kernel lineari H è l'insieme delle funzioni della forma $f_w(x) =< w, x >_{\mathbb{C}^d}$ per un certo $w \in \mathbb{C}^d$ fissato. Tale insieme è dotato del seguente prodotto interno: $\forall w, v \in \mathbb{C}^d, < f_w, f_v >_H =< w, v >_{\mathbb{C}^d}$ e della corrispondente norma: $\forall w \in \mathbb{C}^d, \|f_w\|_H = \|w\|_{\mathbb{C}^d}$.

Dimostrazione. Lo spazio H delle funzioni descritte dal teorema è il duale di $\mathbb{C}^d$, quindi è uno spazio di Hilbert: $H = \left\{ f_w(x) =< w, x >_{\mathbb{C}^d}; w \in \mathbb{C}^d \right\}$. H contiene tutte le funzioni della forma $k_w : x \mapsto < w, x >_{\mathbb{C}^d}$ e $\forall x \in \mathbb{C}^d$ e $\forall f_w \in H$ si ha che $f_w(x) =< w, x >_{\mathbb{C}^d} =< f_w, k_x >_H$. Quindi H è lo spazio di Hilbert a nucleo riproducente delle funzioni kernel lineari. □

Esempio 1.3 (Kernel polinomiale)**.**

Definizione 1.6. Sia $X \subseteq \mathbb{C}^N$ un insieme e sia $\phi : X \to \mathbb{C}^d$ una funzione definita mediante un polinomio. La funzione $K : X \times X \to \mathbb{C}$ tale che $\forall (x, x') \in X \times X$ si ha $K(x, x') =< \phi(x), \phi(x') >_{\mathbb{C}^d}$ è definita positiva. Tale K si definisce kernel polinomiale.

Osservazione Dimostriamo che $K(x, x')$ definita come sopra è definita positiva:

1. è hermitiana perchè è un prodotto kermitiano;
2. $\forall (a_1, ..., a_N) \in \mathbb{C}^N, \forall (x_1, ..., x_N) \in X^N$ si ha:
   $\sum_{i=1}^N \sum_{j=1}^N a_i \overline{a_j} < \phi(x_i), \phi(x_j) >_{\mathbb{C}^d} = \left\| \sum_{i=1}^N a_i \phi(x_i) \right\|^2 \geq 0.$

Vediamo un esempio di ϕ che dà un kernel polinomiale:

Esempio 1.4. Se $X = \mathbb{C}^2$ e $\phi : \mathbb{C}^2 \to \mathbb{C}^3$ è tale che $\phi(x_1, x_2) = (x_1^2, \sqrt{2} x_1 x_2, x_2^2) = (y_1, y_2, y_3)$. Chi è $K(x, x')$?

$$\begin{aligned} K(x, x') &=< \phi(x), \phi(x') >_{\mathbb{C}^3} = \\ &= y_1 \overline{y_1'}^T + y_2 \overline{y_2'}^T + y_3 \overline{y_3'}^T = \\ &= x_1^2 (\overline{x_1'}^T)^2 + 2 x_1 x_2 \overline{x_1'}^T \overline{x_2'}^T + x_2^2 (\overline{x_2'}^T)^2 = \\ &= (x_1 \overline{x_1'}^T + x_2 \overline{x_2'}^T)^2 = \\ &=< x, x' >_{\mathbb{C}^2}^2 \geq 0. \end{aligned}$$

Graficamente avremo che il cerchio $x_1^2 + x_2^2 \leq 1$ diventa il semispazio $y_1 + y_3 \leq 1$.

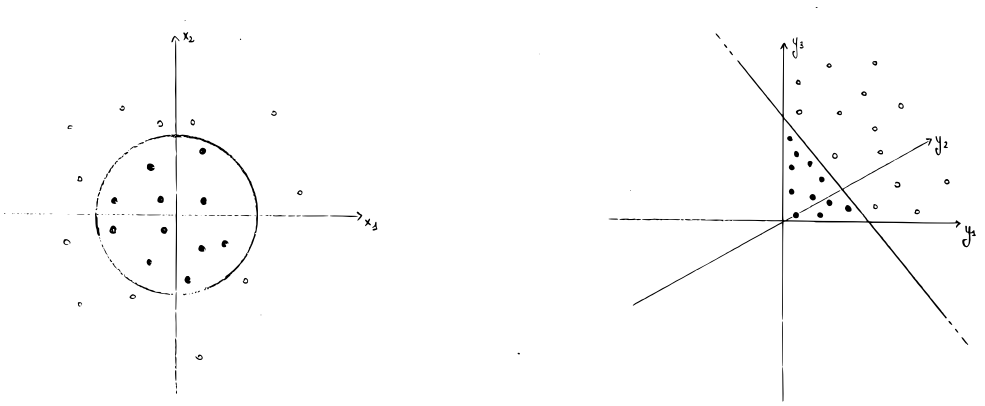

Figura 1.1: Esempio 1.4

Introduciamo ora un lemma di algebra lineare che useremo di seguito.

Lemma 1.1. Siano $x = (x_1, ..., x_n)$ e $y = (y_1, ..., y_n)$ due vettori a coordinate in $\mathbb{C}$. Allora $tr(\overline{x^T} y \overline{y^T} x) = tr(x \overline{x^T} y \overline{y^T})$.

Dimostrazione. Con un conto esplicito vediamo che

$$\begin{aligned} tr(\overline{x^T} y \overline{y^T} x) &= tr\left((\overline{x_1}, ..., \overline{x_n}) \begin{pmatrix} y_1 \\ ... \\ y_n \end{pmatrix} (\overline{y_1}, ..., \overline{y_n}) \begin{pmatrix} x_1 \\ ... \\ x_n \end{pmatrix} \right) = \\ &= tr((\overline{x_1} y_1 + \overline{x_2} y_2 + ... + \overline{x_n} y_n)(\overline{y_1} x_1 + \overline{y_2} x_2 + ... + \overline{y_n} x_n)) = \\ &= (\overline{x_1} y_1 + \overline{x_2} y_2 + ... + \overline{x_n} y_n)(\overline{y_1} x_1 + \overline{y_2} x_2 + ... + \overline{y_n} x_n). \end{aligned}$$

Allo stesso modo si vede che

$$\begin{aligned} tr(x \overline{x^T} y \overline{y^T}) &= tr\left(\begin{pmatrix} x_1 \\ ... \\ x_n \end{pmatrix} (\overline{x_1}, ..., \overline{x_n}) \begin{pmatrix} y_1 \\ ... \\ y_n \end{pmatrix} (\overline{y_1}, ..., \overline{y_n}) \right) = \\ &= tr\left(\begin{pmatrix} x_1\overline{x_1} & x_1\overline{x_2} & \cdots & x_1\overline{x_n} \\ x_2\overline{x_1} & x_2\overline{x_2} & \cdots & x_2\overline{x_n} \\ \vdots & \vdots & \ddots & \vdots \\ x_n\overline{x_1} & x_n\overline{x_2} & \cdots & x_n\overline{x_n} \end{pmatrix} \begin{pmatrix} y_1\overline{y_1} & y_1\overline{y_2} & \cdots & y_1\overline{y_n} \\ y_2\overline{y_1} & y_2\overline{y_2} & \cdots & y_2\overline{y_n} \\ \vdots & \vdots & \ddots & \vdots \\ y_n\overline{y_1} & y_n\overline{y_2} & \cdots & y_n\overline{y_n} \end{pmatrix} \right) = \\ &= tr\begin{pmatrix} x_1\overline{x_1}y_1\overline{y_1} + \ldots + x_1\overline{x_n}y_n\overline{y_1} & \cdots & \cdots \\ \vdots & \ddots & \vdots \\ \cdots & \cdots & x_n\overline{x_1}y_1\overline{y_n} + \ldots + x_n\overline{x_n}y_n\overline{y_n} \end{pmatrix} = \\ &= x_1\overline{y_1}(\overline{x_1}y_1 + \ldots + \overline{x_n}y_n) + \ldots + x_n\overline{y_n}(\overline{x_1}y_1 + \ldots + \overline{x_n}y_n) = \\ &= (x_1\overline{y_1} + \ldots + x_n\overline{y_n})(\overline{x_1}y_1 + \ldots + \overline{x_n}y_n). \end{aligned}$$

