POLITECNICO DI TORINO

Corso di Laurea
in Matematica per l'Ingegneria

Tesi di Laurea

Kernels e Support Vector Machine

Politecnico
di Torino

Relatore
prof. Francesco Vaccarino

Candidato
Luca Rossi

Anno Accademico 2020-2021

Ai miei familiari

Sommario

I recenti progressi nell' hardware dei computer, l'introduzione delle GPU e la grandissima quantità di dati che viene generata ogni giorno hanno risvegliato l'interesse della comunità scientifica verso l'intelligenza artificiale e in particolar modo verso il Machine Learning. Una classe molto importante di algoritmi di Machine Learning è quella di classificazione nella quale spicca l'algoritmo di Support Vector Machine. Questo elaborato si propone di arrivare a formulare in maniera rigorosa il problema di ottimizzazione caratterizzante la Support Vector Machine e di interpretare tale problema geometricamente quando possibile. Per far ciò, nel primo capitolo tratteremo i Reproducing Kernel Hilbert Spaces e i metodi di costruzione dei Reproducing Kernels giungendo infine ad enunciare e dimostrare il Teorema di rappresentazione. Questi risultati verranno poi fortemente utilizzati nel capitolo successivo dove ci occuperemo di derivare il problema di ottimizzazione caratterizzante l'algoritmo di Support Vector Machine. Daremo successivamente un interpretazione di tale problema nel caso di Kernel lineari arrivando a definire i problemi di ottimizzazione per la Hard Margin Support Vector Machine e per la Soft Margin Support Vector Machine. Alla fine del secondo capitolo applicheremo l'algoritmo a qualche dataset e osserveremo come la scelta del Kernel influenzi il risultato della classificazione. Per derivare il problema di ottimizzazione caratterizzante l'algoritmo saranno necessari alcuni risultati di ottimizzazione vincolata (dualità, punti KKT, ecc). A tal proposito, nell'appendice A enunciamo e dimostriamo tali risultati nella speranza di rendere più semplice la lettura dell'elaborato.

Ringraziamenti

L'autore è grato in primo luogo a tutto il corpo docente del Politecnico di Torino e in particolare al professore e relatore Francesco Vaccarino. Ringrazia inoltre i suoi familiari e i suoi amici che lo hanno sempre supportato in questi tre anni di vita universitaria.

Indice

1 Reproducing Kernel Hilbert Spaces e regolarizzazione 9
- 1.1 Introduzione e risultati preliminari . . . 9
- 1.2 Regolarizzazione . . . 11
- 1.3 Reproducing Kernel Hilbert Space . . . 13
- 1.4 Reproducing Kernel via feature mapping . . . 19
- 1.5 Reproducing Kernel da funzioni caratteristiche . . . 19
- 1.6 Reproducing Kernels da features ortonormali . . . 21
- 1.7 Teorema di rappresentazione e applicazioni . . . 24

2 Classificazione tramite Support Vector Machine 29
- 2.1 Introduzione : problema di classificazione e metriche di classificazione . 29
- 2.2 Support Vector Machine . . . 30
- 2.3 Kernel lineare e interpretazione geometrica . . . 36
- 2.4 Hard margin vs Soft Margin Support Vector Machine . . . 38
- 2.5 Applicazione Support Vector Machine ad un dataset . . . 40

A Risultati di ottimizzazione vincolata 45
- A.1 Punti stazionari per ottimizzazione vincolata . . . 45
- A.2 Ottimizzazione vincolata nel caso in cui S sia definito da vincoli di uguaglianza e disuguaglianza . . . 48
- A.3 Punti KKT . . . 50
- A.4 Dualità . . . 51
- A.5 Problemi convessi di ottimizzazione vincolata . . . 53

Il m'est arrivé dans les sciences ce qui arriverait à un homme qui, s'étant levé de grand matin, attendrait avec impatience que l'aurore et le jour vinssent dissiper les ténèbres; et qui, lorsque le soleil aurait paru, se trouverait aveuglé par l'éclat de ses rayons
[MAXIMES ET RÉFLEXIONS, Johann Wolfgang von Goethe]

Capitolo 1

Reproducing Kernel Hilbert Spaces e regolarizzazione

1.1 Introduzione e risultati preliminari

Il Machine Learning è una branca dell'intelligenza artificiale che comprende metodi e algoritmi atti a riconoscere patterns e apprendere automaticamente da una serie di dati (spesso chiamati *features*). Il compito di un algoritmo di Machine Learning è infatti quello di predire un output y a partire da un vettore di input $x = [x_1, x_2, \dots, x_p]^T \in \mathbb{R}^p$, ciò viene fatto cercando una funzione g_τ chiamata spesso funzione di predizione che viene ricercata all'interno di una specifica classe di funzioni $\mathcal{G}$ definite dallo spazio delle features X in $\mathbb{R}$.
Risulta utile suddividere gli algoritmi di Machine Learning in due classi: gli algoritmi di *supervised learning* e gli algoritmi di *unsupervised learning*. Nei primi disponiamo di un training set $\tau = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, cioè di una serie di vettori di features e degli output corrispondenti. Il compito in questo caso è quello di trovare la funzione di predizione g_τ in maniera tale che questa funzione predica successivamente, in maniera accurata, gli outputs di nuovi inputs $\bar{x}_i$ i cui rispettivi outputs sono ignoti. Al contrario, nell'unsupervised learning disponiamo semplicemente delle features x_i e non degli outputs ad esse associate. L'obiettivo è allora semplicemente quello di trovare dei patterns nei dati, per esempio raggruppare le features basandosi sulla loro somiglianza come succede negli algoritmi di K-means o DBSCAN. Ci occuperemo unicamente di problemi di supervised learning. E' utile suddividere ulteriormente i problemi di apprendimento supervisionato in due classi principali: i problemi di *regressione* e quelli di *classificazione*. Nei primi l'output della funzione di predizione è continuo cioè g_τ può assumere valori in tutta la retta reale, nei secondi l'output è discreto cioè g_τ può assumere valori unicamente in un insieme $\{0,1,\dots,c-1\}$ dove $c \in \mathbb{N}$. Tipico esempio di problema di regressione può essere quello di predire il prezzo di un appartamento a partire da informazioni quali la metratura, la località, ecc. Tipico problema di classificazione può essere invece quello di classificare se una mail è spam o meno utilizzando informazioni quali la lunghezza della mail, la frequenza con cui arrivano

mail dallo stesso indirizzo di posta elettronica, ecc. Per scegliere una buona funzione di predizione g_τ è importante definire una *loss function* per il problema in questione. Chiamata infatti $\bar{y}$ la predizione della funzione g_τ e y l'output vero presente nel training set τ, la loss function è una funzione $Loss(y, \bar{y})$ che assume valori tanto più grandi quanto y e $\bar{y}$ sono diversi fra loro. Nel caso della regressione si utilizza spesso lo *scarto quadratico medio*, cioè la funzione di loss è $loss(y, \bar{y}) = (y - \bar{y})^2$, al contrario, per problemi di classificazione, si utilizza piuttosto la *zero-one loss* cioè $l_\tau(g) = \mathbb{1}\{y \neq \bar{y}\}$. La funzione di predizione g_τ è allora la funzione che minimizza la loss function tra tutte le funzioni di una certa classe $\mathcal{G}$. Assumendo ora che ogni coppia $(x_i, y_i) \in \tau$ sia l'output di una coppia di variabili aleatorie (X, Y), per un problema regressivo abbiamo il seguente teorema.

Teorema 1.1. *(Della funzione ottimale per la Squared-Error Loss): Data la Squared-Error Loss $Loss(y, \bar{y}) = (y - \bar{y})^2$, la miglior funzione di predizione g^* è il valore atteso di Y condizionato all'evento $X = x$, cioè*

$$g^*(x) = E[Y|X = x]$$

Dimostrazione. Sia $g^*(x) = E[Y|X = x]$, per ogni funzione g vale:

$$\begin{aligned} E[(Y - g(X))^2] &= E[(Y - g^*(X) + g^*(X) - g(X))^2] \\ &= E[(Y - g^*(X))^2] + 2E[(Y - g^*(X))(g^*(X) - g(X))] \\ &+ E[(g^*(X) - g(X))^2] \\ &\geq E[(Y - g^*(X))^2] + 2E[(Y - g^*(X))(g^*(X) - g(X))] \\ &= E[(Y - g^*(X))^2] + 2E[(g^*(X) - g(X))E[Y - g^*(X)|X]] \end{aligned}$$

Dove nell'ultima equazione abbiamo usato la tower property. Dalla definizione di valore atteso condizionato abbiamo che $E[(Y - g^*(X)|X] = E[Y|X] - g^*(X) = 0$, segue che $E[(Y - g(X))^2] \geq E[(Y - g^*(X))^2]$ cioè g^* minimizza la loss function. □

Al contrario, per problemi di classificazione abbiamo il seguente risultato teorico relativamente alla miglior funzione di predizione.

Teorema 1.2. *Per la loss function definita da $Loss(y, \hat{y}) = \mathbb{1}\{y \neq \hat{y}\}$, la migliore funzione di classificazione è*

$$g^*(x) = \underset{y \in \{0,1,\ldots,c-1\}}{\operatorname{argmax}} \mathbb{P}[Y = y|X = x] \tag{1.1}$$

Dimostrazione. L'obbiettivo è quello di minimizzare $l(g) = \mathbb{E}[\mathbb{1}\{Y \neq g(X)\}]$ tra tutte le funzioni g che assumono unicamente valori appartenenti a $\{0,1,\ldots,c-1\}$. Per la tower property abbiamo che

$$l(g) = \mathbb{E}[\mathbb{1}\{Y \neq g(X)\}] = \mathbb{E}[\mathbb{E}[\mathbb{1}\{Y \neq g(X)\}|X]] = \mathbb{E}[\mathbb{P}(Y \neq g(X))|X]$$

Di conseguenza, minimizzare $\mathbb{E}[\mathbb{P}(Y \neq g(X))|X]$ rispetto a g è equivalente a massimizzare $\mathbb{P}[Y = g(x)|X = x]$ rispetto a $g(x)$ per ogni x fissato. Quindi basta prendere $g(x)$ come la classe y tale per cui $\mathbb{P}[Y = y|X = x]$ è massima. □

Osservazione 1. *L'assegnazione di $g = g^*$ con g^* definita dall'equazione (1.1) divide lo spazio delle features in c regioni*

$$\mathcal{R}_y = \{x : f(y|x) = \max_{z \in \{0,1,\ldots,c-1\}} f(z|x)\} \qquad y = 0,1,\ldots,c-1$$

La classe delle funzioni $\mathcal{G}$ nella quale si ricerca la funzione di predizione dovrebbe essere quanto più semplice possibile in modo da poter permetterne uno studio teorico ma dovrebbe poter essere allo stesso tempo quanto più ricca possibile in modo da contenere g^*. Si può raggiungere questo scopo prendendo come $\mathcal{G}$ uno spazio di Hilbert (uno spazio completo rispetto alla norma indotta dal prodotto scalare). E' utile ricordare il seguente teorema enunciato e dimostrato nel corso di Analisi Funzionale.

Teorema 1.3. *(Riesz-Fréchet) Sia H uno spazio di Hilbert e sia H^* il suo duale (lo spazio vettoriale costituito da tutti i funzionali lineari e continui da H in un campo $\mathbb{F}$). L'applicazione*

$$\Phi : H \to H^*$$
$$y \to f_y$$

dove f_y è definita da $f_y(x) = \langle x, y \rangle$ è un'isometria, antilineare suriettiva.

Per esempio se scegliamo $\mathcal{G}$ come lo spazio delle funzioni lineari da $\mathbb{R}^p$ in $\mathbb{R}$, esso è uno spazio di Hilbert. Allora, in tal caso, presa una qualsiasi funzione $f \in \mathcal{G}$, per il teorema di Riesz-Fréchet esiste una feature x tale che $f(u) = \langle x, u \rangle \; \forall u \in \mathbb{R}^p$ possiamo allora identificare f con $x \in \mathbb{R}^p$.

1.2 Regolarizzazione

Se lo spazio di Hilbert $\mathcal{G}$ è abbastanza grande da permetterci di trovare una funzione di predizione g_τ tale che la training loss $l_\tau(g)$ sia zero o molto vicino a zero, si rischia di incorrere nell'*overfitting*. L'overfitting si verifica quando il modello (la funzione di predizione) impara dal training set τ ogni dettaglio e noise portando a zero (o quasi) la loss mentre il modello risulta insoddisfacente nel predire output a partire da $\bar{x}_i$ non presenti nell'insieme τ. Uno dei modi per superare questo problema è quello di introdurre un nuovo funzionale $J : \mathcal{G} \to \mathbb{R}_+$ che penalizza le funzioni complesse. Allora, introducendo una costante $c > 0$ il nuovo problema diventa:

$$\min\{l_\tau(g) : g \in \mathcal{G}, J(g) \leq c\}$$

Un tipico approccio per risolvere più facilmente il seguente problema quando è convesso e valgono le ipotesi del corollario A.5 dell'appendice è quello di massimizzare la Lagrangiana duale L_D. Dai risultati sulla dualità enunciati nell'appendice, il problema diventa quindi

$$\max_\lambda \mathcal{L}_D(\lambda)$$
$$\text{vincolato a : } \lambda \geq 0$$

dove $\mathcal{L}_D(\lambda) = \min\{l_\tau(g) + \lambda(J(g) - c) : g \in \mathcal{G}\}$. Possiamo ora analizzare un caso particolare di *regolarizzazione*:

Esempio: Ridge Regression
La Ridge Regression non è nient'altro che il problema della regressione semplice con un termine di regolarizzazione. In particolare, dato un training set
$\tau = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, con $x_i \in \mathbb{R}^p$ il problema è il seguente

$$\min_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} (y_i - g(x_i))^2 + \gamma ||g||^2$$

Prendendo come $\mathcal{G}$ lo spazio di Hilbert delle funzioni lineari da $\mathbb{R}^p$ in $\mathbb{R}$ e identificando la funzione g con un elemento $\beta \in \mathbb{R}^p$ tale per cui $g(x) = \langle \beta, x \rangle$ per ogni $x \in \mathbb{R}^p$, il problema precedente può essere espresso come

$$\min_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} (y_i - g(x_i))^2 + \gamma ||\beta||^2 \tag{1.2}$$

Introducendo la matrice $X = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix}$ (che prende il nome di *feature matrix*) dove $x_i \in \mathbb{R}^p$, il problema (1.2) diventa

$$\begin{aligned} &\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \gamma ||\beta||^2 \\ &= \min_{\beta \in \mathbb{R}^p} \frac{1}{n} ||y - X\beta||^2 + \gamma ||\beta||^2 \end{aligned}$$

La funzione che vogliamo minimizzare è una funzione convessa, infatti la sua Hessiana è semidefinita positiva, allora sappiamo che ne possiamo trovare i minimi ponendo il gradiente della funzione uguale a zero. Otteniamo quindi

$$X^T(X\beta - y) + n\gamma\beta = 0$$

Se la matrice $X^T X + n\gamma I_p$ è invertibile, abbiamo

$$\hat{\beta} = (X^T X + n\gamma I_p)^{-1} X^T y$$

Esempio: Ridge Regression con decomposizione di $\mathcal{G}$
Talvolta, quando si regolarizza rispetto ad uno spazio di Hilbert $\mathcal{G}$, è comodo decomporre $\mathcal{G}$ in due spazi ortogonali $\mathcal{H}$ e $\mathcal{C}$ tali che $\mathcal{G} = \mathcal{H} \oplus \mathcal{C}$ ciò risulta utile quando vogliamo penalizzare funzioni in $\mathcal{H}$ ma non in $\mathcal{C}$ per esempio. Vediamo ora come applicare questo concetto alla "ridge regression". Supponiamo allora che una delle features sia costante (senza perdita di generalità possiamo assumere che sia 1), in tal caso

$\tilde{x} = [1, x_1, x_2, \ldots, x_p]^T$. Sia $\mathcal{G}$ lo spazio delle funzioni lineari da $\mathbb{R}^{p+1}$ in $\mathbb{R}$, ogni $g \in \mathcal{G}$ può essere scritta come $g(\tilde{x}) = \beta_0 + x^T\beta$ che è la somma di due funzioni : $c(\tilde{x}) = \beta_0$ e $h(\tilde{x}) = x^T\beta$. Inoltre le due funzioni sono ortogonali rispetto al prodotto scalare definito su $\mathcal{G}$, infatti $\langle c, h\rangle = [\beta_0, 0^T] \cdot [0, \beta^T] = 0$. Definendo allora con $\mathcal{H}, \mathcal{C}$ rispettivamente il sottospazio vettoriale dei vettori in $\mathbb{R}^{p+1}$ con la prima componente nulla e dei vettori in $\mathbb{R}^{p+1}$ con tutte le componenti diverse dalla prima nulle, è banale vedere che $\mathcal{H}, \mathcal{C}$ sono spazi di Hilbert il cui prodotto scalare discende da quello di $\mathcal{G}$. Allora, non volendo penalizzare il termine costante il precedente problema viene così riformulato

$$\min_{g \in \mathcal{H} \oplus \mathcal{C}} \frac{1}{n} \sum_{i=1}^{n} (y_i - g(\tilde{x}_i))^2 + \gamma ||g||^2_{\mathcal{H}} = \min_{\beta_0, \beta} \frac{1}{n} ||y_i - \beta_0 1 - X\beta||^2 + \gamma ||\beta||^2$$

dove abbiamo indicato con 1 il vettore colonna di dimensioni $n \times 1$ composto unicamente da uni. Seguendo lo stesso ragionamento di prima (il problema resta convesso), calcolando il gradiente e ponendo uguale a zero si ottengono

$$X^T(\beta_0 1 + X\beta - y) + n\gamma\beta = 0 \tag{1.3}$$

$$n\beta_0 = 1^T(y - X\beta) \tag{1.4}$$

Risolvendo per β: sostituendo β_0 nella prima equazione otteniamo

$$(X^TX - n^{-1}X^T 1 1^T X + n\gamma I_p)\beta = (X^T - n^{-1}X^T 1 1^T)y$$

Se ora assumiamo che $n \geq p$ e che X abbia massimo rango, allora ogni elemento di $\mathbb{R}^p$ può essere scritto come combinazione lineare delle features $\{x_i\}$. Sostituendo allora $\beta = X^T\alpha$ dove $\alpha \in \mathbb{R}^n$, l'equazione diventa

$$(XX^T - n^{-1}1 1^T XX^T + n\gamma I_n)\alpha = (I_n - n^{-1}1 1^T)y$$

da cui se la matrice $(XX^T - n^{-1}1 1^T XX^T + n\gamma I_n)$ è invertibile

$$\hat{\alpha} = (XX^T - n^{-1}1 1^T XX^T + n\gamma I_n)^{-1}(I_n - n^{-1}1 1^T)y$$

Infine abbiamo che la soluzione per il termine costate è $\hat{\beta_0} = \frac{1}{n}1^T(y - XX^T\hat{\alpha})$ e la soluzione diventa

$$g_\tau(\tilde{x}) = \hat{\beta_0} + x^T X^T \hat{\alpha}$$

Si può allora notare che g_τ dipende dalle features unicamente tramite la matrice XX^T chiamata anche matrice di Gram.

1.3 Reproducing Kernel Hilbert Space

Definizione 1.1. *Sia X un insieme non vuoto, sia $\mathcal{G}$ uno spazio di Hilbert di funzioni $g : X \to \mathbb{R}$ con prodotto scalare $\langle\ ,\ \rangle_{\mathcal{G}}$, $\mathcal{G}$ è chiamato Reproducing Kernel Hilbert Space (RKHS) con reproducing kernel $\mathcal{K} : X \times X \to \mathbb{R}$ se:*

1. *per ogni* $x \in X$, $\mathcal{K}_x = \mathcal{K}(x, \cdot)$ *è in* $\mathcal{G}$.
2. $\mathcal{K}(x, x) < \infty$ *per ogni* $x \in X$.
3. *per ogni* $x \in X$ *e* $g \in \mathcal{G}$, $g(x) = \langle g, \mathcal{K}_x \rangle_{\mathcal{G}}$.

Proposizione 1.1. *Il reproducing Kernel di uno spazio di Hilbert di funzioni se esiste è unico.*

Dimostrazione. Supponiamo che esistano due reproducing kernels $\mathcal{K}_1 : X \times X \to \mathbb{R}$ e $\mathcal{K}_2 : X \times X \to \mathbb{R}$. Dalla proprietà 3) dei reproducing kernels, $\forall x \in X$ e $g \in \mathcal{G}$, $g(x) = \langle g, \mathcal{K}_{1_x} \rangle_{\mathcal{G}}$ e $g(x) = \langle g, \mathcal{K}_{2x} \rangle_{\mathcal{G}}$ ma allora:

$$\langle g, \mathcal{K}_{2x} \rangle_{\mathcal{G}} = \langle g, \mathcal{K}_{1x} \rangle_{\mathcal{G}} \, \forall g \in \mathcal{G}, \forall x \in X$$

Il che implica

$$\mathcal{K}_{2_x} = \mathcal{K}_{1_x} \, \forall x \in X$$

Usando ora la proprietà 1) ciò si riscrive come

$$\mathcal{K}_2(x, \cdot) = \mathcal{K}_1(x, \cdot) \, \forall x \in X$$

Che significa infine che $\mathcal{K}_1 = \mathcal{K}_2$ e quindi il reproducing kernel se esiste è unico. □

La terza condizione della definizione ci permette di valutare una funzione $g \in \mathcal{G}$ ad un certo $x \in X$ semplicemente calcolando il prodotto scalare $\langle g, \mathcal{K}_x \rangle_{\mathcal{G}}$. Ciò è molto importante, infatti,per trovare la miglior funzione di predizione nello spazio $\mathcal{G}$, non è necessario costruire esplicitamente la funzione g ma piuttosto è necessario valutare g ad ogni feauture x_i del training set cosa che è possibile fare valutando un semplice prodotto scalare. Se identifichiamo ora g con $\mathcal{K}_{x'}$, segue immediatamente il seguente fatto.

Proposizione 1.2. *I reproducing kernels sono funzioni simmetriche.*

Dimostrazione. E' sufficiente osservare che $\mathcal{K}(x, x') = \langle \mathcal{K}_{x'}, \mathcal{K}_x \rangle_{\mathcal{G}} = \langle \mathcal{K}_x, \mathcal{K}_{x'} \rangle_{\mathcal{G}} = \mathcal{K}(x', x)$. □

Proposizione 1.3. *I reproducing kernels sono funzioni semidefinite positive.*

Dimostrazione.

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \mathcal{K}(x_i, x_j) \alpha_j = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \mathcal{K}_{x_i}(x_j) \alpha_j = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \left\langle \mathcal{K}_{x_i}, \mathcal{K}_{x_j} \right\rangle_{\mathcal{G}} \alpha_j$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} \left\langle \alpha_i \mathcal{K}_{x_i}, \alpha_j \mathcal{K}_{x_j} \right\rangle_{\mathcal{G}} = \left\langle \sum_{i=1}^{n} \alpha_i \mathcal{K}_{x_i}, \sum_{j=1}^{n} \alpha_j \mathcal{K}_{x_j} \right\rangle_{\mathcal{G}} = || \sum_{i=1}^{n} \alpha_i \mathcal{K}_{x_i} ||^2 \geq 0$$

□

Osservazione 2. *Ciò implica che ogni matrice di Gram (matrice tale per cui* $K_{ij} = \mathcal{K}(x_i, x_j)$*) associata a* $\mathcal{K}$ *è semidefinita positiva.*

Teorema 1.4. *(Caratterizzazione dei RKHS) Un RKHS $\mathcal{G}$ su un insieme non vuoto X è uno spazio di Hilbert in cui ogni funzionale di valutazione $\delta_x : g \to g(x)$ è limitato. Al contrario, uno spazio di Hilbert $\mathcal{G}$ di funzioni definite da $X \to \mathbb{R}$ per cui ogni funzionale di valutazione è limitato è un RKHS.*

Dimostrazione. Mostrare la limitatezza dei funzionali di valutazione δ_x è analogo a mostrarne la continuità in quanto sono operatori lineari. Immaginiamo allora di avere un RKHS e sia g_n una successione di funzioni in $\mathcal{G}$ convergenti ad un dato g. In tal caso,

$$|\delta_x g_n - \delta_x g| = |g_n(x) - g(x)| = |\langle g_n - g, \mathcal{K}_x \rangle_{\mathcal{G}}|$$

Adesso usando la disuguaglianza di Cauchy-Schwartz:

$$\begin{aligned} |\delta_x g_n - \delta_x g| = |\langle g_n - g, \mathcal{K}_x \rangle_{\mathcal{G}}| &\leq ||g_n - g||_{\mathcal{G}} ||\mathcal{K}_x||_{\mathcal{G}} \\ &= ||g_n - g||_{\mathcal{G}} \sqrt{\langle \mathcal{K}_x, \mathcal{K}_x \rangle_{\mathcal{G}}} = ||g_n - g||_{\mathcal{G}} \sqrt{\mathcal{K}(x,x)} \end{aligned}$$

Ora visto che $\sqrt{\mathcal{K}(x,x)} < \infty$ e che $||g_n - g||_{\mathcal{G}} \to 0$ quando $n \to \infty$ allora abbiamo che $|\delta_x g_n - \delta_x g| \to 0$ cioè δ_x è continua. Supponiamo contrariamente ora che tutti i funzionali siano limitati, allora per il teorema di Riesz-Fréchet esiste $g_{\delta_x} \in \mathcal{G}$ tale che $\delta_x g = \langle g, g_{\delta_x} \rangle_{\mathcal{G}}$ per ogni $g \in \mathcal{G}$. Definendo ora $\mathcal{K}(x, x') = g_{\delta_x}(x')$ per ogni $x, x' \in X$, allora:

- $\mathcal{K}_x = \mathcal{K}(x, \cdot) = g_{\delta_x}$ è un elemento di $\mathcal{G}$ per ogni $x \in X$.
- $\mathcal{K}(x,x) = \langle \mathcal{K}_x, \mathcal{K}_x \rangle_{\mathcal{G}} = \langle g_{\delta_x}, g_{\delta_x} \rangle_{\mathcal{G}} = ||g_{\delta_x}||^2_{\mathcal{G}} < \infty$
- $\langle g, \mathcal{K}_x \rangle_{\mathcal{G}} = \delta_x g = g(x)$ cioè anche la proprietà 3 della definizione dei RKHS è soddisfatta.

□

Il seguente teorema mostra come ogni funzione $\mathcal{K} : X \times X \to \mathbb{R}$ che sia finita, simmetrica e semidefinita positiva, può essere un reproducing kernel.

Teorema 1.5. *(Moore-Aronszanj) Sia X un insieme non vuoto, sia una funzione $\mathcal{K} : X \times X \to \mathbb{R}$, finita, simmetrica e semidefinita positiva. Esiste un reproducing kernel Hilbert space $\mathcal{G}$ di funzioni $g : X \to \mathbb{R}$ con reproducing kernel $\mathcal{K}$. In più, $\mathcal{G}$ è unico.*

Dimostrazione. Abbiamo già enunciato e provato che se un reproducing Kernel esiste allora è unico. Ci resta allora solamente da provare l'esistenza, sotto queste ipotesi, di un tale RKHS.
Costruiamo un pre-RKHS $\mathcal{G}_0$ a partire dalla funzione $\mathcal{K}$ presente nelle ipotesi del teorema e poi estenderemo $\mathcal{G}_0$ ad un RKHS $\mathcal{G}$. Definiamo $\mathcal{G}_0$ come l'insieme delle combinazioni lineari finite delle funzioni $\mathcal{K}_x$, $x \in X$. In formule

$$\mathcal{G}_0 = \left\{ g = \sum_{i=1}^{n} \alpha_i \mathcal{K}_{x_i} : x_1, x_2, \ldots, x_n \in X, \alpha_i \in \mathbb{R}, n \in \mathbb{N} \right\}$$

Definiamo ora su $\mathcal{G}_0$ il seguente prodotto scalare:

$$\langle f, g \rangle_{\mathcal{G}_0} = \left\langle \sum_{i=1}^{n} \alpha_i \mathcal{K}_{x_i}, \sum_{j=1}^{m} \beta_j \mathcal{K}_{x'_j} \right\rangle_{\mathcal{G}_0} = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j \mathcal{K}(x_i, x'_j).$$

E' facile osservare che tale operazione è effettivamente un prodotto scalare su $\mathcal{G}_0$. Inoltre, $\mathcal{G}_0$ gode di alcune importanti proprietà:

1. I funzionali di valutazione sono continui/limitati
2. Le sequenze di Cauchy in $\mathcal{G}_0$ convergenti puntualmente sono anche convergenti in norma

Proviamo ora 1) e 2).

1) Notiamo che preso $g \in \mathcal{G}_0$, $g = \sum_{i=1}^{n} \alpha_i \mathcal{K}_{x_i}$, inoltre,

$$\langle g, \mathcal{K}_x \rangle_{\mathcal{G}_0} = \sum_{i=1}^{n} \alpha_i \langle \mathcal{K}_{x_i}, \mathcal{K}_x \rangle_{\mathcal{G}_0} = \sum_{i=1}^{n} \alpha_i \mathcal{K}(x_i, x) = g(x)$$

Allora possiamo scrivere il funzionale di valutazione $\delta_x g$ come $\delta_x g = \langle g, \mathcal{K}_x \rangle_{\mathcal{G}_0}$. Possiamo ora procedere a mostrare che $\delta_x g$ è limitato in $\mathcal{G}_0$. Infatti,

$$|\delta_x g| = |\langle g, \mathcal{K}_x \rangle_{\mathcal{G}_0}| \leq ||g||_{\mathcal{G}_0} \cdot ||\mathcal{K}_x||_{\mathcal{G}_0}$$

Dove abbiamo sfruttato la disuguaglianza di Cauchy-Schwartz. Ora, visto che $\mathcal{K}$ è finita per ipotesi, abbiamo,

$$||\mathcal{K}_x||_{\mathcal{G}_0} = \sqrt{\langle \mathcal{K}_x, \mathcal{K}_x \rangle_{\mathcal{G}_0}} = \sqrt{\mathcal{K}(x, x)} < \infty$$

Il che finisce la dimostrazione che $\delta_x g$ è limitato.

2) Sia f'_n una successione di Cauchy in $\mathcal{G}_0$ tale che $f'_n(x) \xrightarrow{n \to \infty} h(x)\ \forall x \in X$, mostriamo che $||f'_n - h||_{\mathcal{G}_0} \xrightarrow{n \to \infty} 0$. Per semplicità possiamo dimostrare il seguente fatto che è equivalente, sia f_n una successione di Cauchy tale che $|f_n(x)| \xrightarrow{n \to \infty} 0\ \forall x \in X$, mostriamo che $||f_n||_{\mathcal{G}_0} \xrightarrow{n \to \infty} 0$ (infatti basta prendere $f'_n - h = f_n$ per ricondursi al caso di sopra).

Prima di tutto, sia f_n una successione di Cauchy in $\mathcal{G}_0$, chiaramente f_n è limitata cioè $||f_n||_{\mathcal{G}_0} \leq B$ per qualche valore di $B \in \mathbb{R}$. Inoltre, per definizione di successione di Cauchy abbiamo che $\forall \epsilon > 0\ \ \exists N \in \mathbb{N}$ tale che $\forall n, m \geq N$ con $n, m \in \mathbb{N}$, abbiamo $||f_n - f_m|| < \epsilon$. Consideriamo ora $f_N \in \mathcal{G}_0$, abbiamo $f_N = \sum_{t=1}^{T} \alpha_t \mathcal{K}_{x_t}$, visto che f_n converge puntualmente a zero, abbiamo che $\exists N_0 : f_n(x_t) < \epsilon\ \forall t \in \{1, 2, \ldots, T\}$ e per ogni $n > N_0$. Allora, quando $n > \max\{N_0, N\}$, abbiamo:

$$\begin{aligned} ||f_n||^2_{\mathcal{G}_0} &= |\langle f_n - f_N + f_N, f_n \rangle_{\mathcal{G}_0}| \leq |\langle f_n - f_N, f_n \rangle_{\mathcal{G}_0}| + |\langle f_N, f_n \rangle_{\mathcal{G}_0}| \\ &\leq ||f_n||_{\mathcal{G}_0} ||f_n - f_N||_{\mathcal{G}_0} + |\sum_{t=1}^{T} \alpha_t \langle \mathcal{K}_{x_t}, f_n \rangle_{\mathcal{G}_0}| \\ &\leq ||f_n||_{\mathcal{G}_0} ||f_n - f_N||_{\mathcal{G}_0} + \sum_{t=1}^{T} |\alpha_t| |f_n(x_t)| < \epsilon' \end{aligned}$$

Dove l'ultima disuguaglianza vale perchè $|f_n(x_t)| < \epsilon$ e $||f_n - f_N|| < \epsilon$. Il che conclude la dimostrazione.
Adesso possiamo allora estendere $\mathcal{G}_0$ all'insieme $\mathcal{G}$ di tutte le funzioni $g : X \to \mathbb{R}$ per le quali esiste una successione di Cauchy in $\mathcal{G}_0$ convergente puntualmente a g e possiamo allora definire come prodotto scalare su $\mathcal{G}$ il seguente:

$$\langle f, g \rangle = \lim_{n\to\infty} \langle f_n, g_n \rangle_{\mathcal{G}_0}$$

dove $f_n \to f$ e $g_n \to g$. Per provare che $\mathcal{G}$ è un RKHS rimangono da verificare le seguenti affermazioni:

1. Questo prodotto scalare è ben definito.
2. I funzionali di valutazione rimangono limitati.
3. Lo spazio $\mathcal{G}$ è completo.

In modo da poter utilizzare il teorema di caratterizzazione dei RKHS. Proviamo le affermazioni 1), 2), 3).

1) Mostriamo innanzitutto che $\lim_{n\to\infty} \langle f_n, g_n \rangle_{\mathcal{G}_0}$ converge. Sia $a_n = \langle f_n, g_n \rangle_{\mathcal{G}_0} \in \mathbb{R}$. Abbiamo che

$$\begin{aligned} a_n - a_m &= \langle f_n, g_n \rangle_{\mathcal{G}_0} - \langle f_m, g_m \rangle_{\mathcal{G}_0} \\ &= \langle f_n, g_n \rangle_{\mathcal{G}_0} - \langle f_m, g_n \rangle_{\mathcal{G}_0} + \langle f_m, g_n \rangle_{\mathcal{G}_0} - \langle f_m, g_m \rangle_{\mathcal{G}_0} \\ &= \langle f_n - f_m, g_n \rangle_{\mathcal{G}_0} + \langle f_m, g_n - g_m \rangle_{\mathcal{G}_0} \\ &\leq ||f_n - f_m||_{\mathcal{G}_0} \cdot ||g_n||_{\mathcal{G}_0} + ||f_m||_{\mathcal{G}_0} \cdot ||g_n - g_m||_{\mathcal{G}_0} \xrightarrow{n,m\to\infty} 0 \end{aligned}$$

Poichè f_n e g_n sono successioni di Cauchy, allora a_n è di Cauchy in $\mathbb{R}$ completo, il che implica a_n convergente. Mostriamo ora che il limite è indipendente dalla successione di Cauchy usata. Siano f_n e f'_n due successioni di Cauchy convergenti a f, sia $a_n = \lim_{n\to\infty} \langle f_n, g_n \rangle$ e sia $a'_n = \lim_{n\to\infty} \langle f'_n, g_n \rangle$. Allora

$$a_n - a'_n = \lim_{n\to\infty} \langle f_n - f'_n, g_n \rangle \leq \lim_{n\to\infty} ||f_n - f'_n||_{\mathcal{G}_0} \cdot ||g_n||_{\mathcal{G}_0} = 0$$

Perchè f_n e f'_n convergono entrambe ad f. Un ragionamento analogo segue se consideriamo ora f_n e f'_n due successioni di Cauchy convergenti a f e g_n e g'_n due successioni di Cauchy convergenti a g. Abbiamo quindi mostrato che il limite è indipendente dalla successione di Cauchy scelta. Verifichiamo ora che le proprietà del prodotto scalare sono soddisfatte:

- multilinearità :

$$\langle \alpha f + \beta g, w \rangle = \lim_{n\to\infty} \langle \alpha f_n + \beta g_n, w_n \rangle_{\mathcal{G}_0} = \lim_{n\to\infty} \alpha \langle f_n, w_n \rangle_{\mathcal{G}_0} + \beta \langle g_n, w_n \rangle_{\mathcal{G}_0}$$
$$\lim_{n\to\infty} \alpha \langle f_n, w_n \rangle_{\mathcal{G}_0} + \lim_{n\to\infty} \beta \langle g_n, w_n \rangle_{\mathcal{G}_0} = \alpha \langle f, w \rangle + \beta \langle g, w \rangle$$

- simmetria:

$$\langle f, g \rangle = \lim_{n\to\infty} \langle f_n, g_n \rangle_{\mathcal{G}_0} = \lim_{n\to\infty} \langle g_n, f_n \rangle_{\mathcal{G}_0} = \langle g, f \rangle$$

- definito positivo:

$$\langle f, f\rangle = \lim_{n\to\infty} \langle f_n, f_n\rangle_{\mathcal{G}_0} \geq 0 \quad \text{perchè } \langle f_n, f_n\rangle_{\mathcal{G}_0} \geq 0 \text{ , inoltre,}$$
$$\langle f, f\rangle = 0 \iff \lim_{n\to\infty} \langle f_n, f_n\rangle_{\mathcal{G}_0} = 0 \iff \lim_{n\to\infty} ||f_n||^2 = 0 \iff$$
$$||\lim_{n\to\infty} f_n||^2 = 0$$

  Che significa che $\lim_{n\to\infty} f_n = 0$ cioè $f = 0$.