□

Per le funzioni kernel polinomiali di grado due $K(x, y) =< x, y >^2_{\mathbb{C}^d} \forall x, y \in \mathbb{C}^d$ vale il seguente procedimento per determinarne lo spazio di Hilbert a nucleo riproducente:

1. Per prima cosa cerchiamo un prodotto interno:

$$\begin{aligned} K(x, y) &=< x, y >_{\mathbb{C}^d} = (\overline{y}^T x)^2 = \\ &= (\overline{y}^T x \overline{y}^T x) = tr(\overline{y}^T x \overline{y}^T x) \underbrace{=}_{lemma} \\ &\underbrace{=}_{lemma} tr(\overline{x}^T y \overline{y}^T x) = tr(x \overline{x}^T y \overline{y}^T) = \\ &=< x\overline{x}^T, y\overline{y}^T >_F, \end{aligned}$$

   dove l'ultima uguaglianza è data dalla definizione della norma di Frobenius, che è una norma matriciale: $< x\overline{x}^T, y\overline{y}^T >_F := tr(x\overline{x}^T y\overline{y}^T)$.

2. A questo punto proponiamo un candidato spazio di Hilbert a nucleo riproducente H. Sappiamo che H contiene tutte le funzioni della forma $f(x) = \sum_{i=1}^{n} a_i K(x_i, x) = \sum_{i=1}^{n} a_i < x_i\overline{x_i}^T, x\overline{x}^T >_F =< \sum_{i=1}^{n} a_i x_i\overline{x_i}^T, x\overline{x}^T >_F$. Dato che ogni matrice hermitiana in $\mathbb{C}^{n\times n}$ può essere scritta con la decomposizione della forma $\sum_{i=1}^{n} a_i x_i\overline{x_i}^T$; il nostro spazio di Hilbert a nucleo riproducente candidato, H, sarà l'insieme delle funzioni quadratiche $f_{\mathbf{S}}(x) =< \mathbf{S}, x\overline{x}^T >_F = \overline{x}^T\mathbf{S}x$ per $\mathbf{S} \in S^{n\times n}$, dove $S^{n\times n}$ è l'insieme delle matrici hermitiane in $\mathbb{C}^{n\times n}$. H è dotato del prodotto interno $< f_{\mathbf{S}_1}, f_{\mathbf{S}_2} >_H =< \mathbf{S}_1, \mathbf{S}_2 >_F$.

3. Ora controlliamo che il candidato spazio di Hilbert a nucleo riproducente sia davvero uno spazio di Hilbert. In questo caso particolare è banale da verificare perchè si vede facilmente che H è uno spazio euclideo isomorfo a $S^{n\times n}$ tramite l'isomorfismo $\phi : \mathbf{S} \mapsto f_{\mathbf{S}}$.

4. Infine controlliamo che H sia lo spazio di Hilbert a nucleo riproducente. H contiene tutte le funzioni $K_x : t \mapsto K(x,t) =< x\overline{x}^T, t\overline{t}^T >_F$ e inoltre per ogni $f_{\mathbf{S}} \in H$ e $\forall x \in X$ si ha $f_{\mathbf{S}}(x) =< \mathbf{S}, x\overline{x}^T >_F =< f_{\mathbf{S}}, f_{x\overline{x}^T} >=< f_{\mathbf{S}}, K_x >_H$.

Da questo esempio abbiamo ottenuto i passi da seguire per determinare lo spazio di Hilbert a nucleo riproducente di una funzione kernel polinomiale qualunque.

L'esempio di ϕ polinomiale visto era di grado 2. Se volessimo determinare la funzione kernel polinomiale K per un polinomio di grado $p \geq 2$, dovremmo procedere nel seguente modo: se $p = 3$, ad esempio, si ha $\phi : \mathbb{C}^2 \to \mathbb{C}^4$ data da $\phi(x_1, x_2) = (x_1^3, \sqrt{3}x_1^2x_2, \sqrt{3}x_1x_2^2, x_2^3)$. Allora

$$\begin{aligned} K(x,x') &=< \phi(x), \phi(x') >= \\ &= x_1^3((\overline{x_1'})^T)^3 + 3x_1^2((\overline{x_1'})^T)^2x_2((\overline{x_2'})^T) + 3x_1((\overline{x_1'})^T)x_2((\overline{x_2'})^T)^2 + x_2^3((\overline{x_2'})^T)^3 = \\ &= (x_1(\overline{x_1'})^T + x_2(\overline{x_2'})^T)^3 = \\ &=< x, x' >^3_{\mathbb{C}^2} \geq 0. \end{aligned}$$

In generale se ϕ è un polinomio qualsiasi di grado $p \geq 1$, la generica funzione polinomiale K riultante sarà del tipo $K(x,x') = (a\overline{x}^Tx' + b)^p$ con $a, b \in \mathbb{C}$ e $p \geq 1$.

Esempio 1.5 (Kernel gaussiano)**.**

Definizione 1.7. Sia X un insieme. La funzione $K : X \times X \to \mathbb{C}$ data da $K(x,x') = exp(-\frac{1}{2\sigma^2}\|x - x'\|^2)$ è definita positiva e si definisce kernel gaussiano.

Osservazione Dobbiamo verificare che anche questa K è definita positiva:

1. Per dimostrare che K è hermitiana osserviamo prima di tutto che $\|x - x'\|^2 = < x - x', x - x' >_{\mathbb{C}^d} = \overline{< x - x', x - x' >_{\mathbb{C}^d}}$. Quindi

$$\begin{aligned} K(x, x') &= exp(-\frac{1}{2\sigma^2} \|x - x'\|^2) = \\ &= exp(-\frac{1}{2\sigma^2} < \overline{x - x', x - x'} >_{\mathbb{C}^d}) = \overline{K(x', x)}; \end{aligned}$$

2. inoltre, per la funzione kernel gaussiana vale il seguente teorema che ci assicura che sia definita positiva.

   Teorema 1.7 (Baby-Bochner)**.** Sia $\mu \geq 0$ una misura su $\mathbb{R}$ e sia $\hat{\mu} = \int_{\mathbb{R}} e^{-ixy} d\mu(x)$ la sua trasformata di Fourier. Allora $\hat{\mu}$ è definita positiva.