Per mostrare 2) e 3) è necessario prima mostrare il seguente lemma:

Lemma 1.1. *: $\mathcal{G}_0$ è denso in $\mathcal{G}$ nel senso che ogni $f \in \mathcal{G}$ è il limite (rispetto alla norma di $\mathcal{G}$) di una successione di Cauchy (f_n) di $\mathcal{G}_0$.*

Dimostrazione. : Per definizione di $\mathcal{G}$ ogni f è il limite puntuale di una successione di Cauchy (f_n) di $\mathcal{G}_0$, dobbiamo allora solamente provare la successione converge ad f in norma $||\cdot||_{\mathcal{G}}$ (sappiamo già che converge in norma $||\cdot||_{\mathcal{G}_0}$). Cioè per ogni $\epsilon > 0$ esiste un $N > 0$ tale che $\forall n > N$, $||f - f_n||_{\mathcal{G}_0} < \epsilon$, visto che $\lim_{m\to\infty} ||f_m - f_n||_{\mathcal{G}_0} = ||f - f_n||_{\mathcal{G}}$, allora: $||f - f_n||_{\mathcal{G}} = \lim_{m\to\infty} ||f_m - f_n||_{\mathcal{G}_0} < \epsilon$, e ciò mostra la convergenza in norma $||\cdot||_{\mathcal{G}}$. □

Mostriamo ora 2). Sarà sufficiente mostrare la continuità alla funzione 0, la continuità all'intero spazio di funzioni seguirà dalla linearità. In formule, ci basterà provare che

$$\forall \epsilon > 0, \exists \delta > 0 : \forall f \in \mathcal{G} : ||f||_{\mathcal{G}} < \delta \Longrightarrow |f(x)| < \epsilon$$

Sia f_n una successione di Cauchy convergente a f, Sia $\epsilon > 0$, per convergenza puntuale abbiamo che esiste un N_1 tale per cui $\forall n \geq N_1$, $|f_n(x) - f(x)| < \epsilon/2$. Inoltre per continuità dei funzionali di valutazione in $\mathcal{G}_0$, esiste un δ tale per cui se $||g||_{\mathcal{G}_0} < \delta$ allora $|g(x)| < \epsilon/2$. Infine, come risultato del lemma, abbiamo che possiamo fissare un $N \geq N_1$ tale che $||f_n - f||_{\mathcal{G}} < \delta/2\ \forall n \geq N$. Vale allora che se $||f_N||_{\mathcal{G}_0} < \delta$ allora $|f_N(x)| < \epsilon/2$. Si noti che f_N è parte della successione di Cauchy che converge ad f, cioè $||f_N||_{\mathcal{G}} = ||f_N||_{\mathcal{G}_0}$. Ora se prendiamo $||f||_{\mathcal{G}} < \delta/2$, abbiamo
$||f_N||_{\mathcal{G}} \leq ||f||_{\mathcal{G}} + ||f - f_N||_{\mathcal{G}} < \delta/2 + \delta/2 = \delta$. Ciò implica
$|f(x)| \leq |f(x) - f_N(x)| + |f_N(x)| < \epsilon/2 + \epsilon/2 = \epsilon$, il che significa che
$||f||_{\mathcal{G}} < \delta/2 \longrightarrow |\delta_x(f)| = |f(x)| < \epsilon$ il che mostra che i funzionali di valutazione sono continui in $\mathcal{G}$.

Proviamo infine 3). Prendiamo una generica successione di Cauchy (f_n) in $\mathcal{G}$. Visto che i funzionali di valutazione sono limitati, per ogni $x \in X$, $||f_n - f_m||_{\mathcal{G}} < \delta \iff |f_n(x) - f_m(x)| < \epsilon$. Allora $(f_n(x))$ è una successione di Cauchy in $\mathbb{R}$. Per completezza di $\mathbb{R}$ esiste $f(x) = \lim_{n\to\infty} f_n(x) \in \mathbb{R}$ per ogni $x \in X$. Per il lemma esiste per ogni n una sequenza $(g_{n,j}\ j = 1,2\ldots)$ in $\mathcal{G}_0$ tale per cui, per ogni $j > N_n$, vale $||f_n - g_{n,j}||_{\mathcal{G}} < 1/n$, definendo $g_n = g_{n,N_n}$, g_n converge puntualmente ad f. Infatti,

$$|g_n(x) - f(x)| \leq |g_n(x) - f_n(x)| + |f_n(x) - f(x)| \xrightarrow{n\to\infty} 0$$

Notiamo che,

$$||g_m - g_n||_{\mathcal{G}_0} = ||g_m - g_n||_{\mathcal{G}} \leq ||g_m - f_m||_{\mathcal{G}} + ||f_m - f_n||_{\mathcal{G}} + ||f_n - g_n||_{\mathcal{G}}$$

e tutti i termini a destra tendono a zero se n, m tendono ad infinito quindi (g_n) è di Cauchy in $\mathcal{G}_0$. Inoltre, abbiamo studiato dal lemma che data una sequenza in $\mathcal{G}_0$ con limite puntuale in $\mathcal{G}$, converge a tale limite anche in $||\cdot||_{\mathcal{G}}$. Allora $||g_n - f||_{\mathcal{G}} \to 0$. Infine, visto che $||f_n - g_n||_{\mathcal{G}} \to 0$, ne segue che $||f_n - f||_{\mathcal{G}} \to 0$ e quindi $\mathcal{G}$ è completo. □

1.4 Reproducing Kernel via feature mapping

Il metodo più classico per costruire un RKHS è quello di utilizzare una feature map cioè un'applicazione $\Phi : X \to \mathbb{R}^p$ e di definire $\mathcal{K}(x, x') = \langle \Phi(x), \Phi(x') \rangle$ dove il prodotto scalare usato è quello standard di $\mathbb{R}^p$. Chiaramente $\mathcal{K}$ è finita e simmetrica, inoltre è anche semidefinita positiva, infatti, sia Φ la matrice con righe $\Phi(x_1)^T, \Phi(x_2)^T, \ldots, \Phi(x_n)^T$ e sia $\alpha = [\alpha_1, \ldots, \alpha_n]^T \in \mathbb{R}^n$, allora vale,

$$\sum_{i=0}^{n}\sum_{j=0}^{n} \alpha_i \mathcal{K}(x_i, x_j)\alpha_j = \sum_{i=0}^{n}\sum_{j=0}^{n} \alpha_i \Phi(x_i)^T \Phi(x_j)\alpha_j = \alpha^T \Phi\Phi^T \alpha = ||\Phi^T \alpha||^2 \geq 0$$

Sorge spontaneo chiedersi se ad ogni reproducing kernel corrisponde un'unica feature map. Ciò non vale.

Controesempio Sia $X = \mathbb{R}$ e consideriamo le due feature maps:

1. $\phi_1 : X \to \mathbb{R}$ dove $\phi_1(x) = x$
2. $\phi_2 : X \to \mathbb{R}^2$ dove $\phi_2(x) = \frac{1}{\sqrt{2}}[x, x]^T$.

Allora vale che

1. $\mathcal{K}_{\phi_1}(x, x') = \langle \phi_1(x), \phi_1(x') \rangle = xx'$.
2. $\mathcal{K}_{\phi_2}(x, x') = \langle \phi_2(x), \phi_2(x') \rangle = xx'$

Quindi lo stesso reproducing Kernel Hilbert space è associato a due feature maps diverse.

1.5 Reproducing Kernel da funzioni caratteristiche

Teorema 1.6. *(Reproducing kernel da funzione caratteristica): Sia Z un vettore aleatorio a valori in $\mathbb{R}^p$ che sia simmetrico rispetto all'origine (Z e $-Z$ sono identicamente distribuiti). Sia ora ψ_z la sua funzione caratteristica, dove $\psi_z(t) = \mathbb{E}[e^{it^T z}]$ per ogni $t \in \mathbb{R}^p$, allora $\mathcal{K}(x, x') = \psi_z(x - x')$ è un valido RKHS su $\mathbb{R}^p$.*

Dimostrazione. Sarà sufficiente dimostrare che $\mathcal{K}$ è finita, a valori reali, simmetrica e semidefinita positiva, il risultato segue allora dal teorema di Moore-Aronszanj.

- $\mathcal{K}$ è finita:

$$|\mathcal{K}(x,x')| = |\psi_z(x-x')| = |\mathbb{E}\left[e^{i(x-x')^T X}\right]| \leq \mathbb{E}\left[|e^{i(x-x')^T X}|\right] = \mathbb{E}[1] = 1$$

- $\mathcal{K}$ è a valori reali: Dal fatto che

$$\begin{aligned}\mathcal{K}(x,x') =& \psi_z(x-x') = \psi_{-z}(x-x') = \mathbb{E}\left[e^{i(x-x')^T(-z)}\right] = \mathbb{E}\left[\overline{e^{i(x-x')^T z}}\right] = \\ & \overline{\mathbb{E}\left[e^{i(x-x')^T(z)}\right]} = \overline{\psi_z(x-x')}\end{aligned}$$

  Da cui abbiamo che $\psi_z(x-x') = \psi_{-z}(x-x') = \overline{\psi_z(x-x')}$, da cui $\psi_z(x-x') = \mathcal{K}(x,x') \in \mathbb{R}$.

- $\mathcal{K}$ è simmetrica:

$$\begin{aligned}\mathcal{K}(x,x') = \psi_z(x-x') = \mathbb{E}\left[e^{i(x-x')^T z}\right] = \mathbb{E}\left[e^{-i(x'-x)^T z}\right] = \\ = \psi_{-z}(x'-x) = \psi_z(x'-x) = \mathcal{K}(x',x)\end{aligned}$$

  Dove l'ultima uguaglianza vale per simmetria di Z

- $\mathcal{K}$ è semidefinita positiva:

$$\begin{aligned}&\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\mathcal{K}(x_i,x_j)\alpha_j = \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\psi_z(x_i-x_j)\alpha_j = \\ &\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\mathbb{E}\left[e^{i(x_i-x_j)^T z}\right]\alpha_j = \mathbb{E}\left[\left(\sum_{i=1}^{n}\alpha_i e^{ix_i^T z}\right)\cdot\left(\overline{\sum_{j=1}^{n}\alpha_j e^{ix_j^T z}}\right)\right] = \\ &\mathbb{E}\left[\left|\sum_{i=1}^{n}\alpha_i e^{ix_i^T z}\right|^2\right] \geq 0\end{aligned}$$

□

L'esempio più classico di RKHS da una funzione caratteristica è il *Kernel Gaussiano*. Infatti, la distribuzione gaussiana multivariata con media $\overline{0}$ e matrice di varianza e covarianza $\frac{1}{\sigma^2}I_p$ è simmetrica attorno all'origine, inoltre la sua funzione caratteristica è

$$\psi(t) = e^{-\frac{1}{2\sigma^2}||t||^2} \; t \in \mathbb{R}^p$$

Allora $\mathcal{K}(x,x') = e^{-\frac{1}{2\sigma^2}||x-x'||^2} \; t \in \mathbb{R}^p$. Un motivo che rende i Kernel gaussiani oggetto di interesse è il fatto che essi godano della *proprietà di approssimazione universale*: lo spazio delle funzioni generate dal Kernel Gaussiano è denso nello spazio delle funzioni continue con il medesimo supporto della gaussiana multivariata.

1.6 Reproducing Kernels da features ortonormali

Un altro modo per costruire RKHS è quello di lavorare con funzioni della classe $\mathcal{L}^2(X, \mu)$, l'insieme delle funzioni al quadrato integrabili su X rispetto alla misura μ. Sia per semplicità μ la misura di Lebesgue, $\{\nu_1, \nu_2, \dots\}$ una base ortonormale di $\mathcal{L}^2(X, \mu) = \mathcal{L}^2(X)$ e $\{c_1, c_2, \dots\}$ una sequenza di reali positivi. Da quanto discusso per i reproducing kernels generati da feature maps, con le stesse notazioni utilizzate in precedenza, abbiamo definito $\mathcal{K}(x, x') = \phi(x)^T \phi(x') = \sum_{i=1}^{p} \phi_i(x)\phi_i(x')$. Ispirati da ciò, se definiamo $\phi_i = c_i \nu_i \;\; i = 1,2, \dots$ definiamo allora

$$\mathcal{K}(x, x') = \sum_{i \geq 1} \phi_i(x)\phi_i(x') = \sum_{i \geq 1} c_i^2 \nu_i(x)\nu_i(x') = \sum_{i \geq 1} \lambda_i \nu_i(x)\nu_i(x') \tag{1.5}$$

Dove abbiamo posto $\lambda_i = c_i^2$, $\mathcal{K}$ è ben definito se assumiamo $\sum_{i \geq 1} \lambda_i < \infty$ (cosa che faremo da adesso in poi per ipotesi). Definiamo adesso

$$\mathcal{H} = \{f = \sum_{i \geq 1} \alpha_i \nu_i \;:\; \sum_{i \geq 1} \alpha_i^2 / \lambda_i < \infty\}$$

Dal momento che ogni $f \in \mathcal{L}^2(X)$ può essere scritta come $f = \sum_{i \geq 1} \langle f, \nu_i \rangle \nu_i$, vediamo che $\mathcal{H}$ è un sottospazio vettoriale di $\mathcal{L}^2(X)$. Definiamo il seguente prodotto scalare:

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{i \geq 1} \frac{\langle f, \nu_i \rangle \langle g, \nu_i \rangle}{\lambda_i}$$

Abbiamo che la norma definita a partire da questo prodotto scalare è $||f||_{\mathcal{H}}^2 = \sum_{i \geq 1} \alpha_i^2 / \lambda_i < \infty$. Mostriamo ora che $\mathcal{H}$ è un RKHS verificando le proprietà caratterizzanti.

1. $\mathcal{K}_x = \sum_{i \geq 1} \lambda_i \nu_i(x) \nu_i \in \mathcal{H}$, poichè $\sum_{i \geq 1} \lambda_i < \infty$
2. $\mathcal{K}(x, x') = \sum_{i \geq 1} \lambda_i \nu_i(x) \nu_i(x') < \infty$ sempre poichè $\sum_{i \geq 1} \lambda_i < \infty$
3. infine vale la reproducing property:

$$\langle \mathcal{K}_x, f \rangle_{\mathcal{H}} = \sum_{i \geq 1} \frac{\langle \mathcal{K}_x, \nu_i \rangle \langle f, \nu_i \rangle}{\lambda_i} = \sum_{i \geq 1} \frac{\lambda_i \nu_i(x) \alpha_i}{\lambda_i} = \sum_{i \geq 1} \alpha_i \nu_i(x) = f(x)$$

Ciò mostra che validi kernels possono essere costruiti a partire da (1.5). Il teorema di Mercer mostra invece che sotto opportune ipotesi, ogni kernel $\mathcal{K}$ può essere scritto nella forma (1.5).

Teorema 1.7. *(Mercer): Sia $\mathcal{K} : X \times X \to \mathbb{R}$ un reproducing kernel per un sottoinsieme X compatto di $\mathbb{R}^p$. Se valgono le seguenti condizioni:*

1. *$\mathcal{K}$ è continuo in $X \times X$*
2. *la funzione $\tilde{\mathcal{K}}(x) = \mathcal{K}(x, x)$ per $x \in X$ è integrabile.*

Allora esiste una sequenza numerabile di reali non negativi $\{\lambda_l\}$ decrescenti a zero e funzioni $\{\nu_l\}$ ortonormali in $\mathcal{L}^2(X)$ tali che:

$$\mathcal{K}(x, x') = \sum_{l \geq 1} \lambda_l \nu_l(x) \nu_l(x') \quad \text{per ogni} \quad x, x' \in X \tag{1.6}$$

Dove la convergenza è assoluta e uniforme su $X \times X$. Inoltre, se $\lambda_l > 0$, allora (λ_l, ν_l) è una coppia autovalore-autofunzione per l'operatore con nucleo integrale $K : \mathcal{L}^2(X) \to \mathcal{L}^2(X)$ definito da $[Kf](x) = \int_X \mathcal{K}(x, y) f(y) dy$ per $x \in X$.

Osservazione 3. *Un fatto chiave del suddetto teorema consiste nel fatto che la serie (1.6) converge assolutamente ed uniformemente su $X \times X$, questa proprietà di convergenza uniforme permette di poter trasferire alla serie le proprietà di integrabilità e di continuità delle somme parziali.*

Un altro modo per costruire Kernel è quello di costruirli a partire da altri kernels. Il teorema seguente formalizza tale concetto.

Teorema 1.8. *(Regole di costruzione di Kernels da altri Kernels)*

1. *Se $\mathcal{K} : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ è un reproducing kernel e $\phi : X \to \mathbb{R}$ è una funzione, allora $\mathcal{K}(\phi(x), \phi(x'))$ è un reproducing kernel da $X \times X \to \mathbb{R}$.*
2. *Se $\mathcal{K} : X \times X \to \mathbb{R}$ è un reproducing kernel e $f : X \to \mathbb{R}_+$ è una funzione, allora $f(x)\mathcal{K}(x, x')f(x')$ è un reproducing kernel da $X \times X \to \mathbb{R}$.*
3. *Se $\mathcal{K}_1$ e $\mathcal{K}_2$ sono reproducing kernels da $X \times X \to \mathbb{R}$, lo è anche la loro somma $\mathcal{K}_1 + \mathcal{K}_2$.*
4. *Se $\mathcal{K}_1$ e $\mathcal{K}_2$ sono reproducing kernels da $X \times X \to \mathbb{R}$, lo è anche il loro prodotto $\mathcal{K}_1\mathcal{K}_2$.*
5. *Se $\mathcal{K}_1$ e $\mathcal{K}_2$ sono reproducing kernels rispettivamente da $X \times X \to \mathbb{R}$ e da $\mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, allora $\mathcal{K}_+((x, y), (x', y')) = \mathcal{K}_1(x, x') + \mathcal{K}_2(y, y')$ e $\mathcal{K}_\times((x, y), (x', y')) = \\ = \mathcal{K}_1(x, x')\mathcal{K}_2(y, y')$ sono reproducing kernels da $(X \times \mathcal{Y}) \times (X \times \mathcal{Y}) \to \mathbb{R}$.*

Dimostrazione. :

1. In luce del teorema di Moore-Aronszajn, sarà sufficiente mostrare che $\mathcal{K}'(x, x') = \mathcal{K}(\phi(x), \phi(x'))$ è una funzione finita, simmetrica e semidefinita positiva da $X \times X \to \mathbb{R}$.
   - $\mathcal{K}'(x, x') = \mathcal{K}(\phi(x), \phi(x'))$ è finita in quanto se chiamiamo $y = \phi(x)$ e $y' = \phi(x')$, abbiamo che $\mathcal{K}'(x, x') = \mathcal{K}(y, y') < \infty$ perchè $\mathcal{K}$ è reproducing kernel per ipotesi e quindi è finito.
   - $\mathcal{K}'(x, x') = \mathcal{K}(\phi(x), \phi(x'))$ è simmetrica in quanto se chiamiamo $y = \phi(x)$ e $y' = \phi(x')$, abbiamo che $\mathcal{K}'(x, x') = \mathcal{K}(y, y') = \mathcal{K}(y', y) = \mathcal{K}'(x', x)$ perchè $\mathcal{K}$ è reproducing kernel per ipotesi e quindi è simmetrico.
   - $\mathcal{K}'(x, x') = \mathcal{K}(\phi(x), \phi(x'))$ è semidefinita positiva, infatti, se chiamiamo $y = \phi(x)$ e $y' = \phi(x')$, abbiamo che $\sum_{i=1}^n \sum_{j=1}^n \alpha_i \mathcal{K}'(x_i, x_j')\alpha_j = \\ = \sum_{i=1}^n \sum_{j=1}^n \alpha_i \mathcal{K}(y_i, y_j')\alpha_j \geq 0$ perchè $\mathcal{K}$ è reproducing kernel per ipotesi e quindi è semidefinito positivo.

2. Nuovamente mostreremo che $\mathcal{K}'$ è una funzione finita, simmetrica e semidefinita positiva.

   - $\mathcal{K}'(x,x') = f(x)\mathcal{K}(x,x')f(x')$ è finita perchè prodotto di funzioni finite.
   - $\mathcal{K}'(x,x') = f(x)\mathcal{K}(x,x')f(x')$ è simmetrica, infatti: $\mathcal{K}'(x,x') = f(x)\mathcal{K}(x,x')f(x') = f(x')\mathcal{K}(x',x)f(x) = \mathcal{K}'(x',x)$ dove abbiamo usato la commutatività del prodotto e la simmetria di $\mathcal{K}$ in quanto è un reproducing kernel per ipotesi.
   - $\mathcal{K}'(x,x') = f(x)\mathcal{K}(x,x')f(x')$ è semidefinita positiva, infatti:

   $$\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\mathcal{K}'(x_i,x_j')\alpha_j = \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i f(x_i)\mathcal{K}(x_i,x_j')f(x_j')\alpha_j =$$
   $$= \sum_{i=1}^{n}\sum_{j=1}^{n}c_{ij}\alpha_i\mathcal{K}(x_i,x_j')\alpha_j \geq \sum_{i=1}^{n}\sum_{j=1}^{n}\min_{i,j}\{c_{ij}\}\alpha_i\mathcal{K}(x_i,x_j')\alpha_j =$$
   $$= n^2\min_{i,j}\{c_{ij}\}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\mathcal{K}(x_i,x_j')\alpha_j \geq 0$$

   Perchè $\mathcal{K}$ è un reproducing kernel e quindi è semidefinito positivo e $c_{ij} = f(x_i)f(x_j) > 0$ per ogni coppia (i,j).

3. Come fatto per i precedenti punti mostriamo $\mathcal{K}'$ è una funzione finita, simmetrica e semidefinita positiva.

   - $\mathcal{K}'$ è finita perchè somma di funzioni finite
   - $\mathcal{K}'$ è simmetrica, infatti $\mathcal{K}'(x,x') = \mathcal{K}_1(x,x') + \mathcal{K}_2(x,x')$, usando ora la simmetria di $\mathcal{K}_1$ e $\mathcal{K}_2$ abbiamo: $\mathcal{K}'(x,x') = \mathcal{K}_1(x',x) + \mathcal{K}_2(x',x) = \mathcal{K}'(x',x)$.
   - $\mathcal{K}'$ è semidefinito positivo, infatti

   $$\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\mathcal{K}'(x_i,x_j')\alpha_j = \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i(\mathcal{K}_1(x_i,x_j') + \mathcal{K}_2(x_i,x_j'))\alpha_j =$$
   $$= \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\mathcal{K}_1(x_i,x_j')\alpha_j + \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\mathcal{K}_2(x_i,x_j')\alpha_j \geq 0$$

   In quanto $\mathcal{K}_1$ e $\mathcal{K}_2$ sono semidefiniti positivi.

4. Mostriamo per semplicità tale proprietà nel caso in cui $\mathcal{K}_1$ e $\mathcal{K}_2$ ammettano una rappresentazione del tipo (1.5), in tal caso, visto che :

   $$\mathcal{K}_1(x,x')\mathcal{K}_2(x,x') = \left(\sum_{i\geq 1}\phi_i^{(1)}(x)\phi_i^{(1)}(x')\right)\left(\sum_{j\geq 1}\phi_j^{(2)}(x)\phi_j^{(2)}(x')\right) =$$
   $$\sum_{i,j\geq 1}\phi_i^{(1)}(x)\phi_j^{(2)}(x)\phi_i^{(1)}(x')\phi_j^{(2)}(x') = \sum_{k\geq 1}\phi_k(x)\phi_k(x') = \mathcal{K}'(x,x')$$

   Il che mostra che anche $\mathcal{K}_1\mathcal{K}_2$ può essere scritto nella forma (1.5) e quindi è reproducing kernel.

5. Per mostrare $\mathcal{K}_+$ è un valido reproducing kernel mostriamo nuovamente che $\mathcal{K}_+ : (X \times Y) \times (X \times Y) \to \mathbb{R}$ è una funzione finita, simmetrica e semidefinita positiva.

   - finita: è finita poichè somma di funzioni finite (sono entrambi kernel).
   - simmetrica :

     $$\mathcal{K}_+((x,y),(x',y')) = \mathcal{K}_1(x,x') + \mathcal{K}_2(y,y') = \mathcal{K}_1(x',x) + \mathcal{K}_2(y',y) = \mathcal{K}_+((x',y'),(x,y))$$

   - semidefinita positiva:

     $$\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\mathcal{K}_+((x_i,y_i),(x'_j,y'_j))\alpha_j = \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i(\mathcal{K}_1(x_i,x'_j) + \mathcal{K}_2(y_i,y'_j))\alpha_j$$
     $$\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\mathcal{K}_1(x_i,x'_j)\alpha_j + \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\mathcal{K}_2(y_i,y'_j)\alpha_j \geq 0$$

     Poichè $\mathcal{K}_1, \mathcal{K}_2$ sono kernels.

   Mostreremo $\mathcal{K}_\times : (X \times Y) \times (X \times Y) \to \mathbb{R}$ è un reproducing kernel nel caso in cui $\mathcal{K}_1$ e $\mathcal{K}_2$ ammettano una rappresentazione della forma (1.5). In tal caso,

   $$\mathcal{K}_\times((x,y),(x',y')) = \mathcal{K}_1(x,x')\mathcal{K}_2(y,y') =$$
   $$\left(\sum_{i\geq 1}\Phi_i^{(1)}(x)\Phi_i^{(1)}(x')\right)\left(\sum_{j\geq 1}\Phi_j^{(2)}(x)\Phi_j^{(2)}(x')\right) =$$
   $$\sum_{i,j\geq 1}\Phi_i^{(1)}(x)\Phi_j^{(2)}(y)\Phi_i^{(1)}(x')\Phi_j^{(2)}(y') = \sum_{k\geq 1}\Phi_k(x,y)\Phi_k(x',y')$$

   Il che mostra che anche $\mathcal{K}_\times$ può essere scritto nella forma (1.5) ed è quindi anche esso un reproducing kernel.

□

1.7 Teorema di rappresentazione e applicazioni

Ritorniamo ora al problema originale, dato il training set $\tau = \{(x_i, y_i)\ i = 1,2,\dots n\}$ ed una loss function, il nostro obiettivo è quello di minimizzare la loss con l'aggiunta del termine di regolarizzazione. Per far ciò assumiamo però che $\mathcal{G}$, la classe delle funzioni di predizione, possa ora essere decomposta nella somma diretta di un RKHS $\mathcal{H}$ (definito da un kernel $\mathcal{K}$) e uno spazio $\mathcal{H}_0$, un altro spazio vettoriale di funzioni da X in $\mathbb{R}$. In formule:

$$\mathcal{G} = \mathcal{H} \oplus \mathcal{H}_0$$

Il che significa che per ogni $g \in \mathcal{G}$, abbiamo:

$$g = h + h_0$$

dove $h \in \mathcal{H}$, $h_0 \in \mathcal{H}_0$. Nella loss vorremo ora penalizzare il termine h di g ma non il termine h_0, il problema allora diventa:

$$\min_{g \in \mathcal{H} \oplus \mathcal{H}_0} \frac{1}{n} \sum_{i=1}^{n} Loss(y_i, g(x_i)) + \gamma ||g||^2_{\mathcal{H}}. \tag{1.7}$$

dove per $||g||_{\mathcal{H}}$ intendiamo semplicemente $||h||_{\mathcal{H}}$ se $g = h + h_0$. Tipicamente si sceglie lo spazio $\mathcal{H}_0$ come uno spazio di dimensione finita sufficientemente piccola. Considerando ora un tale problema, abbiamo un importante risultato che prende il nome di Teorema di Rappresentazione.

Teorema 1.9. *(Teorema di rappresentazione) La soluzione al problema (1.7) è della forma seguente :*

$$g(x) = \sum_{i=1}^{n} \alpha_i \mathcal{K}(x_i, x) + \sum_{j=1}^{m} \nu_j q_j(x) \tag{1.8}$$

Dove $\{q_1, q_2, \dots q_m\}$ è una base di $\mathcal{H}_0$, ciò è noto come kernel trick.

Dimostrazione. Sia $\mathcal{F} = Span\{\mathcal{K}_{x_i},\ i = 1,2,\dots,n\}$, naturalmente abbiamo $\mathcal{F} \subseteq \mathcal{H}$. Lo spazio di Hilbert $\mathcal{H}$ può allora essere rappresentato come $\mathcal{H} = \mathcal{F} \oplus \mathcal{F}^{\perp}$, dove $\mathcal{F}^{\perp}$ è il complemento ortogonale di $\mathcal{F}$. Da definizione di $\mathcal{F}^{\perp}$, abbiamo

$$\mathcal{F}^{\perp} = \{f^{\perp} \in \mathcal{H} : \left\langle f^{\perp}, f \right\rangle_{\mathcal{H}} = 0, f \in \mathcal{F}\} = \{f^{\perp} \in \mathcal{H} : \left\langle f^{\perp}, \mathcal{K}_{x_i} \right\rangle_{\mathcal{H}} = 0 \, \forall i = 1,2,\dots,n\}$$

Da cui segue facilmente per la proprietà dei reproducing kernel

$$\mathcal{F}^{\perp} = \{f^{\perp} \in \mathcal{H} : f^{\perp}(x_i) = 0 \, \forall i = 1,2 \dots, n\}$$

Allora, presa una qualsiasi funzione $g \in \mathcal{H} \oplus \mathcal{H}_0$ possiamo scrivere $g = f + f^{\perp} + h_0$ con $f \in \mathcal{F}$, $f^{\perp} \in \mathcal{F}^{\perp}$, $h_0 \in \mathcal{H}_0$. Allora abbiamo

$$||g||^2_{\mathcal{H}} = ||f + f^{\perp}||^2_{\mathcal{H}} = ||f||^2_{\mathcal{H}} + ||f^{\perp}||^2_{\mathcal{H}}$$

Dove abbiamo usato la definizione di $||\cdot||_{\mathcal{H}}$ e la perpendicolarità di f e $f^{\perp}$. Segue allora che

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} Loss(y_i, g(x_i)) + \gamma ||g||^2_{\mathcal{H}} &= \frac{1}{n} \sum_{i=1}^{n} Loss(y_i, f(x_i) + h_0(x_i)) + \gamma(||f||^2_{\mathcal{H}} + ||f^{\perp}||^2_{\mathcal{H}}) \\ &\geq \frac{1}{n} \sum_{i=1}^{n} Loss(y_i, f(x_i) + h_0(x_i)) + \gamma ||f||^2_{\mathcal{H}} \end{aligned}$$

Adesso, visto che otteniamo l'uguaglianza quando poniamo $f^{\perp} = 0$, ciò implica che la soluzione del problema (1.7) giace nel sottospazio vettoriale $\mathcal{F} \oplus \mathcal{H}_0$ di $\mathcal{G} = \mathcal{H} \oplus \mathcal{H}_0$ ed è quindi nella forma (1.8). □

Sostituendo allora la rappresentazione nella forma (1.8) di g nel problema di partenza (1.7), otteniamo il problema di ottimizzazione seguente:

$$\min_{\alpha \in \mathbb{R}^n, \nu \in \mathbb{R}^m} \frac{1}{n} \sum_{i=1}^{n} Loss(y_i, (K\alpha + Q\nu)_i) + \gamma \alpha^T K \alpha$$

Dove

- K è una matrice $n \times n$ chiamata matrice di Gram, tale per cui $K_{ij} = \mathcal{K}(x_i, x_j)\ i, j = 1,2, \ldots, n$
- Q è una matrice $n \times m$ con entrate $Q_{ij} = q_j(x_i)\ i = 1, 2, \ldots n,\ j = 1,2 \ldots m$.

Ora nel caso particolare della squared-error loss, abbiamo :

$$\min_{\alpha \in \mathbb{R}^n, \nu \in \mathbb{R}^m} \frac{1}{n} ||y - (K\alpha + Q\nu)||^2 + \gamma \alpha^T K \alpha \tag{1.9}$$

Si osserva che il problema è convesso (per esempio calcolando l'Hessiana e vedendo che essa risulta semidefinita positiva), allora differenziando (1.9) rispetto ad α e ν e uguagliando a zero otteniamo:

$$\begin{bmatrix} KK^T + n\gamma K & KQ \\ Q^T K^T & Q^T Q \end{bmatrix} \begin{bmatrix} \alpha \\ \nu \end{bmatrix} = \begin{bmatrix} K^T \\ Q^T \end{bmatrix} y$$

Continuazione ridge regression: Ritornando all'esempio della ridge regression, sia ora $\mathcal{H}$ il RKHS delle funzioni kernel lineari cioè $\mathcal{K}(x, x') = x^T x'$ e $\mathcal{H}_0$ lo spazio lineare delle funzioni costanti ($\mathcal{H}_0$ è generato dalla funzione $q_1 = 1$). Abbiamo allora $K = XX^T$ e $Q = 1$. Richiamando ora il teorema di rappresentazione, possiamo riscrivere il problema (1.2) nel modo seguente

$$\min_{\alpha, \nu_0} \frac{1}{n} ||y - \nu_0 1 - XX^T \alpha||^2 + \gamma ||X^T \alpha||^2$$

Che è nuovamente un problema di ottimizzazione convessa, ponendo allora il gradiente a zero otteniamo:

$$\begin{bmatrix} XX^T XX^T + n\gamma XX^T & XX^T 1 \\ 1^T XX^T & n \end{bmatrix} \begin{bmatrix} \alpha \\ \nu \end{bmatrix} = \begin{bmatrix} XX^T \\ 1^T \end{bmatrix} y$$

Questo è un sistema di $n + 1$ equazioni la cui soluzione è quella trovata risolvendo il sistema (1.3), (1.4). Tuttavia, il beneficio di questa formulazione è che in questo modo il problema può essere espresso direttamente in termini della matrice di Gram K senza dover calcolare esplicitamente i vettori delle features, ciò ci permette di usare anche una possibile infinita rappresentazione dello spazio delle features.

Esempio:(Stima massimi e minimi di una funzione)
Data una funzione di due variabili,per esempio

$$f(x_1,x_2) = 3(1-x_1)^2 e^{-x_1^2-(x_2+1)^2} - 10(\frac{x_1}{5} - x_1^3 - x_2^5)e^{-x_1^2-x_2^2} - \frac{1}{3}e^{-(x_1+1)^2-x_2^2}$$

L'obiettivo è quello di quello di predire i valori di $y = f(x)$ basandosi su un piccolo insieme di valori $\tau = (x_i, y_i)$ della funzione f, esso costituisce il nostro training set. In particolare, ci potremmo restringere all'insieme $[-3,3] \times [-3,3] \in \mathbb{R}^2$ e usare dei quasi random point $x_i \in [-3,3] \times [-3,3]$ e calcolare $y_i = f(x_i)$ per ogni $x_i \in [-3,3] \times [-3,3]$. L'esempio mostra quanto le curve di livello della funzione f possano essere ben approssimate anche avendo accesso ad un piccolo training set, considereremo infatti $n = 20$. Usando un kernel gaussiano in $\mathbb{R}^2$, e denotando con $\mathcal{H}$ l'unico RKHS associato al kernel e omettendo il termine di regolarizzazione, siamo interessati alla soluzione di

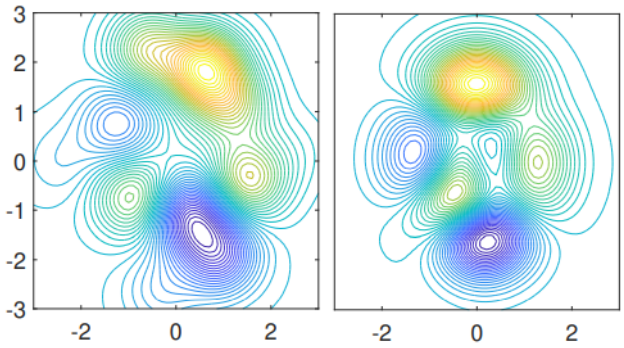
$$\min_{g\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}(y_i - g(x_i))^2$$

Dal teorema di rappresentazione la funzione ottimale è della forma

$$g(x) = \sum_{i=1}^{n} \alpha_i e^{-\frac{1}{2}\frac{||x-x_i||^2}{\sigma^2}}$$

Dove $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_n]^T$ è soluzione di $KK^T\alpha = Ky$ come visto nell'esempio della ridge regression senza regolarizzazione. L'implementazione di tale metodo porta al risultato illustrato. Come si può osservare dalla figura le curve di livello delle due

Figura 1.1: curve di livello della funzione di predizione g (a sinistra) e curve di livello per f (a destra).



funzioni si assomigliano molto. Il risultato è sorprendente se pensiamo che abbiamo utilizzato solo un training set con 20 punti distinti.

Capitolo 2

Classificazione tramite Support Vector Machine

2.1 Introduzione : problema di classificazione e metriche di classificazione

Supponiamo ora di affrontare un problema di classificazione come descritto nella sezione 1.1. Cioè dato un training set $\tau = \{(x_i, y_i)\}_{i=1}^n$ dove ogni y_i assume unicamente valori appartenenti a $\{0,1,\dots,c-1\}$. L'obbiettivo di un algoritmo di classificazione è quello di trovare una funzione $g : X \to \{0,1,\dots,c-1\}$ chiamata funzione di classificazione o classificatore che minimizzi la loss media

$$l(g) = \mathbb{E}[Loss(Y, g(X))] \tag{2.1}$$

Se ora però supponiamo che il dati del training set $\{(x_i, y_i)\}_{i=1}^n$ siano iid (indipendenti e identicamente distribuiti), si può minimizzare

$$l_\tau(g) = \frac{1}{n}\sum_{i=1}^{n} Loss(y_i, g(x_i)) \tag{2.2}$$

rispetto a $g \in \mathcal{G}$. Abbiamo diverse possibilità di scelta della loss function, una delle più naturali è la scelta di $Loss(y, \hat{y}) = \mathbb{1}\{y \neq \hat{y}\}$ dove $\hat{y} = g(x)$ come accennato nel capitolo 1. Infatti, $Loss(y, \hat{y}) = 1$ se e solo se $y \neq \hat{y}$ cioè se e solo se $y \neq g(x)$. Per tale scelta della funzione di loss vale il teorema 1.2 come descritto nella sezione 1.1.

La scelta della funzione indicatrice come loss function può non sempre essere la scelta più appropriata. Prendiamo infatti come esempio quello della diagnosi di una malattia, commettere un errore classificando un non malato come malato può in tal caso essere più grave che classificare un malato come non malato. In casi come questi si valuta la bontà di una funzione di classificazione attraverso dei particolari indicatori chiamati *metriche di classificazione*. Una delle più importanti è l'accuracy che è il rapporto tra il numero di features correttamente classificate e il numero totale delle features.

2.2 Support Vector Machine

Supponiamo ora di dover risolvere un problema di classificazione binario: dato il training set $\tau = \{(x_i, y_i)\}_{i=1}^n$ ogni y_i assume unicamente i valori ± 1. Notiamo che tale problema di classificazione è esattamente lo stesso problema descritto nella sezione 2.1 quando $c = 1$ dopo aver rinominato le classi 0,1 come -1,1. Come diretta conseguenza abbiamo che la migliore funzione di classificazione per la loss data dalla funzione indicatrice $\mathbb{1}\{y \neq \hat{y}\}$ è la seguente:

$$g^*(x) = \begin{cases} 1 & \text{se } \mathbb{P}[Y = 1|X = x] \geq \frac{1}{2} \\ -1 & \text{se } \mathbb{P}[Y = 1|X = x] < \frac{1}{2} \end{cases}$$

Inoltre abbiamo il seguente teorema.