   Dimostrazione.

$$\begin{aligned} \sum_{j,l=1}^{n} c_j \overline{c_l} \hat{\mu}(y_j - y_l) &= \sum_{j,l=1}^{n} c_j \overline{c_l} \int_{\mathbb{R}} e^{-iy_j x} e^{-iy_l x} d\mu(x) = \\ &= \int_{\mathbb{R}} d\mu(x) \sum_{j,l=1}^{n} c_j e^{-iy_j x} \overline{c_l e^{-iy_l x}} = \\ &= \int_{\mathbb{R}} d\mu(x) \sum_{j=1}^{n} c_j e^{-iy_j x} \sum_{l=1}^{n} \overline{c_l e^{-iy_l x}} = \\ &= \int_{\mathbb{R}} \left| \sum_{j=1}^{n} c_j e^{-iy_j x} \right| d\mu(x) \geq 0. \end{aligned}$$

   □

   Nel nostro caso μ è la misura di probabilità di una variabile aleatoria con distribuzione normale e quindi ha come trasformata di Fourier la gaussiana.

In particolare se $X = \mathbb{R}$ il kernel gaussiano agisce così: $\forall x \in \mathbb{R}$ la funzione gaussiana ϕ fa corrispondere ad x la gaussiana centrata in tale punto. Chiaramente lo spazio di Hilbert a nucleo riproducente con $K(x, x')$ tali gaussiane è più ampio e contiene molto più delle funzioni gaussiane.

Dopo aver visto questi esempi, studiamo cosa succede quando proviamo a fare operazioni con le funzioni kernel definite positive.

Teorema 1.8. 1. Siano K_1, K_2 due funzioni kernel definite positive, allora $K_1 + K_2$, $K_1 K_2$ e cK_1 con $c \geq 0$ sono ancora funzioni kernel definite positive;

2. Se $(K_i)_{i\geq 1}$ è una successione di funzioni kernel definite positive che convergono puntualmente ad una funzione kernel K definita positiva, allora K è tale che $\forall (x, x') \in X \times X$ si ha $K(x, x') = \lim_{i\to\infty} K_i(x, x')$;

3. Se K è una funzione kernel definita positiva allora e^K è ancora una funzione kernel definita positiva.

Dimostrazione. 1. Dato che ogni funzione kernel definita positiva ammette una matrice hermitiana definita positiva e dato che la somma di due matrici hermitiane definite positive è ancora una matrice hermitiana definita positiva allora vale la tesi. Per quanto riguarda il prodotto di matrici hermitiane definite positive bisogna utilizzare il prodotto di Hadamard che conserva positività e simmetria.

□

Negli spazi di Hilbert a nucleo riproducente H vale la seguente disuguaglianza:

$$\begin{aligned} |f(x) - f(x')| &= |< f, K_x > - < f, K'_x >| = \\ &= |< f, K_x - K'_x >| \leq \|f\|_H \|K_x - K'_x\|_H = \\ &= \|f\|_H \, d_K(x, x') \end{aligned}$$

dove $d_K(x, x')$ è una distanza su X che si definisce tramite la funzione nucleo riproducente K:

$$\begin{aligned} d_K &\colon X \times X \to \mathbb{R} \\ (x, x') &\mapsto \|K_x - K'_x\| . \end{aligned}$$

Quindi da questa disuguaglianza si può osservare che la norma di una funzione nello spazio di Hilbert a nucleo riproducente controlla con che velocità varia la funzione su X rispetto alla geometria definita dalla distanza kernel d_K. Quindi più è piccola la norma di f, meno variano le funzioni su X.

Capitolo 2

Metodi kernel: kernel trick e representer theorem

Ora vedremo alcune applicazioni delle funzioni kernel e di spazi di Hilbert a nucleo riproducente perchè ci servirà poi per risolvere il problema della scelta, ed eventualmente della progettazione, delle giuste funzioni kernel.
Alla base di una famiglia di potenti algoritmi per l'analisi dei dati utilizzando i kernel definiti positivi ci sono due risultati teorici noti come metodi kernel:

1. il kernel trick che è basato sulla rappresentazione di kernel definiti positivi come prodotti interni;

2. il representer theorem basato su alcune proprietà del funzionale di regolarizzazione definito dalla norma dello spazio di Hilbert a nucleo riproducente.

Esempio 2.1. Consideriamo l'apprendiamento automatico supervisionato. Sia F un insieme di "funzioni di previsione" $f : X \to Y$. Per ogni famiglia di dati di prova etichettati $(x_i, y_i)_{i=1,...,n}$ tali che $x_i \in X$ e $y_i \in Y$ $\forall i = 1, ..., n$, dove y_i è l'esito di f sui dati di prova. Cerchiamo la funzione $\hat{f}$ che soddisfa il seguente minimo:

$$min_{f \in F}\left\{\frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda\Omega(f)\right\}, \tag{2.1}$$

dove $L(y_i, f(x_i))$ è una funzione obiettivo, ad esempio $L(y_i, f_(x_i)) = (y_i - f(x_i))^2$ e $\Omega(f)$ è il coefficiente di regolarizzazione di f. Le etichette y_i sono, per esempio, in:

- $Y = \{-1, 1\}$ per i problemi binari;
- $Y = \{1, ..., k\}$ per i problemi di classificazione multi-classe;
- $Y = \mathbb{R}$ per i problemi di regressione;

- $Y = \mathbb{R}^k$ per i problemi di regressione a più variabili.

Vediamo cosa vuol dire la formula (2.1) nel caso delle funzioni kernel linerari:

- assumiamo che esista una relazione lineare tra y e $x \in \mathbb{R}^p$;
- assumiamo che $f \in F$ sia del tipo $f(x) = \overline{w}^T x + b$;
- supponiamo che L sia una funzione convessa;
- supponiamo che $\Omega(f) = \|w\|^2$.

Quindi vogliamo minimizzare:

$$min_{f \in H} \left\{ \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda \|f\|_H^2 \right\}. \tag{2.2}$$

Il metodo kernel consente di mappare i dati $x \in X$ in uno spazio di Hilbert H e di lavorare con formule lineari: $\phi : X \to H$ e $f(x) =< \phi(x), f >_H$.

Il primo scopo è quello di mappare i dati $x \in X$ in uno spazio vettoriale H (lo spazio di Hilbert a nucleo riproducente) in cui:

- esistono molte operazioni geometriche (calcolo degli angoli, proiezione, distanza);
- si possono avere modelli infinito-dimensionali potenzialmente ricchi;
- la $\|f\|_H^2$ deve essere scelta in modo da essere teoricamente o empiricamente fondata.

Si osservi che questo modo di procedere è generico e non assume nulla sulla natura dell'insieme dei dati X.

Il secondo scopo è quello di migliorare ancora lo spazio vettoriale di arrivo in cui abbiamo mappato i dati, quindi:

- trasformiamo i dati in uno spazio di dimensioni superiori con proprietà migliori (ad esempio proprietà di raggruppamento);
- in questo spazio la forma lineare $f(x) =< \phi(x), f >_H$ può corrispondere a un modello che non era lineare in X.

2.1 Kernel trick

Proposizione 2.1. Qualsiasi algoritmo per elaborare vettori di dimensione finita, esprimibile in termini di prodotti interni di coppie di dati, può essere applicato a vettori potenzialmente infiniti nello spazio delle caratteristiche di un kernel definito positivo sostituendo ogni valutazione interna del prodotto con una valutazione della funzione kernel.

Osservazioni

- La dimostrazione di questa proposizione è banale, perchè la funzione kernel è esattamente il prodotto interno nello spazio delle caratteristiche.
- Questo "trucchetto del kernel" ha enormi applicazioni pratiche.
- I vettori nello spazio delle caratteristiche si manipolano solo implicitamente, attraverso prodotti interni di coppie.

Quello che facciamo mappando i dati da uno spazio X ad uno spazio di dimensione maggiore non è conveniente a livello di costi computazionali. Vediamolo nel seguente esempio:

Esempio 2.2. Se $X = \mathbb{R}^3$ con $x = (x_1, x_2, x_3) \in X$ e $y = (y_1, y_2, y_3) \in X$ e se $\phi : \mathbb{R}^3 \to \mathbb{R}^9$ è la funzione che agisce così:

$$\phi(x_1, x_2, x_3) = (x_1^2, x_1x_2, x_1, x_3, x_2x_1, x_2^2, x_2x_3, x_3x_1, x_3x_2, x_3^3) \tag{2.3}$$

cioè è la funzione che mappa i dati in uno spazio "più comodo".
Se $K(u, v) = u^T v \Rightarrow K(\phi(u), \phi(v)) = \phi(u)^T \phi(v) = \sum_{i,j=1}^{3} x_i x_j y_i y_j$ Osserviamo che fare il prodotto $K(u, v) = u^T v$ ha un costo computazionale pari a $O(n)$ perchè si eseguono n prodotti, mentre il prodotto $K(\phi(u), \phi(v))$ ha un costo computazionale pari a $O(n^2)$ perchè si eseguono $n \times n$ prodotti.

In sostanza, ciò che il kernel trick fa è offrirci un modo più efficiente e meno costoso per trasformare i dati in dimensioni superiori. Questo kernel trick è possibile grazie alle funzioni kernel $K : X \times X \to \mathbb{C}$.

Il kernel trick sembra perfetto, tuttavia è fondamentale tenere a mente che quando mappiamo i dati su una dimensione superiore, ci sono possibilità che il modello venga sovra-adattato. Quindi la scelta della giusta funzione kernel e la regolarizzazione sono di grande importanza.

Vediamo ora alcuni esempi di utilizzo del kernel trick.

Esempio 2.3 (Calcolo delle distanze nello spazio delle caratteristiche (Feature Space))**.** Se la funzione distanza è data da:

$$\begin{aligned} d_K(x_1, x_2)^2 &= \|\phi(x_1) - \phi(x_2)\|_H^2 = \\ &= < \phi(x_1) - \phi(x_2), \phi(x_1) - \phi(x_2) >_H = \\ &= < \phi(x_1), \phi(x_1) >_H + < \phi(x_2), \phi(x_2) >_H - 2 < \phi(x_1), \phi(x_2) >_H . \end{aligned}$$

Allora $d_K(x_1, x_2)^2 = K(x_1, x_1) + K(x_2, x_2) - 2K(x_1, x_2)$.

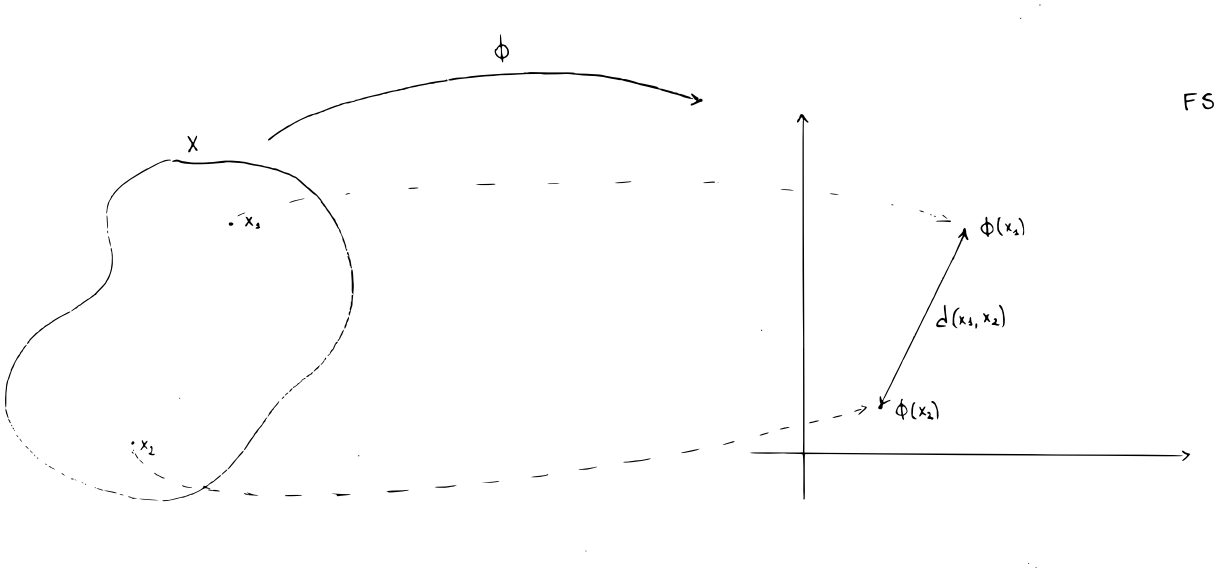

Figura 2.1: Distanza d nello spazio delle caratteristiche (FS)

Vediamo quanto vale tale distanza nel caso del kernel gaussiano. Sia K il kernel gaussiano di parametro σ su $\mathbb{R}^d$:
$K(x,y) = exp(-\frac{\|x-y\|^2}{2\sigma^2})$. Allora osserviamo che $K(x,x) = 1 = \|\phi(x)\|_H^2$, quindi tutti i punti sono sulla sfera unitaria nello spazio delle caratteristiche. La distanza tra le immagini di due punti x e y nello spazio delle caratteristiche è data da:
$d_K(x,y)^2 = 1 + 1 - 2exp(-\frac{\|x-y\|^2}{2\sigma^2}) = 2 - 2exp(-\frac{\|x-y\|^2}{2\sigma^2})$
$\Rightarrow d_K(x,y) = \sqrt{2(1 - exp(-\frac{\|x-y\|^2}{2\sigma^2}))}$.