Teorema 2.1. *La funzione g^* può essere vista come il minimo del valore atteso della hinge loss function, $Loss(y, \hat{y}) = \max\{0,1 - y\hat{y}\}$ tra tutte le funzioni di predizione. In formule:*

$$g^* = \operatorname*{argmin}_g \mathbb{E}[\max\{0,1 - Yg(X)\}]$$

Dimostrazione. Sarà sufficiente provare che per ogni funzione h vale:

$$\mathbb{E}[\max\{0,1 - Yh(X)\}] \geq \mathbb{E}[\max\{0,1 - Yg^*(X)\}]$$

Dalla definizione di valore atteso abbiamo che:

$$\mathbb{E}[\max\{0,1 - Yh(X)\}|X = x] = \sum_{y \in \{-1,1\}} \max\{0,1 - yh(x)\}\mathbb{P}[Y = y|X = x] = \\ \max\{0,1 - h(x)\}\alpha(x) + \max\{0,1 + h(x)\}(1 - \alpha(x))$$

con $\alpha(x) = \mathbb{P}[Y = 1|X = x]$. Se ora valutiamo l'espressione per $h = g^*$ otteniamo $\mathbb{E}[\max\{0,1 - Yh(X)\}|X = x] = 2\min\{\alpha(x), 1 - \alpha(x)\}$.
Infatti $g^*(x) = -1$ se $\alpha(x) < \frac{1}{2}$ e in tal caso $\mathbb{E}[\max\{0,1 - Yh(X)\}|X = x] = 2\alpha(x)$ e $g^*(x) = 1$ if $\alpha(x) \geq \frac{1}{2}$ e in tal caso $\mathbb{E}[\max\{0,1 - Yh(X)\}|X = x] = 2(1 - \alpha(x))$. Disegnando il grafico di $\mathbb{E}[\max\{0,1 - Yh(X)\}|X = x]$ come funzione di $\alpha(x)$ si nota che $\mathbb{E}[\max\{0,1 - Yh(X)\}|X = x] = 2\min\{\alpha(x), 1 - \alpha(x)\}$. Quindi sarà sufficiente provare che per ogni funzione $h(x)$ e ogni $\alpha(x) \in [0,1]$ abbiamo:

$$\max\{0,1 - h(x)\}\alpha(x) + \max\{0,1 + h(x)\}(1 - \alpha(x)) \geq 2\min\{\alpha(x), 1 - \alpha(x)\}$$

Fissando ora un x,

- Se $h(x) \geq 1$

  $$\max\{0,1 - h(x)\}\alpha(x) + \max\{0,1 + h(x)\}(1 - \alpha(x)) = (1 + h(x))(1 - \alpha(x)) \\ \geq 2(1 - \alpha(x)) \geq 2\min\{\alpha(x), 1 - \alpha(x)\}$$

- Se $h(x) \leq -1$

$$\max\{0,1-h(x)\}\alpha(x) + \max\{0,1+h(x)\}(1-\alpha(x)) = (1-h(x))\alpha(x)$$
$$\geq 2\alpha(x) \geq 2\min\{\alpha(x), 1-\alpha(x)\}$$

- Se $h(x) \in [-1,1]$ abbiamo

$$\max\{0,1-h(x)\}\alpha(x) + \max\{0,1+h(x)\}(1-\alpha(x)) = (1-h(x))\alpha(x) + (1+h(x))(1-\alpha(x)$$
$$\geq (1-h(x))\min\{\alpha(x),1-\alpha(x)\} + (1+h(x))\min\{\alpha(x),1-\alpha(x)\} \geq 2\min\{\alpha(x), 1-\alpha(x)\}$$

Quindi, per ogni x, qualsiasi sia la funzione h il bound inferiore è dimostrato.

□

Similmente a quanto visto in (2.2) al posto di minimizzare $l(g) = \mathbb{E}[\max\{0, 1-Yg(X)\}]$ possiamo minimizzare

$$l_\tau(g) = \frac{1}{n}\sum_{i=1}^{n} \max\{0,1-y_i g(x_i)\}$$

per g in una (piccola) classe di funzioni ottenendo la funzione minimizzatrice g_τ. Visto che in generale g_τ non è una funzione di classificazione (non assume solamente valori ± 1) si pone come funzione di classificazione $g(x) = \text{sign}g_\tau(x)$. Come conseguenza, un feature vector x è classificato come 1 o -1 rispettivamente se $g_\tau(x) \geq 0$ e $g_\tau(x) < 0$. Si definisce *bordo decisionale* l'insieme delle x tali per cui $g_\tau(x) = 0$. Similmente a quanto visto nel capitolo 1, dato il training data τ, possiamo trovare la miglior funzione classificatrice risolvendo tale problema:

$$\min_{g \in \mathcal{H}\oplus\mathcal{H}_0} \frac{1}{n}\sum_{i=1}^{n} \max\{0,1-y_i g(x_i)\} + \hat{\gamma}||g||^2_{\mathcal{H}}$$

dove $\hat{\gamma}$ è il parametro di regolarizzazione, $\mathcal{H}$ è un RKHS e $\mathcal{H}_0$ è uno spazio di funzioni reali da $X \to \mathbb{R}$. Se ora definiamo $\gamma = 2n\hat{\gamma}$ possiamo risolvere il seguente problema equivalente:

$$\min_{g \in \mathcal{H}\oplus\mathcal{H}_0} \sum_{i=1}^{n} \max\{0,1-y_i g(x_i)\} + \frac{\gamma}{2}||g||^2_{\mathcal{H}}$$

Se ora assumiamo che $\mathcal{K}$ sia il reproducing kernel associato a $\mathcal{H}$ e se assumiamo che $\mathcal{H}_0$ sia lo spazio delle funzioni costanti, per il teorema di rappresentazione sappiamo che la funzione minimizzatrice g è della forma:

$$g(x) = \alpha_0 + \sum_{i=1}^{n} \alpha_i \mathcal{K}(x_i, x)$$

Sostituendo ora tale g nell'espressione da minimizzare il nuovo problema diventa il seguente:

$$\min_{\alpha,\alpha_0} \sum_{i=1}^{n} \max\{0,1 - y_i(\alpha_0 + \{K\alpha\}_i)\} + \frac{\gamma}{2}\alpha^T K\alpha \tag{2.3}$$

con K la matrice di Gram, il procedimento è analogo a quanto visto nella sezione 1.7. Notiamo che il problema è un problema di ottimizzazione convesso in quanto la funzione da minimizzare è convessa perchè somma di una funzione convessa ($\frac{\gamma}{2}\alpha^T K\alpha$) e di termini lineari in α e α_0. Inoltre, chiamando $\varepsilon_i = \max\{0,1 - y_i(\alpha_0 + \{K\alpha\}_i)\}$ possiamo riscrivere il problema come un problema vincolato nel seguente modo:

$$\min_{\alpha,\alpha_0,\varepsilon} \sum_{i=1}^{n} \varepsilon_i + \frac{\gamma}{2}\alpha^T K\alpha \quad \text{soggetto a} \quad \begin{cases} \varepsilon \geq 0 \\ y_i(\alpha_0 + \{K\alpha\}_i) \geq 1 - \varepsilon_i, \;\; i = 1,\ldots,n \end{cases} \tag{2.4}$$

dove con $\varepsilon \geq 0$, intendiamo che ogni sua componente è non negativa. Notiamo ora che la funzione da minimizzare resta anche in questo caso convessa per le stesse ragioni enunciate in precedenza, inoltre, tutte le funzioni che definiscono i vincoli (di disuguaglianza) sono lineari quindi convesse ed è possibile trovare un punto $(\alpha, \alpha_0, \varepsilon)$ che soddisfi le disuguaglianze strettamente. Allora valgono Slater's contraint qualification (Teorema A.17) in ogni $(\alpha, \alpha_0, \varepsilon)$ nell'insieme convesso (sempre perchè i vincoli sono lineari vedasi corollario A.3)

$$S = \{(\alpha, \alpha_0, \varepsilon) \in \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n \;\text{ tali che }\; \varepsilon \geq 0,\;\; y_i(\alpha_0 + \{K\alpha\}_i) \geq 1 - \varepsilon_i, \;\; i = 1,\ldots,n\}$$

Di conseguenza $(\alpha, \alpha_0, \varepsilon) \in S$ è un ottimo globale per il problema convesso (2.4) se e solo se è un punto KKT (corollario A.4). Inoltre, può essere conveniente passare al problema duale. A tal proposito ricaviamo la funzione Lagrangiana associata al problema (2.4). Se chiamiamo $\mu, \lambda \in \mathbb{R}^n$ i moltiplicatori di Lagrange (non negativi) associati rispettivamente al primo e al secondo vincolo di disuguaglianza, la Lagrangiana è

$$\mathcal{L} : \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^{2n}_+ \to \mathbb{R}$$

definita da

$$\mathcal{L}(\alpha_0, \alpha, \varepsilon, \lambda, \mu) = \sum_{i=1}^{n} \varepsilon_i + \frac{\gamma}{2}\alpha^T K\alpha - \sum_{i=1}^{n} \varepsilon_i\mu_i + \sum_{i=1}^{n} \lambda_i(1 - \varepsilon_i - y_i(\alpha_0 + \{K\alpha\}_i))$$

Chiamando ora $\alpha_0 + \{K\alpha\}_i = g(x_i)$, l'espressione può essere riscritta come

$$\mathcal{L}(\alpha_0, \alpha, \varepsilon, \lambda, \mu) = \sum_{i=1}^{n} \varepsilon_i + \frac{\gamma}{2}\alpha^T K\alpha - \sum_{i=1}^{n} \varepsilon_i\mu_i + \sum_{i=1}^{n} \lambda_i(1 - \varepsilon_i - y_i g(x_i))$$

Notiamo ora che anche la funzione Lagrangiana è una funzione convessa in $(\alpha, \alpha_0, \varepsilon)$ in quanto somma di $\sum_{i=1}^{n} \varepsilon_i + \frac{\gamma}{2}\alpha^T K$ che è convessa e di funzioni lineari (quindi convesse) moltiplicate per coefficienti non negativi (λ_i, μ_i). Ora, ricordando che la funzione Lagrangiana duale è la funzione

$$\mathcal{L}_D : \mathbb{R}^n_+ \times \mathbb{R}^n_+ \to \mathbb{R}$$

definita da

$$\mathcal{L}_D = \inf_{\alpha_0,\alpha,\varepsilon} \mathcal{L}(\alpha_0, \alpha, \varepsilon, \lambda, \mu)$$

Possiamo calcolare tale funzione semplicemente ponendo a zero il gradiente rispetto a $(\alpha_0, \alpha, \epsilon)$ di $\mathcal{L}$ in quanto $\mathcal{L}$ è convessa in $(\alpha_0, \alpha, \epsilon)$. Sapendo ora che:

- $$\nabla_\alpha(\mathcal{L}(\alpha_0, \alpha, \varepsilon, \lambda, \mu)) = \nabla_\alpha(\sum_{i=1}^n \varepsilon_i + \frac{\gamma}{2}\alpha^T K\alpha - \sum_{i=1}^n \varepsilon_i\mu_i + \sum_{i=1}^n \lambda_i(1 - \varepsilon_i - y_i(\alpha_0 + \{K\alpha\}_i))) =$$
  $$= \gamma K\alpha + \nabla_\alpha(\sum_{i=1}^n \lambda_i(1 - \varepsilon_i - y_i(\alpha_0 + \{K\alpha\}_i))) = \gamma K\alpha - \sum_{i=1}^n \lambda_i y_i(\sum_{j=1}^n \nabla_\alpha(K_{ij}\alpha_j)) =$$
  $$= \gamma K\alpha - \begin{bmatrix} \sum_{i=1}^n \lambda_i y_i K_{i1} \\ \sum_{i=1}^n \lambda_i y_i K_{i2} \\ \vdots \\ \sum_{i=1}^n \lambda_i y_i K_{in} \end{bmatrix} = \begin{bmatrix} \gamma\sum_{j=1}^n K_{1j}\alpha_j - \sum_{j=1}^n \lambda_j y_j K_{j1} \\ \gamma\sum_{j=1}^n K_{2j}\alpha_j - \sum_{j=1}^n \lambda_j y_j K_{j2} \\ \vdots \\ \gamma\sum_{j=1}^n K_{nj}\alpha_j - \sum_{j=1}^n \lambda_j y_j K_{jn} \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^n K_{1j}(\gamma\alpha_j - \lambda_j y_j) \\ \sum_{j=1}^n K_{2j}(\gamma\alpha_j - \lambda_j y_j) \\ \vdots \\ \sum_{j=1}^n K_{nj}(\gamma\alpha_j - \lambda_j y_j) \end{bmatrix}$$
  $$= K(\gamma\alpha - \lambda \odot y)$$

  Dove $\odot$ indica il prodotto elemento per elemento dei due vettori.

- $$\nabla_\varepsilon(\mathcal{L}(\alpha_0, \alpha, \varepsilon, \lambda, \mu)) = \nabla_\varepsilon(\sum_{i=1}^n \varepsilon_i + \frac{\gamma}{2}\alpha^T K\alpha - \sum_{i=1}^n \varepsilon_i\mu_i + \sum_{i=1}^n \lambda_i(1 - \varepsilon_i - y_i(\alpha_0 + \{K\alpha\}_i))) =$$
  $$= \begin{bmatrix} 1 - \mu_1 - \lambda_1 \\ 1 - \mu_2 - \lambda_2 \\ \vdots \\ 1 - \mu_n - \lambda_n \end{bmatrix} = 1 - \mu - \lambda$$

- $$\nabla_{\alpha_0}(\mathcal{L}(\alpha_0, \alpha, \varepsilon, \lambda, \mu)) =$$
  $$\nabla_{\alpha_0}(\sum_{i=1}^n \varepsilon_i + \frac{\gamma}{2}\alpha^T K\alpha - \sum_{i=1}^n \varepsilon_i\mu_i + \sum_{i=1}^n \lambda_i(1 - \varepsilon_i - y_i(\alpha_0 + \{K\alpha\}_i))) =$$
  $$= -\sum_{i=1}^n \lambda_i y_i$$

Quindi

$$\nabla_{(\alpha_0,\alpha,\varepsilon)}(\mathcal{L}(\alpha_0, \alpha, \varepsilon, \lambda, \mu)) = \begin{bmatrix} -\sum_{i=1}^n \lambda_i y_i \\ K(\gamma\alpha - \lambda \odot y) \\ 1 - \mu - \lambda \end{bmatrix}$$

Ponendo il gradiente uguale a zero abbiamo che $\alpha_0, \alpha, \varepsilon$ soddisfano il seguente sistema

$$\begin{cases} \lambda^T y = 0 \\ \alpha = \frac{y \odot \lambda}{\gamma} \\ \mu = 1 - \lambda \end{cases}$$

Allora, sostituendo l'espressione di α e μ nella Lagrangiana e usando il fatto che $\lambda^T y = 0$ otteniamo

$$\begin{aligned}
&\mathcal{L}(\alpha_0, \alpha, \varepsilon, \lambda, \mu) = \sum_{i=1}^{n} \varepsilon_i + \frac{\gamma}{2}\alpha^T K\alpha - \sum_{i=1}^{n} \varepsilon_i \mu_i + \sum_{i=1}^{n} \lambda_i (1 - \varepsilon_i - y_i(\alpha_0 + \{K\alpha\}_i)) = \\
&= \sum_{i=1}^{n} \varepsilon_i + \frac{\gamma}{2}\alpha^T K\alpha - \sum_{i=1}^{n} \varepsilon_i (1 - \lambda_i) + \sum_{i=1}^{n} \lambda_i (1 - \varepsilon_i - y_i(\alpha_0 + \{K\alpha\}_i)) = \\
&= \sum_{i=1}^{n} \lambda_i \varepsilon_i + \frac{\gamma}{2}\alpha^T K\alpha + \sum_{i=1}^{n} \lambda_i (1 - \varepsilon_i - y_i(\alpha_0 + \{K\alpha\}_i)) = \\
&= \frac{\gamma}{2}\alpha^T K\alpha + \sum_{I=1}^{n} \lambda_i - \sum_{i=1}^{n} \lambda_i y_i \alpha_0 - \sum_{i=1}^{n} \lambda_i y_i \{K\alpha\}_i = \frac{\gamma}{2}\alpha^T K\alpha + \sum_{I=1}^{n} \lambda_i - \sum_{i=1}^{n} \lambda_i y_i \{K\alpha\}_i = \\
&= \sum_{i=1}^{n} \lambda_i + \frac{\gamma}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i K_{ij} \alpha_j - \sum_{i=1}^{n} \lambda_i y_i \sum_{j=1}^{n} K_{ij} \alpha_j = \\
&= \sum_{i=1}^{n} \lambda_i + \frac{\gamma}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\lambda_i y_i}{\gamma} K_{ij} \frac{\lambda_j y_j}{\gamma} - \sum_{i=1}^{n} \lambda_i y_i \sum_{j=1}^{n} K_{ij} \frac{\lambda_j y_j}{\gamma} = \sum_{i=1}^{n} \lambda_i - \frac{1}{2\gamma} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j \mathcal{K}(x_i, x_j)
\end{aligned}$$

Quindi abbiamo che la Lagrangiana duale è

$$L_D(\lambda) = \sum_{i=1}^{n} \lambda_i - \frac{1}{2\gamma} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j \mathcal{K}(x_i, x_j)$$

Il problema duale è allora

$$\max_{\lambda} \sum_{i=1}^{n} \lambda_i - \frac{1}{2\gamma} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j \mathcal{K}(x_i, x_j) \quad \text{soggetto a} \quad \left\{ \lambda^T y = 0, \;\; 0 \leq \lambda \leq 1 \right.$$

Dove il fatto che $0 \leq \lambda \leq 1$ deriva dal fatto che λ e μ sono non negativi e $\mu = 1 - \lambda$. Possiamo allora risolvere il problema duale ottenendo i moltiplicatori di Lagrange λ^*. Visto che le ipotesi del teorema di dualità forte (corollario A.5) nel caso di problemi convessi sono rispettate: il problema (2.4) ammette un minimo globale in quanto il valore della funzione obiettivo è non negativa (limitata dal basso) ed è convessa su un insieme non vuoto (definito da vincoli di disuguaglianza lineari) e esistono validi moltiplicatori di Lagrange per il primale poichè valgono i constraint qualifications qualsiasi $(\alpha_0, \alpha, \varepsilon) \in S$, allora per il corollario A.5 abbiamo che l'ottimo duale è uguale all'ottimo del problema (2.4) e i valori di $\alpha, \alpha_0, \varepsilon$ che minimizzano il primale sono i minimi del seguente problema non vincolato:

$$\min_{\alpha_0, \alpha, \varepsilon} \mathcal{L}(\alpha_0, \alpha, \varepsilon, \mu^*, \lambda^*)$$

la funzione $(\alpha_0, \alpha, \varepsilon) \to \mathcal{L}(\alpha_0, \alpha, \varepsilon, \mu^*, \lambda^*)$ è convessa quindi basterà porre il gradiente rispetto ad $\alpha_0, \alpha, \varepsilon$ uguale a zero. Otteniamo allora (stessi calcoli fatti in precedenza) $\alpha_i = \frac{\lambda_i y_i}{\gamma} \ \ \forall i \in \{1,2,\dots,n\}$ da cui abbiamo

$$g_\tau(x) = \alpha_0 + \frac{1}{\gamma}\sum_{i=1}^{n} y_i \lambda_i^* \mathcal{K}(x_i, x) \tag{2.5}$$

Come abbiamo già osservato $(\alpha_0, \alpha, \varepsilon)$ è un ottimo globale del problema (2.4) se e solo se è un punto KKT. Determiniamo allora le condizioni necessarie affinchè un punto sia KKT. Dalle KKT conditions abbiamo che devono valere le seguenti:

1. $\nabla_{(\alpha_0,\alpha,\varepsilon)} = 0 \implies \begin{cases} \lambda^{*T} y = 0 \\ \alpha = \frac{y \odot \lambda^*}{\gamma} \\ \mu^* = 1 - \lambda^* \end{cases}$
2. $\varepsilon \geq 0$
3. $y_i g_\tau(x_i) + \varepsilon_i - 1 \geq 0$
4. $\lambda^* \geq 0, \mu^* \geq 0 \implies 0 \leq \lambda^* \leq 1$
5. $$\mu_i^* \varepsilon_i = 0 \ \text{ e } \ \lambda_i^*(1 - \varepsilon_i - y_i g_\tau(x_i)) = 0 \ \ \forall i \ \in \{1,\dots,n\} \implies$$ $$(1 - \lambda^*) \odot \varepsilon = 0, \ \ \lambda_i^*(1 - \varepsilon_i - y_i g_\tau(x_i)) = 0 \ \ \forall i \ \in \{1,\dots,n\}$$

Utilizzando ora la condizione 5 con $\lambda_i^* \in (0,1)$ abbiamo che $\alpha_0 = y_j - \frac{1}{\gamma}\sum_{i=1}^{n} y_i \lambda_i \mathcal{K}(x_i, x_j)$ per ogni j tale che $\lambda_j \in (0,1)$.