Le funzioni distanza possono essere diverse. Se la distanza è:

$$\begin{aligned}
\delta(x_1,x_2)^2 &= \left\| \frac{\phi(x_1)}{\|\phi(x_1)\|} - \frac{\phi(x_2)}{\|\phi(x_2)\|} \right\|_H^2 = \\
&= < \frac{\phi(x_1)}{\|\phi(x_1)\|} - \frac{\phi(x_2)}{\|\phi(x_2)\|}, \frac{\phi(x_1)}{\|\phi(x_1)\|} - \frac{\phi(x_2)}{\|\phi(x_2)\|} >_H = \\
&= < \frac{\phi(x_1)}{\|\phi(x_1)\|}, \frac{\phi(x_1)}{\|\phi(x_1)\|} >_H + < \frac{\phi(x_2)}{\|\phi(x_2)\|} - \frac{\phi(x_2)}{\|\phi(x_2)\|} >_H - 2 < \frac{\phi(x_1)}{\|\phi(x_1)\|} - \frac{\phi(x_2)}{\|\phi(x_2)\|} >_H = \\
&= \frac{K(x_1,x_1)}{\|\phi(x_1)\|} + \frac{K(x_2,x_2)}{\|\phi(x_2)\|} - \frac{2K(x_1,x_2)}{\|\phi(x_1)\|\,\|\phi(x_2)\|}.
\end{aligned}$$

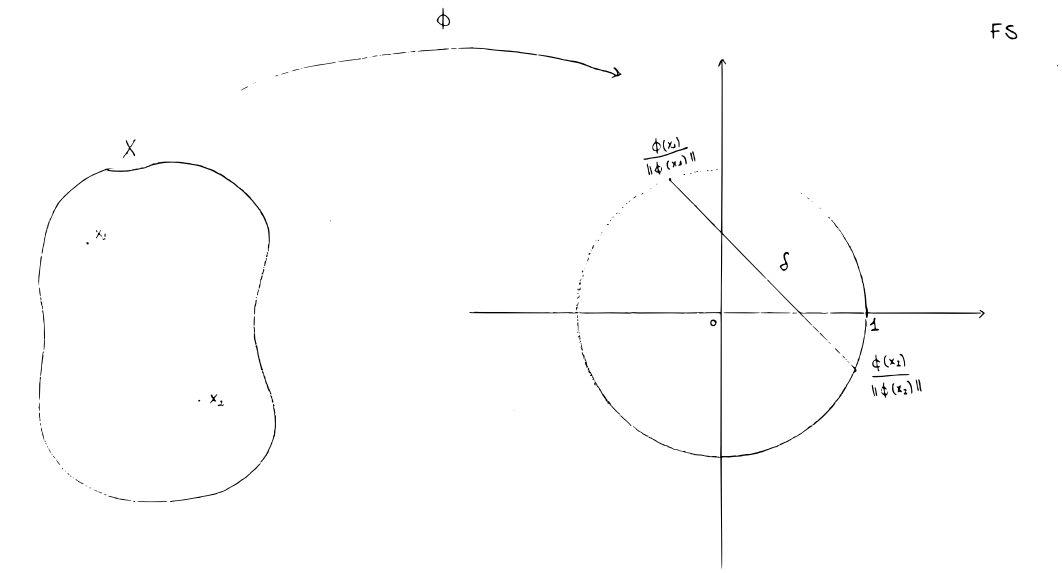

Figura 2.2: Distanza δ nello spazio delle caratteristiche (FS)

Vediamo quanto vale tale distanza nel caso di K kernel gaussiano con parametro σ su $\mathbb{R}^d$:

$K(x,y) = exp(-\frac{\|x-y\|^2}{2\sigma^2})$, allora

$$\delta(x_1,x_2)^2 = \frac{1}{\|\phi(x_1)\|} + \frac{1}{\|\phi(x_2)\|} - \frac{exp(-\frac{\|x_1-x_2\|^2)}{2\sigma^2}}{\|\phi(x_1)\|\,\|\phi(x_2)\|} = \\ = \frac{\|\phi(x_1)\| + \|\phi(x_2)\| - 2exp(-\frac{\|x_1-x_2\|^2)}{2\sigma^2}}{\|\phi(x_1)\|\,\|\phi(x_2)\|}.$$

Quindi $\delta(x_1,x_2) = \sqrt{\frac{\|\phi(x_1)\|+\|\phi(x_2)\|-2exp(-\frac{\|x_1-x_2\|^2)}{2\sigma^2}}{\|\phi(x_1)\|\|\phi(x_2)\|}}$.

Esempio 2.4 (Distanza tra un punto e un insieme)**.** Sia $S = \{x_1, ..., x_n\}$ un insieme finito di punti in X. Per definire e calcolare la distanza tra ogni punto $x \in X$ e l'insieme S possiamo procedere nel seguente modo:

- mappiamo tutti i punti nello spazio delle caratteristiche;
- consideriamo S come il suo baricentro: $\mu := \frac{1}{n}\sum_{i=1}^{n}\phi(x_i)$;
- definiamo la distanza tra x e S così: $d_K(x,S) := \|\phi(x) - \mu\|_H$.

Quindi, facendo i calcoli, si ottiene:

$$
\begin{aligned}
d_K(x,S)^2 &= \left\| \phi(x) - \frac{1}{n}\sum_{i=1}^{n}\phi(x_i) \right\|_H^2 = \\
&= \left(< \phi(x) - \frac{1}{n}\sum_{i=1}^{n}\phi(x_i), \phi(x) - \frac{1}{n}\sum_{i=1}^{n}\phi(x_i) >_H \right)^2 = \\
&= \left(< \phi(x), \phi(x) >_H - 2 < \phi(x), \frac{1}{n}\sum_{i=1}^{n}\phi(x_i) >_H + < \frac{1}{n}\sum_{i=1}^{n}\phi(x_i), \frac{1}{n}\sum_{i=1}^{n}\phi(x_i) >_H \right)^2 = \\
&= \left(K(x,x) - \frac{2}{n}\sum_{i=1}^{n}K(x,x_i) + \frac{1}{n^2}\sum_{i,j=1}^{n}K(x_i,x_j) \right)^2 .
\end{aligned}
$$

Allora abbiamo ottenuto che

$$
d_K(x,S) = \sqrt{K(x,x) - \frac{2}{n}\sum_{i=1}^{n}K(x,x_i) + \frac{1}{n^2}\sum_{i,j=1}^{n}K(x_i,x_j)}. \tag{2.4}
$$

Osservazione In generale il baricentro μ esiste solo nello spazio delle caratteristiche: non è detto che μ abbia una preimmagine x_μ tale che $\phi(x_\mu) = \mu$.

Esempio 2.5 (Centrare i dati nello spazio delle caratteristiche)**.** In questo esempio vogliamo far coincidere il baricentro dello spazio delle caratteristiche con l'origine del sistema di riferimento.

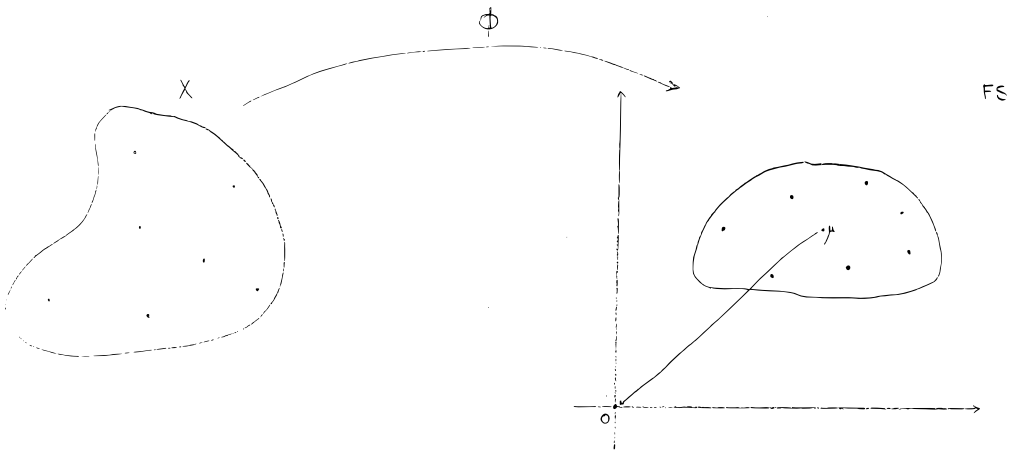

Figura 2.3: Centrare i dati nello spazio delle caratteristiche (FS)

Sia $S = \{x_1, ..., x_n\}$ un insieme finito di punti in X dotato di un kernel definito positivo K. Sia $\mathbf{K}$ la sua matrice di Gram $n \times n$: $[\mathbf{K}]_{i,j} = K(x_i, x_j)$ e sia $\mu = \frac{1}{n}\phi(x_i)$ il

baricentro di S nello spazio delle caratteristiche e siano $u_i = \phi(x_i) - \mu$ per $i = 1, ..., n$ i dati centrati in H. Vogliamo calcolare la matrice di Gram $\mathbf{K}^c$ per i dati centrati in H.

Un conto diretto è dato dalla seguente formula: per $0 \leq i, j \leq n$

$$\begin{aligned}[\mathbf{K}^c]_{i,j} =& < \phi(x_i) - \mu, \phi(x_j) - \mu >_H = \\ &= < \phi(x_i), \phi(x_j) >_H + < \mu, \mu >_H - < \mu, \phi(x_i) + \phi(x_j) >_H = \\ &= \mathbf{K}_{i,j} - \frac{1}{n} \sum_{k=1}^{n} (\mathbf{K}_{i,k} + \mathbf{K}_{j,k}) + \frac{1}{n^2} \sum_{k,l=1}^{n} \mathbf{K}_{k,l}.\end{aligned}$$

Questo, in forma matriciale, si può scrivere nel seguente modo:

$$\mathbf{K}^c = \mathbf{K\text{-}UK\text{-}KU+UKU=(I\text{-}U)K(I\text{-}U)}, \tag{2.5}$$

dove $\mathbf{U}_{i,j} = \frac{1}{n}$ per $1 \leq i, j \leq n$.

Quindi il trucchetto del kernel è un'affermazione matematicamente banale, ma con importanti applicazioni:

- può essere utilizzato per ottenere versioni non lineari di noti algoritmi lineari, ad esempio sostituendo il classico prodotto interno con un kernel gaussiano;
- può essere utilizzato per applicare algoritmi classici a dati non vettoriali (ad esempio stringhe e grafici) sostituendo nuovamente il classico prodotto interno con un kernel valido per i dati;
- in alcuni casi consente di incorporare lo spazio iniziale in uno spazio delle caratteristiche più grande e di coinvolgere punti nello spazio delle caratteristiche senza pre-immagine (es: il baricentro).

2.2 Representer theorem

Introduciamo questo argomento per i seguenti motivi:

- uno spazio di Hilbert a nucleo riproducente è uno spazio di funzioni (potenzialmente non lineari) e negli spazi di Hilbert a nucleo riproducente la norma di una funzione è scelta in modo da misurare la differenziabilità della funzione;
- dato un insieme di dati $(x_i, y_i)_{i=1,...,n} \in X \times \mathbb{C}$ un modo naturale per cercare una funzione di regressione $f : X \to \mathbb{C}$ è trovare la minimizzante per:

$$min_{f \in H} \left\{ \frac{1}{n} \sum_{i=1}^{n} l(y_i, f(x_i)) + \lambda \left\| f \right\|_H^n \right\} \tag{2.6}$$

  per una funzione obiettivo l, come ad esempio $l(y, t) = (y - t)^2$;

- come si risolve in pratica questo problema, che potrebbe essere infinto dimensionale?

Teorema 2.1. Sia X un insieme dotato di un kernel definito positivo K, sia H lo spazio di Hilbert a nucleo riproducente corrispondente a K e sia $S = \{x_1, ..., x_n\} \subseteq X$ un insieme finito di punti. Sia $\Psi : \mathbb{C}^n \times \mathbb{R} \to \mathbb{R}$ una funzione di $n+1$ variabili, strettamente crescente rispetto all'ultima variabile. Allora ogni funzione soluzione del seguente problema di ottimizzazione:

$$min_{f \in H} \left\{ \Psi(f(x_1), ..., f(x_n), \|f\|_H) \right\}, \tag{2.7}$$

ammette una rappresentazione di questa forma:

$$\forall x \in X, f(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x) = \sum_{i=1}^{n} \alpha_i K_{x_i}(x). \tag{2.8}$$

In altre parole, la soluzione f è nel sottospazio $Span(K_{x_1}, ..., K_{x_n})$ finito dimensionale.

Dimostrazione. Sia $\xi(f)$ il funzionale che soddisfa (2.7) e sia H_S lo $Span$ lineare in H dei vettori K_{x_i}:
$H_S = \{f \in H; f(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x), (\alpha_1, ...\alpha_n) \in \mathbb{R}^n\}$.
H_S è un sottospazio finito-dimensionale, quindi qualsiasi funzione $f \in H$ può essere decomposta in modo unico così: $f = f_S + f_\perp$, dove $f_S \in H_S$ e $f_\perp \perp H_S$ (attraverso la proiezione ortogonale).
Dato che H è uno spazio di Hilbert a nucleo riproducente, si ha che
$\forall i = 1, ..., n \; f_\perp(x_i) =< f_\perp, K_{x_i} >_H = 0$
perchè $K_{x_i} = (x_i, \cdot) \in H$, perciò: $\forall i = 1, ..., n \; f(x_i) = f_S(x_i)$.
Il teorema di Pitagora in H ci mostra che $\|f\|_H^2 = \|f_S\|_H^2 + \|f_\perp\|_H^2$.
Come conseguenza $\xi(f) \geq \xi(f_S)$, con $\xi(f) = \xi(f_S) \Leftrightarrow \|f_\perp\| = 0$.
Il minimo di ψ è quindi necessariamente in H_S. □

Osservazioni Spesso la funzione ψ ha la seguente forma: $\psi(f(x_1), ..., f(x_n), \|f\|_H) = c(f(x_1), ..., f(x_n)) + \lambda\Omega(\|f\|_H)$, dove $c(\cdot)$ misura l'adattamento di f ad un dato problema (regressione, classificazione, riduzione della dimensione) e Ω è strettamente crescente. Questa formulazione ha due conseguenze importanti:

- teoricamente la minimizzazione imporrà che anche la norma di f, $\|f\|_H$, sia "piccola", il che può essere vantaggioso e garantire un livello di regolarità sufficiente per la soluzione (effetto regolarizzazione);

- praticamente grazie al teorema di rappresentazione sappiamo che la soluzione vive in un sottospazio di dimensione n, il che può portare ad algoritmi efficienti nonostante lo spazio di Hilbert a nucleo riproducente possa essere di dimensione infinita.