Osservazione 4. *Per evitare possibili problemi numerici generalmente il coefficiente α_0 viene calcolato tramite una media su tutti gli indici j tali per cui $\lambda_j \in (0,1)$ di $y_j - \frac{1}{\gamma} y_i \lambda_i \mathcal{K}(x_i, x_j)$. In formule:*

$$\alpha_0 = \frac{1}{|\mathcal{T}|}\left[\sum_{j \in \mathcal{T}} y_j - \sum_{i=1}^{n} y_i \lambda_i \mathcal{K}(x_i, x)\right]$$

Dove $\mathcal{T} := \{j : \lambda_j \in (0,1)\}$.

Osservazione 5. *Dalla formula (2.5) notiamo che $g_\tau(x)$ dipende unicamente dai vettori x_i per i quali $\lambda_i^* \neq 0$, questi vettori sono chiamati vettori di supporto della Support Vector Machine.*

Notiamo che il problema duale può essere riscritto in maniera differente se definiamo $\nu_i = \frac{\lambda_i}{\gamma}$ infatti, il problema duale

$$\max_{\lambda} \sum_{i=1}^{n} \lambda_i - \frac{1}{2\gamma}\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j \mathcal{K}(x_i, x_j) \quad \text{soggetto a} \quad \{\lambda^T y = 0, \ \ 0 \leq \lambda \leq 1\}$$

è equivalente al seguente problema

$$\min_{\lambda} \frac{1}{2\gamma} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j \mathcal{K}(x_i, x_j) - \sum_{i=1}^{n} \lambda_i \quad \text{soggetto a} \quad \{\lambda^T y = 0,\ 0 \leq \lambda \leq 1\}$$

Usando ora la definizione di ν_i, il problema diventa

$$\min_{\nu} \frac{\gamma}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \nu_i \nu_j y_i y_j \mathcal{K}(x_i, x_j) - \gamma \sum_{i=1}^{n} \nu_i \quad \text{soggetto a} \quad \{\gamma \nu^T y = 0,\ 0 \leq \nu \leq \frac{1}{\gamma}\}$$

Visto che minimizziamo rispetto a ν e γ è una costante, ciò è equivalente al seguente problema

$$\min_{\nu} \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \nu_i \nu_j y_i y_j \mathcal{K}(x_i, x_j) - \sum_{i=1}^{n} \nu_i \quad \text{soggetto a} \quad \{\nu^T y = 0,\ 0 \leq \nu \leq \frac{1}{\gamma} = C\}$$

2.3 Kernel lineare e interpretazione geometrica

Ricordiamo che il Kernel lineare è quell'applicazione $\mathcal{K} : X \times X \to \mathbb{R}$ definita da $\mathcal{K}(x, x') = x^T x'$. Dalla formula (2.5) abbiamo allora che

$$g_\tau(x) = \alpha_0 + \frac{1}{\gamma} \sum_{i=1}^{n} y_i \lambda_i^* \mathcal{K}(x_i, x) = \alpha_0 + \frac{1}{\gamma} \sum_{i=1}^{n} y_i \lambda_i^* x_i^T x = \beta_0 + \beta^T x$$

Dove $\beta_0 = \alpha_0$ e $\beta = \frac{1}{\gamma} \sum_{i=1}^{n} \lambda_i^* y_i x_i = \sum_{i=1}^{n} \alpha_i x_i$. Come conseguenza, il bordo decisionale cioè l'insieme delle $\{x : g_\tau(x) = 0\}$ è un iperpiano affine in $\mathbb{R}^n$. Inoltre i due insiemi $\{x : g_\tau(x) = -1\}$ e $\{x : g_\tau(x) = 1\}$ sono chiamati margini.

Proposizione 2.1. *La distanza del margine dal bordo decisionale è* $\frac{1}{||\beta||}$.

Dimostrazione. Siano $x_1 \in \{x : g_\tau(x) = 1\}$ e $x_2 \in \{x : g_\tau(x) = 0\}$ tali per cui il vettore congiungente x_1 e x_2 (chiamiamo tale vettore v) sia ortogonale al bordo decisionale. Ne segue che $v \parallel \beta$. Da definizione, $\beta_0 + \beta^T x_1 = 1$ e $\beta_0 + \beta^T x_2 = 0$. Sottraendo le due espressioni otteniamo $\beta^T (x_1 - x_2) = 1$, considerando il valore assoluto di entrambi i membri abbiamo $|\beta^T (x_1 - x_2)| = 1$. Sfruttando infine il parallelismo tra β e $(x_1 - x_2)$ abbiamo $||\beta^T|| \cdot ||(x_1 - x_2)|| = 1$ cioè $||x_1 - x_2|| = \frac{1}{||\beta||}$. □

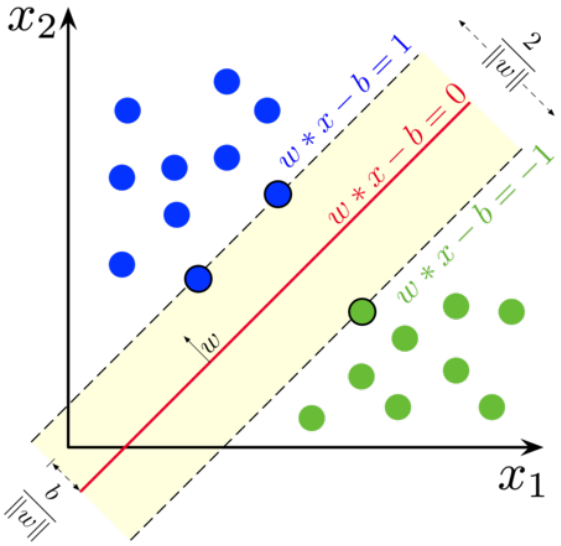


Figura 2.1: Classificazione binaria (blu e verde) tramite Support Vector Machine con kernel lineare

La seguente figura riassume la situazione.

Possiamo ora dividere il training set $\{(x_i, y_i)\}$ basandoci sul valore dei moltiplicatori λ_i^*:

- I punti tali per cui $\lambda_i^* \in (0,1)$ sono punti che giacciono esattamente sul margine. Questi punti sono chiamati vettori di supporto e sono correttamente classificati.
- I punti tali per cui $\lambda_i^* = 1$ sono punti che giacciono all'interno del margine. Sono anche essi chiamati vettori di supporto ma tali punti possono o meno essere classificati correttamente.
- I punti tali per cui $\lambda_i^* = 0$ sono punti che giacciono al di fuori del margine. Questi punti sono chiamati vettori non di supporto e sono correttamente classificati.

Mostriamo ora perchè tale classificazione vale.

Dimostrazione. Prima di tutto notiamo che se λ^* è soluzione del problema duale, allora $(\alpha_0, \alpha, \varepsilon)$ che minimizzano il primale sono i minimi della funzione $(\alpha_0, \alpha, \varepsilon) \to \mathcal{L}(\alpha_0, \alpha, \varepsilon, \mu^*, \lambda^*)$, in particolare allora $\nabla_{(\alpha_0,\alpha,\varepsilon)}\mathcal{L}(\alpha_0, \alpha, \varepsilon, \mu^*, \lambda^*) = 0$ il che implica che $\mu^* = 1 - \lambda^*$. Inoltre, visto che vale un constraint qualification per il problema primale (2.4) su tutto S, allora un minimo globale per il primale è tale se e solo se è un punto KKT (corollario A.4). Occupiamoci allora di determinare le condizioni sotto le quali un punto $(\alpha_0, \alpha, \varepsilon)$ è un KKT point per il primale. Dalle KKT conditions enunciate in precedenza abbiamo che

- Se $\lambda_i^* \in (0,1)$ dalla condizione 5) dei punti KKT abbiamo che $\varepsilon_i = 0$ e quindi $1 - y_i g(x_i) = 0$ cioè $y_i g(x_i) = 1$ cioè il punto è sul margine.

- $g(x_i) = 1$ cioè il punto è sul margine. Nel caso in cui $\lambda_i^* = 1$ dalla condizione 5) abbiamo che $(1 - \varepsilon_i - y_i g_\tau(x_i)) = 0$ il che implica che $y_i g_\tau(x_i) = 1 - \varepsilon_i \leq 1$ cioè il punto giace all'interno del margine.
- Infine se $\lambda_i^* = 0$, abbiamo dalla 5) che $\varepsilon_i = 0$ da cui dalla 3) abbiamo $y_i g_\tau(x_i) \geq 1$ cioè il punto giace al di fuori del margine.

□

2.4 Hard margin vs Soft Margin Support Vector Machine

Nel caso di Kernel lineare il problema (2.4) si riscrive come

$$\min_{\alpha,\alpha_0,\varepsilon} \sum_{i=1}^{n} \varepsilon_i + \frac{\gamma}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i K(x_i, x_j) \alpha_j \quad \text{soggetto a} \quad \begin{cases} \varepsilon \geq 0 \\ y_i(\alpha_0 + \sum_{j=1}^{n} K(x_i, x_j)\alpha_j) \geq 1 - \varepsilon_i, \ \ \forall i \end{cases}$$

$$\min_{\alpha,\alpha_0,\varepsilon} \sum_{i=1}^{n} \varepsilon_i + \frac{\gamma}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i x_i^T x_j \alpha_j \quad \text{soggetto a} \quad \begin{cases} \varepsilon \geq 0 \\ y_i(\alpha_0 + \sum_{j=1}^{n} x_i^T x_j \alpha_j) \geq 1 - \varepsilon_i, \ \ \forall i \end{cases}$$

Usando ora il fatto che $\beta = \sum_{i=1}^{n} \alpha_i x_i = X^T \alpha$, possiamo riscrivere il problema come

$$\min_{\alpha,\alpha_0,\varepsilon} \sum_{i=1}^{n} \varepsilon_i + \frac{\gamma}{2} \alpha^T X X^T \alpha \quad \text{soggetto a} \quad \begin{cases} \varepsilon \geq 0 \\ y_i(\alpha_0 + x_i^T \beta) \geq 1 - \varepsilon_i, \ \ \forall i \end{cases}$$

Finalmente, riscrivendo $\alpha^T X X^T \alpha = \alpha^T X \beta = ||\beta||^2$ e $\alpha_0 = \beta_0$ il problema diventa

$$\min_{\beta,\beta_0,\varepsilon} \sum_{i=1}^{n} \varepsilon_i + \frac{\gamma}{2} ||\beta||^2 \quad \text{soggetto a} \quad \begin{cases} \varepsilon \geq 0 \\ y_i(\beta_0 + x_i^T \beta) \geq 1 - \varepsilon_i, \ \ \forall i \end{cases} \tag{2.6}$$

Questo problema prende il nome di *Soft Margin Support Vector Machine* e ha una facile interpretazione geometrica. Il problema di minimizzazione cerca di minimizzare essenzialmente due quantità: $\frac{\gamma}{2}||\beta||^2$ e $\sum_{i=1}^{n} \varepsilon_i$. Notiamo che la prima quantità è l'inverso della distanza del margine dal bordo decisionale (moltiplicata per una costante di regolarizzazione) quindi l'algoritmo cerca di massimizzare il margine. La seconda quantità è $\sum_{i=1}^{n} \varepsilon_i$. Il significato di ε_i è intuibile dalla figura 2.2 e dai vincoli a cui è sottoposto il problema di minimizzazione. Se supponessimo che il vincolo $y_i(\alpha_0 + x_i^T \beta) = y_i g_\tau(x_i) \geq 1 - \varepsilon_i$ non presenti il termine ε_i (cioè se $\varepsilon_i = 0$), allora il vincolo diventa $y_i g_\tau(x_i) \geq 1$, cioè chiederemmo che $g_\tau(x_i) \geq 1$ quando $y_i = 1$ e $g_\tau(x_i) \leq -1$ quando $y_i = -1$. Ciò significa che staremmo vincolando ogni x_i a giacere fuori dal margine e ad essere correttamente classificato. Ciò è possibile unicamente nel caso di dataset linearmente separabili (dataset che possono essere separati tramite un iperpiano), nel caso il dataset non lo fosse, il problema non avrebbe allora soluzione nello scenario sopra menzionato. Tuttavia, se le variabili di slack ε_i possono essere diverse da zero, come si può vedere dalla figura 2.2, permettiamo ad alcuni x_i di cadere

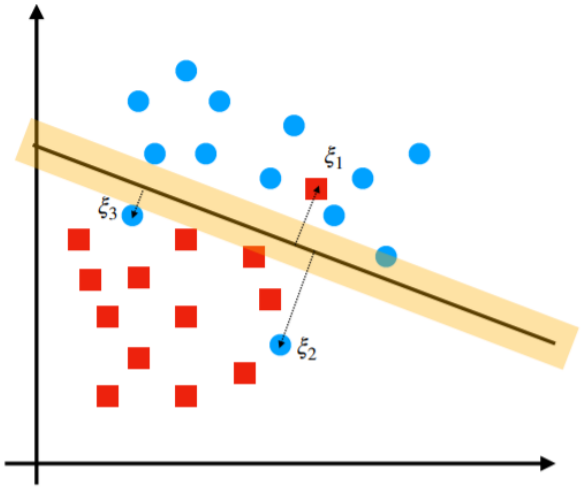


Figura 2.2: Soft Margin Support Vector Machine

all'interno del margine e addirittura nel semipiano sbagliato ma minimizzando $\sum_{i=1}^{n} \varepsilon_i$ nella funzione obiettivo e vincolando ad avere $\varepsilon_i \geq 0 \ \ \forall i = 1,2,\ldots,n$ cerchiamo soluzioni in cui ciò capiti il meno possibile. In tal modo, l'iperpiano trovato risolvendo il problema (2.6) è tale per cui il suo margine è massimizzato e classifica la maggior parte dei punti correttamente e fuori dal margine come si può vedere dalla figura 2.2.

Nel caso i punti $\{x_i : y_i = 1\}$ e $\{x_i : y_i = -1\}$ siano perfettamente separabili da un iperpiano, non ci sono punti strettamente all'interno del margine e tutti i vettori di supporto giacciono sul margine. Questo specifico problema di classificazione prende il nome di *Hard Margin Support Vector Machine* e il problema (2.4), sotto queste ipotesi, può essere riformulato nel seguente modo

$$\min_{\beta,\beta_0} ||\beta||^2 \ \text{ soggetto a } \ y_i(\beta_0 + x_i^T\beta) \geq 1 \ \ \forall i = 1,2,\ldots,n$$

Dove abbiamo usato che, visto che tutti i punti giacciono fuori o sul margine, abbiamo (come conseguenza della dimostrazione di pagina 37) che $\varepsilon_i = 0 \ \ \forall i \in 1,2,\ldots,n$ e inoltre, usando ancora il fatto che $\alpha_0 = \beta_0$ e $K\alpha = XX^T\alpha = X\beta$ e dunque $\alpha^T K\alpha = \alpha^T X\beta = ||\beta||^2$. Tale problema può essere riscritto in maniera equivalente nel seguente modo

$$\max_{\beta,\beta_0} \frac{1}{||\beta||} \ \text{ soggetto a } \ y_i(\beta_0 + x_i^T\beta) \geq 1 \ \ \forall i = 1,2,\ldots,n \tag{2.7}$$

In tal caso, visto che $\frac{1}{||\beta||}$ è metà della lunghezza del margine questo problema di ottimizzazione può essere facilmente interpretato: il problema cerca di separare i punti attraverso un iperpiano tale per cui il margine sia massimizzato.

2.5 Applicazione Support Vector Machine ad un dataset

Consideriamo dapprima il caso di un dataset linearmente separabile (in $\mathbb{R}^2$) come quello nella figura (2.3).
In tal caso, quanto studiato in precedenza ci assicura che, utilizzando un kernel linea-

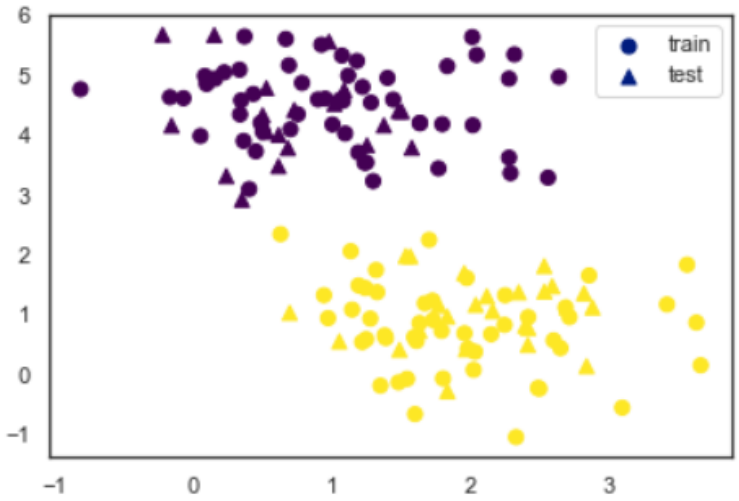


Figura 2.3: Dataset linearmente separabile, in blu gli elementi di una classe, in giallo gli elementi dell'altra

re, risolvendo il problema di ottimizzazione (2.7) riusciamo a trovare un iperpiano (in tal caso una retta) che separa i punti delle due classi. La figura (2.4) mostra il bordo decisionale e i due margini risultanti dall'applicazione dell'algoritmo di Support Vector Machine al dataset. Inoltre, un semplice calcolo dell'accuracy ci mostra che essa è pari al 100% sia nel caso del training che nel caso del test set. Cioè il dataset (come intuibile dalla figura) é linearmente separabile.

Consideriamo ora il caso in cui il dataset non sia separabile (in $\mathbb{R}^2$) attraverso un iperpiano come in figura (2.5). In tal caso qualsiasi retta non sarebbe in grado di separare gli elementi delle due classi. Tuttavia è possibile separare tali elementi in $\mathbb{R}^3$ considerando un nuovo vettore di features $z = [z_1, z_2, z_3]^T = [x_1, x_2, x_1^2 + x_2^2]^T$. Infatti z_3 è una misura della distanza di ogni punto (x_1, x_2) dall'origine e questa informazione risulta essere di vitale importanza per distinguere le due classi. Inoltre, per ogni $(x_1, x_2) \in \mathbb{R}^2$, il vettore delle features z corrispondente giace su un paraboloide come illustrato nella figura (2.6). Sempre dalla stessa figura è facile notare come ora i punti siano linearmente separabili da un piano in $\mathbb{R}^3$.