Vediamo qual è l'uso pratico del teorema di rappresentazione. Quando vale il teorema di rappresentazione, sappiamo che possiamo cercare una soluzione della forma

$$f(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x)$$

per qualche $\alpha \in \mathbb{R}^n$. Per qualsiasi $j = 1, ..., n$ si ha $f(x_j) = \sum_{i=1}^{n} \alpha_i K(x_i, x_j) = [\mathbf{K}\alpha]_j$, inoltre

$$\|f\|_H^2 = \left\| \sum_{i=1}^{n} \alpha_i K_{x_i} \right\|_H^n = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j K(x_i, x_j) = \alpha^T \mathbf{K} \alpha.$$

Quindi un problema della forma $min_{f \in H}(\psi(f(x_1), ..., f(x_n), \|f\|_H^2))$ è equivalente al seguente problema di ottimizzazione n-dimensionale:

$$min_{\alpha \in \mathbb{C}^n} \left\{ \psi([\mathbf{K}\alpha]_1, ..., [\mathbf{K}\alpha]_n, \alpha^T \mathbf{K} \alpha) \right\}.$$

Questo problema si può solitamente risolvere analiticamente o attraverso dei metodi numerici.

Ci sono quindi due interpretazioni dei metodi kernel:

- una geometrica nello spazio delle caratteristiche data dal kernel trick. Anche quando lo spazio delle caratteristiche è grande, molti metodi kernel lavorano nello *span* lineare degli "inserimenti" dei punti disponibili;
- una funzionale, spesso come un problema di ottimizzazione (sottoinsiemi) dello spazio di Hilbert a nucleo riproducente associato al kernel.

Il teorema di rappresentazione ha conseguenze importanti, ma è in realtà piuttosto banale. Stiamo cercando una funzione $f \in H$ tale che $\forall x \in X \ f(x) =< K_x, f >_H$. La parte $f_\perp$ che è ortogonale al kernel K_{x_i} è quindi "inutile" per spiegare i dati di prova.

Capitolo 3

Metodi kernel per l'apprendimento supervisionato

L'obiettivo di un sistema basato sull'apprendimento supervisionato è quello di produrre una funzione in grado di "apprendere" dai risultati forniti durante la fase di esempio e in grado di avvicinarsi a dei risultati desiderati per tutti gli esempi non forniti.

Definizione 3.1. Dati uno spazio di inputs X, uno spazio di outputs Y e un insieme di coppie di dati di prova $S_n = (x_i, y_i)_{i=1,\dots,n}$, il problema dell'apprendimento supervisionato è quello di calcolare una funzione $h : X \to Y$ per predire l'output per tutti i futuri inputs.

Osservazione La classificazione degli outputs può essere molto varia; proprio per questo motivo, a seconda della natura di questi, la natura del problema è varia. Si hanno:

- problemi di regressione se l'output è quantitativo ($Y = \mathbb{R}$);
- problemi di classificazione se l'output è qualitativo ($Y = \{-1, 1\}$);
- problemi più strutturati se l'output è sia quantitativo che qualitativo (regressione e classificazione).

Vediamo degli esempi per ogni tipo di problema:

Esempio 3.1.

- REGRESSIONE
  L'obiettivo è quello di creare una funzione in grado di approssimare i dati; in questo esempio la funzione deve prevedere la capacità di una piccola molecola di inibire il bersaglio di un farmaco. Allora avremo che X=$\{strutture\ molecolari\}$ e $Y = \mathbb{R}$
- CLASSIFICAZIONE
  L'obiettivo è quello di creare una funzione in grado di individuare l'appartenenza

di un elemento a una data classe; in questo esempio la funzione deve riconoscere se una data immagine è un cane o un gatto. Allora avremo che X=$\{immagini\} \subseteq \mathbb{R}^d$ e $Y = \{cane,\ gatto\}$

3.1 Principi generali dell'apprendimento supervisionato con le funzioni kernel

Definizione 3.2. Dato un insieme arbitrario X, un insieme totalmente ordinato Y e una funzione $f : X \to Y$, l' $argmax$ di un sottoinsieme $S \subseteq X$ è l'insieme dei punti di massimo ed è così definito:

$$argmax_S f := argmax_{x \in S} f(x) := \{x \in S; f(s) \leq f(x)\ \forall s \in S\}\,.$$

Definizione 3.3. Dato un insieme arbitrario X, un insieme totalmente ordinato Y e una funzione $f : X \to Y$, l' $argmin$ di un sottoinsieme $S \subseteq X$ è l'insieme dei punti di minimo ed è così definito:

$$argmin_S f := argmin_{x \in S} f(x) := \{x \in S; f(s) \geq f(x)\ \forall s \in S\}\,.$$

Come prima cosa abbiamo bisogno di esprimere $h : X \to Y$ usando una funzione a valori reali $f : Z \to \mathbb{R}$. Nel caso della regressione avremo che $h(x) = f(x)$ con $f : X \to \mathbb{R}$; nel caso della classificazione avremo che $h(x) = sgn(f(x))$ con $f : X \to \mathbb{R}$; infine per gli outputs più strutturati avremo che $h(x) = argmax_{y \in Y} f(x, y)$ con $f : X \times Y \to \mathbb{R}$. Successivamente bisogna definire una funzione di rischio empirico $R_n(f)$ per valutare quanto giusta sia una candidata funzione f sull'insieme dei dati di prova S_n.

Definizione 3.4. Scelta una classe di funzioni $f \in F$ e definita una funzione di perdita $l(f(x), y)$ si definisce rischio empirico $R_n(f) := \frac{1}{n} \sum_{i=1}^{n} l(f(x_i), y_i)$.

Il rischio empirico dipende solo dai dati di prova e dalla funzione scelta. Infine bisogna cercare una funzione kernel definita positiva su Z che risolve

$$\min_{f \in H, \|f\| \leq B} R_n(f) \quad o \quad \min_{f \in H} R_n(f) + \lambda \|f\|_H^2.$$

3.2 Regressione lineare del kernel

Siano X un insieme di punti, $Y = \mathbb{R}$ e $S_n = (x_i, y_i)_{i=1,\ldots,n} \in (X \times \mathbb{R})^n$ un insieme di dati di prova. Vogliamo trovare una funzione $f : X \to \mathbb{R}$ che dà y come risultato di f calcolata in x.

Esaminiamo in particolare la regressione con il metodo dei minimi quadrati. In uno spazio funzionale qualsiasi per quantificare l'errore con il metodo dei minimi quadrati

si utilizza la funzione $l(f(x), y) = (y - f(x))^2$. Quindi risolvere il problema di regressione con il metodo dei minimi quadrati equivale a trovare la funzione $\hat{f}$ che risolve $argmin_{f \in F} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$. Nel risolvere questo problema ci possono essere alcune difficoltà come il sovradattamento, se F è troppo grande, e l'instabilità della funzione.
Consideriamo, ora, come insieme di funzioni uno spazio di Hilbert a nucleo rirpoducente H associato ad una funzione kernel K definita positiva su X. La regressione lineare del kernel si ottiene attraverso la regressione dell'errore quadratico medio attraverso la norma in H:

$$\hat{f} = argmin_{f \in H} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \|f\|_H^2 \right\}. \tag{3.1}$$

In questo modo notiamo subito che si previene il sovradattamento del problema penalizzando le funzioni non lisce. Inoltre grazie al representer theorem, tra le soluzioni di (3.1) ci sono quelle della forma

$$\hat{f}(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x).$$

Quindi otteniamo un secondo vantaggio che è quello di semplificare la soluzione.
Risolviamo il problema della regressione lineare del kernel: siano $y = (y_1, ..., y_n)^T \in \mathbb{R}^n$, $\alpha = (\alpha_1, ..., \alpha_n)^T \in \mathbb{R}^n$ e $\mathbf{K}$ la matrice di Gram $n \times n$: $\mathbf{K}_{i,j} = K(x_i, x_j)$. Possiamo scrivere $(\hat{f}(x_1), ..., \hat{f}(x_n))^T = \mathbf{K}\alpha$ e quindi vale $\left\|\hat{f}\right\|_H^2 = \alpha^T \mathbf{K} \alpha$. Allora il problema della regressione lineare del kernel (3.1) è equivalente al seguente problema:

$$argmin_{\alpha \in \mathbb{R}^n} \left\{ \frac{1}{n} (\mathbf{K}\alpha - y)^T (\mathbf{K}\alpha - y) + \lambda \alpha^T \mathbf{K} \alpha \right\}.$$

Questa è una funzione convessa differenziabile in α. Il suo minimo si può ottenere uguagliando il gradiente in α a 0:

$$0 = \frac{2}{n} \mathbf{K} (\mathbf{K}\alpha - y) + 2\lambda \mathbf{K}\alpha = \mathbf{K} \left[(\mathbf{K} + \lambda n \mathbf{I}) \alpha - y \right]. \tag{3.2}$$

Per $\lambda > 0$, $\mathbf{K} + \lambda n \mathbf{I}$ è invertibile (perchè $\mathbf{K}$ è semidefinita positiva) quindi una possibile soluzione di (3.2) è

$$\alpha = (\mathbf{K} + \lambda n \mathbf{I})^{-1} y.$$

Osservazione [Unicità della soluzione] La soluzione di (3.2) trovata è unica. Infatti $\mathbf{K}$ è una matrice simmetrica e, quindi, può essere diagonalizzata in due basi ortonormali. Inoltre $\mathbf{K}$ è tale che $Ker(\mathbf{K}) \perp Im(\mathbf{K})$. In queste basi ortonormali si vede che $(\mathbf{K} + \lambda n \mathbf{I})^{-1}$ lascia $Ker(\mathbf{K})$ e $Im(\mathbf{K})$ invariati. Quindi risolvere (3.2) è equivalente a risolvere

$$\begin{aligned} \{(\mathbf{K} + \lambda n \mathbf{I}) \alpha - y\} \in ker(\mathbf{K}) \Leftrightarrow \\ \Leftrightarrow \left\{ \alpha - (\mathbf{K} + \lambda n \mathbf{I})^{-1} y \right\} \in ker(\mathbf{K}) \Leftrightarrow \\ \Leftrightarrow \alpha = (\mathbf{K} + \lambda n \mathbf{I})^{-1} y + \epsilon \ \ con \ \ \mathbf{K}\epsilon = 0 \end{aligned}$$

Quindi se $\alpha' = \alpha + \epsilon$, con $\mathbf{K}\epsilon = 0$, allora

$$\|f - f'\|_H^2 = (\alpha - \alpha')^T \mathbf{K} (\alpha - \alpha') = 0 \Rightarrow f = f'.$$

Quindi la soluzione è unica.

Vediamo ora come la regressione lieare del kernel può essere collegata alla regressione lieare "standard". Sia $X = \mathbb{R}^d$, sia K il kernel lineare $K(x, x') = x^T x'$ e sia $\mathbf{X} = (x_1, ..., x_n)^T$ la matrice dei dati di dimensione $n \times d$. Allora la matrice del kernel è $\mathbf{K} = \mathbf{X}\mathbf{X}^T$. La funzione che otteniamo dalla soluzione del problema della regressione del kernel nel caso lineare è:

$$f(x) = w_{KRR}^T x, \quad con \quad w_{KRR} = \sum_{i=1}^{n} \alpha_i x_i = \mathbf{X}^T \alpha = \mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T + \lambda n \mathbf{I}\right)^{-1} y.$$

D'altra parte, lo spazio di Hilbert a nucleo riproducente è l'insieme delle funzioni lineari della forma $f(x) = w^T x$ e la sua norma è $\|f\|_H = \|w\|$ quindi possiamo riscrivere direttamente il problema della regressione lineare del kernel (3.1) nel seguente modo:

$$\begin{aligned} &argmin_{w \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(y_i - w^T x_i\right)^2 + \lambda \|w\|^2 \right\} = \\ &= argmin_{w \in \mathbb{R}^d} \left\{ \frac{1}{n} (y - \mathbf{X}w)^T (y - \mathbf{X}w) + \lambda w^T w \right\} \end{aligned}$$

e, imponendo il gradiente nullo, si ha la seguente soluzione:

$$w_{RR} = \left(\mathbf{X}^T \mathbf{X} + \lambda n \mathbf{I}\right)^{-1} \mathbf{X}^T y.$$

La soluzione w_{RR} non è diversa dalla soluzione w_{KRR} perchè vale il seguente

Lemma 3.1. Siano $\mathbf{B}$ una matrice $n \times k$, $\mathbf{C}$ una matrice $k \times n$ e $\gamma > 0$ vale quanto segue:

$$(\mathbf{B}\mathbf{C} + \gamma \mathbf{I})^{-1} \mathbf{B} = \mathbf{B} (\mathbf{C}\mathbf{B} + \gamma \mathbf{I})^{-1}.$$

Dimostrazione. Sapendo che

$$\mathbf{B}(\mathbf{C}\mathbf{B} + \gamma \mathbf{I}) = (\mathbf{B}\mathbf{C} + \gamma \mathbf{I})\mathbf{B},$$

moltiplicando entrambi i membri di questa uguaglianza a sinistra per $(\mathbf{B}\mathbf{C} + \gamma \mathbf{I})^{-1}$ e a destra per $(\mathbf{C}\mathbf{B} + \gamma \mathbf{I})^{-1}$, si ha che

$$(\mathbf{B}\mathbf{C} + \gamma \mathbf{I})^{-1} \mathbf{B} = \mathbf{B} (\mathbf{C}\mathbf{B} + \gamma \mathbf{I})^{-1}.$$

□

Da questo deduciamo che

$$w_{RR} = \left(\mathbf{X}^T\mathbf{X} + \lambda n\mathbf{I}\right)^{-1}\mathbf{X}^T y = \mathbf{X}^T\left(\mathbf{X}\mathbf{X}^T + \lambda n\mathbf{I}\right)^{-1} y = w_{KRR}.$$

A livello di costi computazionali invertire una matrice è la parte più "costosa", quindi questo ci suggerisce di usare:

- KRR se $d > n$;
- RR se $d < n$.

Bibliografia

[1] N. Aronszajn, *Theory of reproducing kernels, Transactions of the American Mathematical Society*, vol. 68, no. 3 (May, 1950), pp. 337-404

[2] M. Reed, B. Simon, *Functional Analysis, Methods of Modern Mathematical Physics*, 1981

[3] F. R. Bach, *Sharp analysis of low-rank kernel matrix approximations, In COLT*, vol. 30, pp. 185-209, 2013

[4] A. Bietti, J. Mairal, *On the inductive bias of neutral tangent kernels, in Adv. NeurIPS*, 2019b.

[5] B. Schölkopf, A. J. Smola, *Learning with kernels, Support Vector Machines, Regularization, Optimization and Beyond.*

[6] J. Mairal, J.P. Vert, *Machine Learning with Kernel Methods*, `http://members.cbio.mines-paristech.fr/~jvert/svn/kernelcourse/slides/master2017/master2017.pdf`, dicembre 2020, pp. 1-100