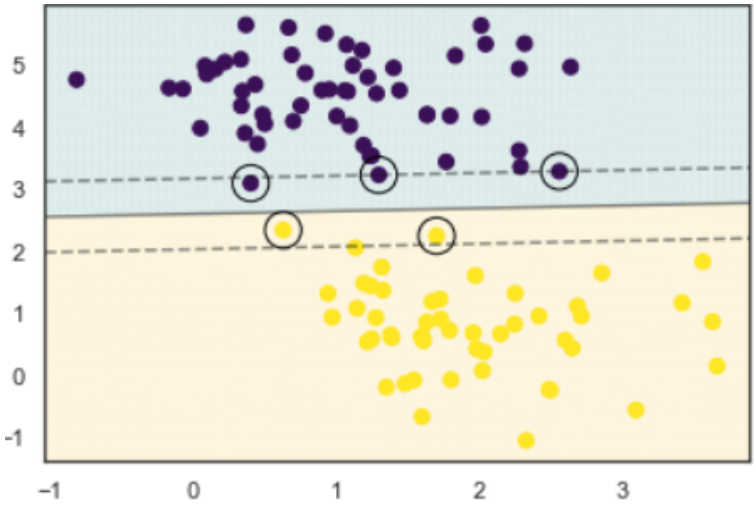


Figura 2.4: Bordo decisionale (linea continua) e margini (tratteggiati)

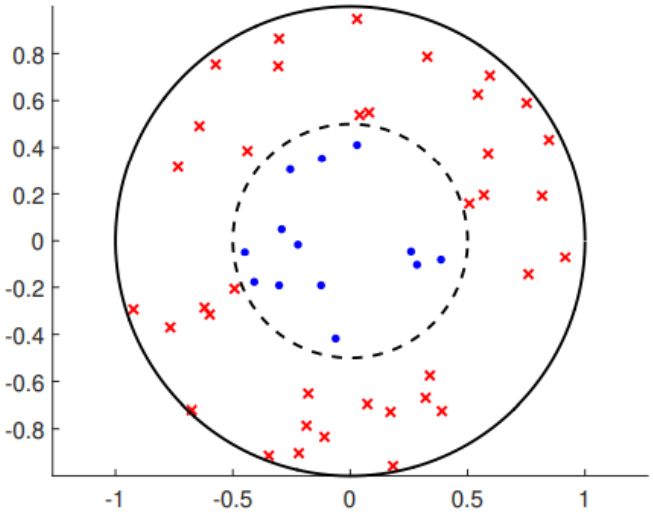


Figura 2.5: Dataset non linearmente separabile, in rosso gli elementi di una classe, in blu gli elementi dell'altra

Applicando ora l'algoritmo di Support Vector Machine ai nuovi punti in $\mathbb{R}^3$, otteniamo che il piano che separa i dati delle due classi è

$$\pi : 5.6179 - 0.9128z_1 + 0.8917z_2 = 24.2764z_3$$

Possiamo ora trovare il bordo decisionale in $\mathbb{R}^2$ semplicemente intersecando tale piano con il paraboloide $z_3 = z_1^2 + z_2^2$ e proiettando sul piano xy. La figura (2.7) mostra

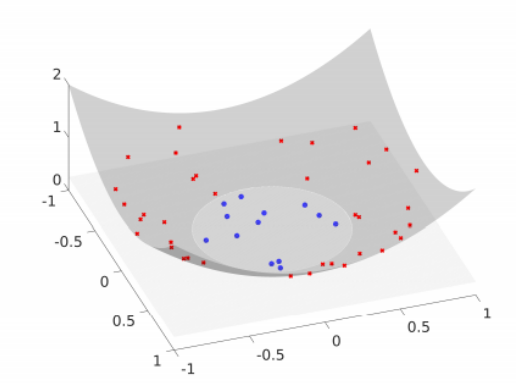


Figura 2.6: In $\mathbb{R}^3$ i punti possono essere separati linearmente

il risultato. Il processo di introdurre nuove features al vettore delle features prende il

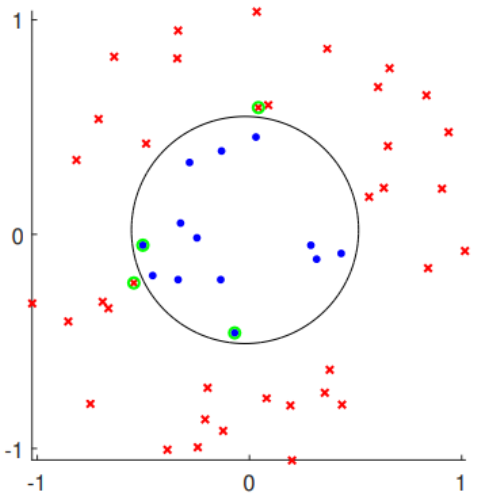


Figura 2.7: Bordo decisionale dataset non linearmente separabile

nome di *feature's expansion* e può risultare molto utile per superare il problema della non linearità del dataset.
Un'alternativa è quella di utilizzare un kernel diverso dal kernel lineare. Nell'esempio del dataset della figura (2.5) possiamo infatti considerare la feature map $\phi(x) = [x_1, x_2, x_1^2 + x_2^2]^T$ in $\mathbb{R}^2$ che definisce il seguente reproducing kernel

$$\mathcal{K}(x, x') = \phi(x)^T \phi(x')$$

il quale origina un unico RKHS $\mathcal{H}$. La funzione di predizione è in tal caso

$$g_\tau(x) = \alpha_0 + \frac{1}{\gamma} \sum_{i=1}^{n} y_i \lambda_i \phi(x_i)^T \phi(x) = \beta_0 + \beta^T \phi(x)$$

Dove $\alpha_0 = \beta_0$ e $\beta = \frac{1}{\gamma}\sum_{i=1}^{n} y_i\lambda_i\phi(x_i)$. In questo caso il bordo decisionale risulta essere esattamente lo stesso mostrato nella figura (2.7).
Una scelta popolare nel trattare dataset non linearmente separabili è quello del kernel gaussiano. Come abbiamo già visto nel capitolo uno, il kernel gaussiano è definito da $\mathcal{K}(x, x') = e^{-c||x-x'||^2}$ dove c è un hyperparameter cioè una costante da determinare per tentativi. Dal teorema di rappresentazione abbiamo che

$$g_\tau(x) = \alpha_0 + \sum_{i=1}^{n} \alpha_i e^{-c||x_i-x||^2}$$

Applicando ora l'algoritmo di Support Vector Machine con kernel gaussiano al dataset non linearmente separabile otteniamo il bordo decisionale in figura (2.8). Come si può

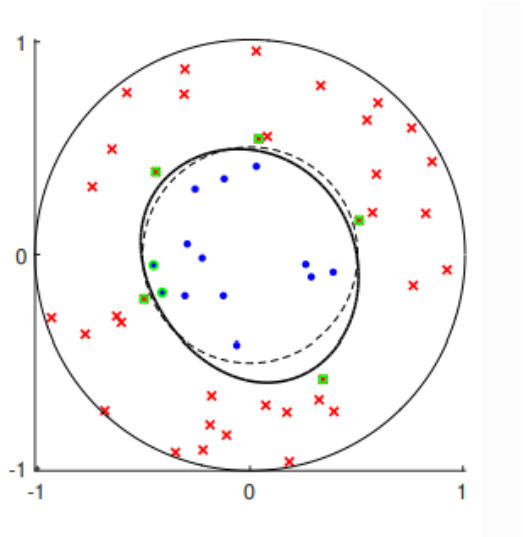


Figura 2.8: Bordo decisionale originato da kernel gaussiano

osservare in tal caso il bordo decisionale non è esattamente una circonferenza. Tuttavia, anche se il bordo decisionale ha un contorno più complesso, ogni elemento del training set risulta anche in questo caso essere classificato correttamente.

Appendice A

Risultati di ottimizzazione vincolata

A.1 Punti stazionari per ottimizzazione vincolata

Definizione A.1. *Data una funzione $f : \mathcal{E} \to \mathbb{R}$ dove $\mathcal{E}$ è uno spazio euclideo (spazio di Hilbert di dimensione finita) e $S \subseteq \mathcal{E}$ (S si assume chiuso e non vuoto), il problema*

$$\min_{x \in S} f(x)$$

si definisce un problema di ottimizzazione vincolato.

Definizione A.2. *Un minimo globale di $f : S \to \mathbb{R}$ è un punto $x^* \in S$ tale che $f(x) \geq f(x^*)$ per ogni $x \in S$.*

Definizione A.3. *Un insieme $U \subseteq S$ è aperto in S se esiste un aperto V in $\mathcal{E}$ tale che $U = S \cap V$.*

Definizione A.4. *Un intorno di $x \in S$ è un aperto di S contenente x.*

Definizione A.5. *Un minimo locale di $f : S \to \mathbb{R}$ è un punto $x^* \in S$ tale che $f(x) \geq f(x^*)$ per ogni x in un intorno di $x^* \in S$.*

Definizione A.6. *Un sottoinsieme $K \subseteq \mathcal{E}$ è un cono se per ogni $v \in K$ e $\alpha > 0$ abbiamo $\alpha v \in K$.*

Definizione A.7. *Considerando $x \in S$, il vettore $v \in \mathcal{E}$ è una direzione tangente ad x per S se esiste una sequenza $(x_k)_{k \geq 0}$ di punti in S e una sequenza di reali positivi t_k tali che:*

1. $\lim_{k \to \infty} x_k = x$
2. $\lim_{k \to \infty} t_k = 0$
3. $\lim_{k \to \infty} \frac{x_k - x}{t_k} = v$

L'insieme T_xS di tutte le direzioni tangenti ad x per S è chiamato il cono tangente per S ad x.

Teorema A.1. *Il cono tangente è un cono chiuso.*

Dimostrazione. Mostriamo che il cono tangente è proprio un cono. Sia $v \in T_xS$ da definizione esiste una sequenza $(x_k)_{k\geq 0} \subseteq S$ convergente ad x e una sequenza di reali positivi $(t_k)_{k\geq 0}$ convergente a zero tale per cui $v = \lim_{k\to\infty} \frac{x_k - x}{t_k}$. Considerando ora αv abbiamo $\alpha v \in T_xS$, sarà infatti sufficiente prendere la medesima successione $(x_k)_{k\geq 0}$ e $t'_k = t_k/\alpha$, in questo modo rimane vero che $\lim_{k\to\infty} t'_k = 0$ ma ora $\alpha v = \lim_{k\to\infty} \alpha \frac{x_k - x}{t_k} = \lim_{k\to\infty} \frac{x_k - x}{t'_k}$. Omettiamo per semplicità la dimostrazione della chiusura di T_xS. □

Teorema A.2. *Dato $x \in S$ consideriamo una curva continua $c : [0, \epsilon] \to \mathcal{E}$ con $\epsilon > 0$ tale che $c(0) = x$ e $c(t) \in S$ per ogni t e tale per cui $c'(0)$ sia ben definita, allora $c'(0) \in T_xS$.*

Dimostrazione. Da ipotesi $c'(0) = \lim_{t\to 0^+} \frac{c(t)-c(0)}{t}$ è ben definita. Considerando ora la sequenza $t_k = \epsilon/k$ e $x_k = c(t_k)$, notiamo che $t_k > 0$ e $x_k \in S$ per ogni $k \in \mathbb{N}$, inoltre $t_k \to 0^+$, $x_k \to x$ e

$$\lim_{k\to\infty} \frac{x_k - x}{t_k} = \lim_{k\to\infty} \frac{c(t_k) - c(0)}{t_k} = \lim_{t\to 0^+} \frac{c(t) - c(0)}{t} = c'(0)$$

Quindi, $c'(0) \in T_xS$. □

Lemma A.1. *Sia $x \in S$ e si consideri una direzione tangente $v \in T_xS$. Sia ancora (x_k) una sequenza in S convergente ad x e sia (t_k) una sequenza di reali positivi convergente a zero tali per cui $v = \lim_{k\to\infty} \frac{x_k - x}{t_k}$. Allora se $f : \mathcal{E} \to \mathbb{R}$ è differenziabile in x, vale che*

$$Df(x)[v] = \lim_{k\to\infty} \frac{f(x_k) - f(x)}{t_k}$$

Dimostrazione. Definiamo una sequenza di vettori (e_k) in $\mathcal{E}$ tali che soddisfino

$$\frac{x_k - x}{t_k} = v + e_k$$

Ovviamente abbiamo che $\lim_{k\to\infty} e_k = 0$. Possiamo ora scrivere il seguente sviluppo di Taylor:

$$\begin{aligned} f(x_k) = f(x + t_k \frac{x_k - x}{t_k}) &= f(x + t_k(v + e_k)) = \\ &= f(x) + t_k Df(x)[v + e_k] + O(t_k^2 ||v + e_k||^2) = \\ &= f(x) + t_k Df(x)[v] + t_k Df(x)[e_k] + O(t_k^2 ||v + e_k||^2) \end{aligned}$$

Quindi

$$\frac{f(x_k) - f(x)}{t_k} = Df(x)[v] + Df(x)[e_k] + O(t_k ||v + e_k||^2)$$

Prendendo il limite per k che tende ad infinito ad entrambi i membri dell'equazione otteniamo che visto che e_k e t_k tendono a zero, abbiamo che $Df(x)[e_k], O(t_k||v+e_k||^2)$ tendono a zero e quindi

$$\lim_{k\to\infty} \frac{f(x_k)-f(x)}{t_k} = Df(x)[v]$$

□

Teorema A.3. *Sia $f : \mathcal{E} \to \mathbb{R}$ differenziabile e $S \subseteq \mathcal{E}$. Se $x^* \in S$ è un minimo locale per f vincolata a S, allora $Df(x^*)[v] \geq 0$ qualsiasi $v \in T_x S$.*

Dimostrazione. Per ipotesi, visto che x^* è un minimo locale, esiste un intorno U di x^* in S tale che $f(x) \geq f(x^*)$ per ogni $x \in U$. Assumiamo per contraddizione che $Df(x^*)[v] < 0$ per un certo $v \in T_{x^*}S$, siano (x_k) e (t_k) le sequenze associate a $v \in T_{x^*}S$ come da definizione di direzione tangente, dal lemma A1 abbiamo che :

$$\lim_{k\to\infty} \frac{f(x_k)-f(x^*)}{t_k} = Df(x^*)[v] < 0$$

Ma allora se chiamiamo $\epsilon = -\frac{1}{2}Df(x^*)[v] > 0$, abbiamo che esiste $K \in \mathbb{N}$ tale che

$$\frac{f(x_k)-f(x^*)}{t_k} \leq -\epsilon \quad \forall k \geq K$$

Da cui

$$\forall k \geq K \quad f(x_k) \leq f(x^*) - t_k\epsilon < f(x^*)$$

Questa è una contraddizione perchè visto che (x_k) converge a x^*, la sequenza dovrà entrare prima o poi nell'intorno U ma sappiamo per ipotesi che $\forall x \in S$ abbiamo $f(x) \geq f(x^*)$. □

La condizione data dal teorema A.3 è di fondamentale importanza, i punti che soddisfano $Df(x)[v]$ per ogni $v \in T_x S$ saranno detti punti stazionari e tra di essi ci saranno i minimi locali di f vincolata ad S.

Definizione A.8. *Sia un cono C in $\mathcal{E}$, il cono duale di C è l'insieme*

$$C^* = \{w \in \mathcal{E} : \langle w, v\rangle \geq 0 \ \textit{ per ogni } \ v \in C\}$$

Il cono polare di C è l'insieme

$$C^\circ = \{w \in \mathcal{E} : \langle w, v\rangle \leq 0 \ \textit{ per ogni } \ v \in C\}$$

Allora le seguenti condizioni sono equivalenti:

1. $Df(x^*)[v] \geq 0$ per ogni $v \in T_{x^*}S$
2. $\nabla f(x^*) \in (T_{x^*}S)^*$

3. $-\nabla f(x^*) \in (T_{x^*}S)^\circ$

Definizione A.9. *Se x^* soddisfa una delle equivalenti condizioni riportate sopra, chiamiamo x^* critico o punto stazionario per il problema di minimizzazione di f vincolato ad S.*

Ogni minimo locale di f vincolata ad S è un punto stazionario per il problema di minimizzazione vincolato (dal teorema A.3).

Definizione A.10. *Il cono normale ad S in $x \in S$ è l'insieme $N_xS = (T_xS)^\circ$.*

A.2 Ottimizzazione vincolata nel caso in cui S sia definito da vincoli di uguaglianza e disuguaglianza

Una particolare classe di problemi è quella in cui l'insieme S sia definito da vincoli di uguaglianza e disuguaglianza cioè il problema è il seguente:

$$\begin{aligned} &\min_{x\in\mathcal{E}} f(x) \\ \text{vincolato a :}\quad & \\ & h_i(x) = 0 \ \ \forall i \in 1,2,\cdots,p \\ & g_i(x) = 0 \ \ \forall i \in 1,2,\cdots,m \end{aligned} \tag{A.1}$$

Dove le funzioni $h_1,\dots h_p : \mathcal{E} \to \mathbb{R}$ e $g_1,\dots g_m : \mathcal{E} \to \mathbb{R}$ sono supposte C^1. In maniera più compatta possiamo anche scrivere $S = \{x \in \mathcal{E} : h(x) = 0 \ \text{ and } \ g(x) \leq 0\}$ dove $h : \mathcal{E} \to \mathbb{R}^p$ e $g : \mathcal{E} \to \mathbb{R}^m$ sono definite da

$$h(x) = \begin{bmatrix} h_1(x) \\ h_2(x) \\ \vdots \\ h_p(x) \end{bmatrix} \ \text{ e } \ g(x) = \begin{bmatrix} g_1(x) \\ g_2(x) \\ \vdots \\ g_m(x) \end{bmatrix}$$

e la notazione $u \leq 0$ per un vettore $u \in \mathbb{R}^m$ significa $u_i \leq 0$ per $i = 1,2,\dots,m$.

Ci possiamo ora chiedere se, nel caso in cui S sia definito da vincoli di uguaglianza e disuguaglianza, possiamo trovare delle particolari espressioni per T_xS e N_xS e di conseguenza delle condizioni più esplicite per trovare i punti stazionari di f vincolata ad S. A tal proposito definiamo il cono delle direzioni possibili linearizzate.

Definizione A.11. *Consideriamo S definito da vincoli di uguaglianza e disuguaglianza come in (A.1). Sia $x \in S$, il cono delle direzioni possibili linearizzate (cone of linearized feasible directions) ad x in S rispetto ai vincoli dati è l'insieme*

$$\begin{aligned} F_xS = \{v \in \mathcal{E} :& \langle \nabla h_i(x), v\rangle = 0, \ i = 1,\dots,p \quad e \\ & \langle \nabla g_i(x), v\rangle \leq 0, \ \ \textit{per ogni } \ i = 1,\dots,m \ \textit{ tali che } \ g_i(x) = 0\} \end{aligned}$$

Si potrebbe mostrare che anch'esso è un cono chiuso.

Teorema A.4. *Il cono tangente è sempre incluso nel cono delle direzioni possibili linearizzate, in formule:*

$$T_x S \subseteq F_x S$$

Dimostrazione. Consideriamo $v \in T_x S$. Per definizione esiste una sequenza $(x_k) \subseteq S$ e una sequenza di reali positivi (t_k) tali che $x_k \to x$, $t_k \to 0^+$ e $\frac{x_k - x}{t_k} \to v$. Per ogni $1 \leq i \leq p$ abbiamo $h_i(x) = 0$ e $h_i(x_k) = 0$ perchè $x \in S$ e $(x_k) \subseteq S$. Dal lemma A.1 ne segue allora che

$$\langle \nabla h_i(x), v \rangle = Dh_i(x)[v] = \lim_{k \to \infty} \frac{h_i(x_k) - h_i(x)}{t_k} = 0$$

In maniera analoga, per $1 \leq i \leq m$ tali per cui $g_i(x) = 0$, abbiamo

$$\langle \nabla g_i(x), v \rangle = Dg_i(x)[v] = \lim_{k \to \infty} \frac{g_i(x_k) - g_i(x)}{t_k} \leq 0$$

poichè $g_i(x_k) \leq 0$ per ogni $k \in \mathbb{N}$. □

Da quanto visto in precedenza sappiamo che

$$x^* \text{ è un minimo locale } \implies -\nabla f(x^*) \in (T_{x^*} S)^\circ$$

inoltre

$$T_x S \subseteq F_x S \implies (F_x S)^\circ \subseteq (T_x S)^\circ$$

dove l'ultima implicazione segue immediatamente dalla definizione di cono polare. Allora, in generale, non possiamo concludere nulla di utile riguardante il legame tra $-\nabla f(x^*)$ e $(F_{x^*})^\circ$. Tuttavia possiamo fare ciò quando $T_{x^*} S = F_{x^*} S$ cioè quando valgono le così dette *constraint qualification conditions* (CQ).

Definizione A.12. *Dato un insieme S definito da vincoli di uguaglianza e disuguaglianza come in A.1, diciamo che valgono le constraint qualification conditions (CQ) in $x \in S$ se $T_x S = F_x S$.*

Esistono diverse (CQ), qui di sotto ne riporteremo le più importanti. Indicheremo con $\mathcal{I}(x)$ l'insieme degli indici per cui i vincoli di disuguaglianza sono attivi, cioè

$$\mathcal{I}(x) = \{i \in \{1, \ldots, m\} : g_i(x) = 0\}$$

Teorema A.5. *Sia $x \in S$ dove S è definita da vincoli di uguaglianza e disuguaglianza, se:*

$$\nabla h_1(x), \ldots, \nabla h_p(x) \;\; e \;\; \nabla g_i(x) \;\; \forall i \in \mathcal{I}(x)$$

sono vettori linearmente indipendenti in $\mathcal{E}$, allora $T_x S = F_x S$ e diciamo che (LICQ) (linear independence constraint qualification condition) vale in x.

Teorema A.6. *Sia $x \in S$ dove S è definito da vincoli di uguaglianza e disuguaglianza, se $h_1, \dots, h_p$ e g_i con $i \in \mathcal{I}(x)$ sono funzioni affini, allora $T_x S = F_x S$.*

Teorema A.7. *Sia $x \in S$ dove S è definito da vincoli di uguaglianza e disuguaglianza. Se i gradienti $\nabla h_1(x), \dots, \nabla h_p(x)$ sono linearmente indipendenti e esiste $\bar{x} \in \mathcal{E}$ tale per cui*

$$\langle \nabla h_i(x), \bar{x} - x \rangle = 0 \ \ \forall i = 1, \dots, p$$
$$\langle \nabla g_i(x), \bar{x} - x \rangle < 0 \ \ \forall i \in \mathcal{I}(x)$$

allora $T_x S = F_x S$ e diciamo che Mangasarian-Fromowitz constraint qualification conditions (MFCQ) valgono in x.

Quando vale uno tra i (CQ) in $x^* \in S$, abbiamo $F_{x^*} S = T_{x^*} S$, allora

$$x^* \text{ è un minimo locale } \implies -\nabla f(x^*) \in (F_{x^*} S)^\circ$$

Può allora risultare utile il seguente teorema

Teorema A.8. *Il polare di $F_x S$ è*

$$(F_x S)^\circ = \{ \sum_{i=1}^{p} \mu_i \nabla h_i(x) + \sum_{i \in \mathcal{I}(x)} \lambda_i \nabla g_i(x) : \mu_i \in \mathbb{R}, \lambda_i \geq 0 \}$$

A.3 Punti KKT

Definizione A.13. *Un punto $x \in S$ è un punto KKT per il problema di minimizzazione vincolato con S definito da vincoli di uguaglianza e disuguaglianza (vedasi A.1) se $x \in S$ e esistono moltiplicatori di Lagrange $\mu \in \mathbb{R}^p$ e $\lambda \in \mathbb{R}^m$ con $\lambda \geq 0$ che soddisfano le condizioni KKT:*

$$-\nabla f(x) = \sum_{i=1}^{p} \mu_i \nabla h_i(x) + \sum_{i=1}^{m} \lambda_i \nabla g_i(x) \ \ e \ \ \lambda_i g_i(x) = 0 \ \ \forall i = 1, \dots, m$$

La seconda condizione è chiamata condizione di complementarietà.

Lemma A.2. *Un punto $x \in S$ è KKT se e solo se $-\nabla f(x) \in (F_x S)^\circ$.*

Dimostrazione. La condizione di complementarietà ci dice che se $g_i(x) \neq 0$ allora $\lambda_i = 0$, questo è un altro modo di dire che la somma $\sum_{i=1}^{m} \lambda_i \nabla g_i(x)$ deve includere solo gli indici $1 \leq i \leq m$ per i quali $g_i(x) = 0$. Tale osservazione unita al teorema A.8 ci permette di concludere. □

Corollario A.1. *Se $x \in S$ è un punto KKT allora è stazionario.*

Dimostrazione. La condizione dei punti KKT implica che $-\nabla f(x) \in (F_x S)^\circ$, inoltre il teorema A.4 ci dice che $T_x S \subseteq F_x S$, allora $(F_x S)^\circ \subseteq (T_x S)^\circ = N_x S$. Allora la condizione KKT implica che $-\nabla f(x) \in N_x S$, cioè x è stazionario. □

Corollario A.2. *Se $x \in S$ è stazionario e un (CQ) vale in x, allora x è un punto KKT.*

Dimostrazione. Il (CQ) implica che $T_x S = F_x S$. Allora, la stazionarietà di x implica che $-\nabla f(x) \in N_x S = (T_x S)^\circ = (F_x S)^\circ$, dal lemma A.2 ne segue che x è un punto KKT. □

Teorema A.9. *Se $x^* \in S$ è un minimo locale e un (CQ) vale in x^*, allora x^* è un punto KKT.*

Dimostrazione. Se x^* è un minimo locale allora è un punto stazionario, allora per il corollario A.2 x^* è KKT. □

A.4 Dualità

Considerando sempre un problema di minimizzazione vincolato con S definito da vincoli di uguaglianza e disuguaglianza (vedasi A.1), si definisce la *Lagrangiana* associata al problema

$$L : \mathcal{E} \times \mathbb{R}^p \times \mathbb{R}^m_+ \to \mathbb{R}$$

definita da $L(x, \mu, \lambda) = f(x) + \sum_{i=1}^{p} \mu_i h_i(x) + \sum_{i=1}^{m} \lambda_i g_i(x)$ dove $\mathbb{R}^m_+ = \{\lambda \in \mathbb{R}^m : \lambda_i \geq 0 \text{ per } i = 1, \ldots, m\}$. Si definisce la *Lagrangiana primale* associata al problema la funzione seguente

$$L_p(x) = \sup_{\mu \in \mathbb{R}^p, \lambda \in \mathbb{R}^m_+} L(x, \mu, \lambda)$$

E' facile verificare che

$$L_p(x) = \begin{cases} f(x) \text{ se } x \in S \\ +\infty \text{ altrimenti} \end{cases}$$

Allora il seguente problema ha la medesima soluzione dell'originale ed è chiamato *problema primale*

$$\min_{x \in \mathcal{E}} L_P(x)$$

Definiamo ora la *Lagrangiana duale* associata al problema di minimizzazione originale come quella funzione definita da

$$L_D(\mu, \lambda) = \inf_{x \in \mathcal{E}} L(x, \mu, \lambda)$$

Si può definire allora il *problema duale* il seguente:

$$\max_{\mu \in \mathbb{R}^p, \lambda \in \mathbb{R}^m_+} L_D(\mu, \lambda)$$

Ci sono dei legami tra i valori della funzione obiettivo del primale e del duale, queste relazioni sono formalizzate dai seguenti teoremi.

Teorema A.10. *(Dualità debole): Per ogni $x \in \mathcal{E}$ e $(\mu, \lambda) \in \mathbb{R}^p \times \mathbb{R}^m_+$, vale che $L_D(\mu, \lambda) \leq L_P(x)$, in particolare,*

$$d^* = \max_{(\mu,\lambda)\in\mathbb{R}^p\times\mathbb{R}^m_+} L_D(\mu, \lambda) \leq \min_{x\in\mathcal{E}} L_P(x) = \min_{x\in S} f(x) = p^*$$

Dimostrazione. Fissati $\bar{x} \in \mathcal{E}$, $\bar{\mu} \in \mathbb{R}^p$, $\bar{\lambda} \in \mathbb{R}^m_+$, abbiamo

$$L_D(\bar{\mu}, \bar{\lambda}) = \inf_{x\in\mathcal{E}} L(x, \bar{\mu}, \bar{\lambda}) \leq L(\bar{x}, \bar{\mu}, \bar{\lambda}) \leq \sup_{\mu\in\mathbb{R}^p, \lambda\in\mathbb{R}^m_+} L(\bar{x}, \mu, \lambda) = L_P(\bar{x})$$

Visto che la disuguaglianza vale per qualsiasi scelta di $\bar{x}, \bar{\mu}, \bar{\lambda}$ allora vale anche per gli ottimi globali del duale e del primale. □

Teorema A.11. *(Dualità forte): Assumendo che valgano le seguenti:*

1. *Il problema primale ammetta un minimo globale $x^* \in S$.*
2. *Esistono validi moltiplicatori di Lagrange $\mu^* \in \mathbb{R}^p, \lambda^* \in \mathbb{R}^m_+$ per x^*.*
3. *La funzione $x \to L(x, \mu^*, \lambda^*)$ è convessa.*

Allora abbiamo la dualità forte cioè $d^* = p^*$. Inoltre, (μ^*, λ^*) sono ottimali per il duali e x^* è il minimo di

$$\min_{x\in\mathcal{E}} L(x, \mu^*, \lambda^*)$$

Dimostrazione. Notiamo prima di tutto che il gradiente della funzione $x \to L(x, \mu, \lambda)$ è

$$\nabla_x L(x, \mu, \lambda) = \nabla f(x) + \sum_{i=1}^{p} \mu_i \nabla h_i(x) + \sum_{i=1}^{m} \lambda_i \nabla g_i(x)$$

Ciò implica che se $\bar{x}$ è un punto KKT per il problema iniziale, allora

$$L(\bar{x}, \bar{\mu}, \bar{\lambda}) = f(\bar{x}) \quad \text{e} \quad \nabla_x L(\bar{x}, \bar{\mu}, \bar{\lambda}) = 0$$

Ora, grazie alla dualità debole già sappiamo che $p^* \geq d^*$, dobbiamo quindi mostrare che $p^* \leq d^*$. Per ipotesi esiste un minimo globale x^* per il problema primale e moltiplicatori di Lagrange associati $\mu^* \in \mathbb{R}^p$ e $\lambda^* \in \mathbb{R}^m_+$, ne segue dal teorema A.9 e da quanto appena visto che

$$L(x^*, \mu^*, \lambda^*) = f(x^*) \quad \text{e} \quad \nabla_x L(x^*, \lambda^*, \mu^*) = 0$$

Inoltre la funzione $x \to L(x, \mu^*, \lambda^*)$ è convessa, visto che x^* è un punto stazionario per questa funzione, ne segue che x^* è un minimo globale per questa funzione, allora

$$d^* = \max_{\mu\in\mathbb{R}^p, \lambda\in\mathbb{R}^m_+} L_D(\mu, \lambda) \geq L_D(\mu^*, \lambda^*) = \inf_{x\in\mathcal{E}} L(x, \mu^*, \lambda^*) = L(x^*, \mu^*, \lambda^*) = f(x^*) = p^*$$

□

Questo è uno dei risultati più importanti nell'ambito dell'ottimizzazione in quanto se le ipotesi del teorema della dualità forte sono soddisfatte, possiamo risolvere il problema duale al posto di quello primale per ottenere il valore ottimo della funzione obiettivo del primale. Talvolta infatti risolvere il problema duale risulta più semplice che risolvere il problema primale. Inoltre vale il seguente teorema.

Teorema A.12. *La funzione Lagrangiana duale L_D è sempre concava, allora il problema duale è convesso anche se il primale non è convesso.*

Dimostrazione. Per provare che L_D è concava è sufficiente notare che L_D è l'inf di una collezione di funzioni affini (e quindi concave). Allora L_D è concava e il problema duale

$$\max_{\mu\in\mathbb{R}^p,\lambda\in\mathbb{R}^m_+} L_D(\mu,\lambda)$$

è equivalente a

$$\min_{\mu\in\mathbb{R}^p,\lambda\in\mathbb{R}^m_+} -L_D(\mu,\lambda)$$

Che è un problema convesso in quanto non vincolato e $-L_D$ è una funzione convessa. □

A.5 Problemi convessi di ottimizzazione vincolata

Definizione A.14. *Un insieme $S \subseteq \mathcal{E}$ è convesso se*

$$\forall x,y \in S,\ \ \forall t \in [0,1],\ \ (1-t)x + ty \in S$$

In altre parole un insieme è convesso se contiene il segmento che connette ogni coppia di punti in S.

Teorema A.13. *Considerando il problema di minimizzazione vincolato $\min_{x\in S} f(x)$, se S è convesso e f è convessa e differenziabile allora*

$$x^* \in S \ \text{ è minimo globale } \iff -\nabla f(x^*) \in N_{x^*}S$$

Dimostrazione. Dal teorema A.3 sappiamo che se x^* è un minimo globale allora $-\nabla f(x^*) \in N_{x^*}S$, dobbiamo solo provare l'altra freccia. Assumiamo che $x^* \in S$ sia ora tale che $-\nabla f(x^*) \in N_{x^*}S$. Per contraddizione supponiamo esista $x \in S$ tale per cui $f(x) < f(x^*)$. Per convessità di S, segue che $c(t) = x^* + t(x - x^*)$ è in S per ogni $t \in [0,1]$. Ma allora $c'(0) = x - x^*$ appartiene al cono tangente $T_{x^*}S$. La nostra ipotesi che $-\nabla f(x^*) \in N_{x^*}S$ implica allora che

$$\langle -\nabla f(x^*), x - x^* \rangle \leq 0$$

Inoltre, la convessità di f implica che

$$f(x) > f(x^*) + \langle \nabla f(x^*), x - x^* \rangle$$

Combinando i due risultati abbiamo che $f(x) > f(x^*)$ che è una contraddizione. □

Definizione A.15. *Un problema di minimizzazione vincolata è chiamato problema di ottimizzazione convesso se entrambi S e f sono convessi.*

Considerando ora il caso specifico in cui l'insieme S sia definito tramite vincoli di uguaglianza e disuguaglianza come in A.1, l'insieme S può essere riscritto in maniera equivalente nel seguente modo

$$\begin{aligned} S =& \{x \in \mathcal{E} : h_1(x) = 0\} \cap \cdots \cap \{x \in \mathcal{E} : h_p(x) = 0\} \\ & \cap \{x \in \mathcal{E} : g_1(x) \leq 0\} \cap \cdots \cap \{x \in \mathcal{E} : g_m(x) \leq 0\} \end{aligned}$$

Inoltre è immediata la verifica dei seguenti teoremi

Teorema A.14. *L'intersezione finita di insiemi convessi è convessa.*

Dimostrazione. Siano $S_1, S_2, \ldots, S_n$ insiemi convessi, allora da definizione vale che

$$\forall x, y \in S_i, \ \ \forall t \in [0,1], \ \ (1-t)x + ty \in S_i$$

Allora presi $x, y \in S_1 \cap \cdots \cap S_n$, vale che $x, y \in S_i$ per ogni $i = 1, \ldots, n$ e quindi

$$\forall x, y \in S_i, \ \ \forall t \in [0,1], \ \ (1-t)x + ty \in S_i \ \text{ per ogni } \ i = 1, \ldots, n$$

Cioè $S_1 \cap \cdots \cap S_n$ è convesso. □

Teorema A.15. *Data $g : \mathcal{E} \to \mathbb{R}$, l'insieme $S = \{x \in \mathcal{E} : g(x) \leq 0\}$ è convesso se g è convessa.*

Dimostrazione. Siccome g è convessa $\forall x, y \in S$ e $\forall t \in [0,1]$ abbiamo

$$g((1-t)x + ty) \leq (1-t)g(x) + tg(y) \leq 0$$

Da cui $(1-t)x + ty \in S$, allora S è convesso. □

Teorema A.16. *Data $h : \mathcal{E} \to \mathbb{R}$, l'insieme $S = \{x \in \mathcal{E} : h(x) = 0\}$ è convesso se h è affine cioè $h(x) = \langle w, x \rangle + b$ per qualche $w \in \mathcal{E}$, $b \in \mathbb{R}$.*

Dimostrazione. Siano $x, y \in S$, allora $\langle w, x \rangle + b = 0$ e $\langle w, y \rangle + b = 0$, inoltre qualsiasi $t \in [0,1]$, vale che $h(tx + (1-t)y) = \langle tx + (1-t)y, w \rangle + b = th(x) + (1-t)h(y) = 0$, quindi $tx + (1-t)y \in S$ cioè S è convesso. □

Corollario A.3. *L'insieme S definito da vincoli di uguaglianza e disuguaglianza come in A.1 è convesso se le funzioni $h_1, \ldots, h_p$ sono affini e $g_1, \ldots, g_m$ sono convesse.*

Teorema A.17. *(Slater's CQ) Sia S definito da vincoli di uguaglianza e disuguaglianza come in A.1, se le funzioni $h_1, \ldots, h_p$ sono affini e $g_1, \ldots, g_m$ sono convesse e esiste un punto $x_s \in S$ che soddisfi tutte le disuguaglianze strettamente allora $T_x S = F_x S$ per ogni $x \in S$ e si dice che valgono le Slater's qualification conditions.*

Inoltre abbiamo speciali risultati riguardanti i punti KKT e la dualità nel caso in cui le funzioni che definiscono i vincoli di uguaglianza siano affini e quelle che definiscono i vincoli di disuguaglianza siano convesse.

Teorema A.18. *(KKT convesso) Sia il problema di ottimizzazione vincolata in cui l'insieme S sia definito da vincoli di uguaglianza e disuguaglianza come in A.1. Supponendo inoltre che le funzioni $h_1, \dots, h_p$ siano affini e $f, g_1, \dots, g_m$ siano convesse (in particolare S è convesso), i seguenti valgono.*

1. *Se le condizioni KKT valgono in $x^* \in S$ allora x^* è un minimo globale.*
2. *Se un (CQ) vale in $x^* \in S$ e x^* è un minimo locale (e quindi globale) allora le condizioni KKT valgono in x^*.*

Dimostrazione. A causa della convessità il teorema A.13 afferma che punti stazionari, minimi locali e minimi globali coincidono, allora il primo punto segue dal corollario A.1, il secondo dal corollario A.2. □

Corollario A.4. *Nelle stesse ipotesi del teorema A.18, se un (CQ) vale per tutti i punti di S allora le condizioni KKT sono necessarie e sufficienti per l'ottimalità globale.*

Corollario A.5. *(Dualità convessa) Nelle stesse ipotesi del teorema A.18, se il problema primale ammette un minimo globale $x^* \in S$ e una condizione di constraint qualification vale in x^* allora la dualità forte vale e i moltiplicatori di Lagrange per il primale sono ottimi per il duale.*

Dimostrazione. Sarà sufficiente verificare che le ipotesi del teorema di dualità forte sono rispettate. Chiaramente la prima ipotesi è verificata in quanto x^* è un minimo globale, inoltre il fatto che valga un (CQ) ad x^* implica l'esistenza di validi moltiplicatori di Lagrange μ^*, λ^* ad x^* (Teorema A.9). Inoltre sotto le ipotesi del corollario $x \to L(x, \mu^*, \lambda^*)$ è convessa in quanto $f, g_1, \dots, g_m$ sono convesse e $h_1, \dots, h_p$ affini. Allora le ipotesi del teorema di dualità forte sono rispettate e il corollario segue direttamente dal teorema di dualità forte. □

Bibliografia

[1] Dirk P.Kroese, Zdravko I. Botev, Thomas Taimre and Radislav Vaisman. Data Science and Machine Learning, Mathematical and Statistical Methods. CRC Press 2020.

[2] Shai Shalev-Shwartz and Shai Ben-David. Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, 2014.

[3] M. P. Deisenroth, A. A. Faisal, and C. S. Ong. Mathematics for Machine Learning. Cambridge University Press, 2020.

[4] Jorge Nocedal and Stephen J. Wright. Numerical Optimization. Springer, 2006

[5] Nicolas Boumal. Lecture notes for the course of Nonlinear Optimization at École polytechnique fédérale de Lausanne (EPFL), 2021.