ANALISI FDA e Kernel FDA

Enrico Greppi

11/07/2022

L'Analisi del Discriminante di Fisher (FDA) è una tecnica matematica vista per la prima volta nel 1936 (quindi precedente alle applicazioni più incisive che assume in ambito Machine Learning). Nasce per affrontare un problema di Catalogazione di fossili: l'obiettivo era quello di assegnare il fossile alla categoria di Umanoidi o di Primati a partire da un insieme di dati genetici a disposizione. È poi diventata molto utilizzata nei problemi di intelligenza artificiale legati alla classificazione e di riduzione dimensionale. Assumiamo di avere un dataset di campionamento $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ di semplice dimensione n(numero di campioni), $\mathbf{x}_i \in \mathbb{R}^d, y_i \in \mathbb{R}$. Le $\{\mathbf{x}_i\}_{i=1}^n$ sono i dati del modello (d i predittori o variabili) mentre le $\{y_i\}_{i=1}^n$ sono le corrispondenti etichette che assegnano un dato a una delle nostre c classi. Definiamo $\mathbb{R}^{d\times n} \ni \mathbf{X} := [\mathbf{x}_1, ..., \mathbf{x}_n]$ e $\mathbb{R}^{1\times n} \ni Y := [y_1, ..., y_n]$.Le y_i appartengono a un set discreto di cardinalità coincidente al numero di classi .

Assumiamo che il dataset sia composto da c classi, $\{\mathbf{x}_i^{(1)}\}_{i=1}^{n_1}, ..., \{\mathbf{x}_i^{(c)}\}_{i=1}^{n_c}$ dove n_j è la cardinalità della j -esima classe.

La FDA ha due obiettivi (simili ma interpretati da due punti di vista differenti):

1)**riduzione dimensionale**: Trovare un sottospazio che separi il più possibile le varie classi e allo stesso tempo che schiacci i dati di una stessa classe il più possibile vicini tra loro.

2)**costruzione di un classificatore**: Costruire un classificatore, ovvero una 'macchina', la cui variabile di risposta sarà discreta, che, una volta inserito un dato nella macchina, ci consenta di assegnargli un' etichetta (che lo classifichi in una delle nostre classi).

La kernel FDA ha gli stessi obiettivi ma li raggiunge lavorando con kernel nello spazio dei funzionali. Sono tecniche molto utilizzate in intelligenza artificiale, alcune applicazioni importanti sono il riconoscimento facciale, action recognition e gesture recognition.

Formule di Proiezione

Assumiamo di avere un punto $\mathbf{x} \in \mathbb{R}^d$ noi vogliamo proiettare questo punto in un sottospazio vettoriale formato da uno span di p vettori : $\{\mathbf{u}_1, ..., \mathbf{u}_p\}$ in cui ogni vettore è d-dimensionale. Inseriamo lo span dei p vettori in una matrice $\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_p] \in \mathbb{R}^{d \times p}$. Vogliamo perciò proiettare $\mathbf{x}$ nel sottospazio formato dalle colonne di $\mathbf{U}$ denotato con $\mathbb{C}ol(\mathbf{U})$. La proiezione di $\mathbf{x} \in \mathbb{R}^d$ in $\mathbb{C}ol(\mathbf{U}) \in \mathbb{R}^p$ e la sua ricostruzione in $\mathbb{R}^d$ possono essere viste come un sistema lineare di equazioni :

$$\mathbb{R}^d \ni \hat{\boldsymbol{x}} := \boldsymbol{U\beta} \tag{1}$$

in cui i coefficienti di $\boldsymbol{\beta} \in \mathbb{R}^p$ sono sconosciuti e siamo interessati a trovarli. Se $\mathbf{x}$ giace nello spazio $\mathbb{C}ol(\mathbf{U})$, il sistema lineare (1) ammette una soluzione esatta: $\hat{\boldsymbol{x}} = \boldsymbol{x} = \boldsymbol{U\beta}$. Invece, se $\mathbf{x}$ non giace in questo spazio, non esiste nessun $\boldsymbol{\beta}$ per il quale $\boldsymbol{x} = \boldsymbol{U\beta}$. In questo caso $\mathbf{x}$ e $\hat{\boldsymbol{x}}$ sono diverse e introduciamo il residuo:

$$\boldsymbol{r} = \boldsymbol{x} - \hat{\boldsymbol{x}} = \boldsymbol{x} - \boldsymbol{U\beta} \tag{2}$$

che desideriamo avere il più piccolo possibile. Si può dimostrare che il residuo più piccolo è quello ortogonale allo spazio $\mathbb{C}ol(\mathbf{U})$. Dunque:

$$\begin{aligned} \boldsymbol{x} - \boldsymbol{U\beta} \perp \boldsymbol{U} \Rightarrow \boldsymbol{U}^T(\boldsymbol{x} - \boldsymbol{U\beta}) = 0 \\ \Rightarrow \mathbf{U}^T\boldsymbol{x} = \boldsymbol{U}^T\boldsymbol{U\beta} \Rightarrow \boldsymbol{\beta} = (\boldsymbol{U}^T\mathbf{U})^{-1}\mathbf{U}^T\mathbf{x} \end{aligned} \tag{3}$$

Combinando la (1) con la (3) otteniamo:
$\hat{\boldsymbol{x}} = \boldsymbol{U}(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\mathbf{x}$
Ora definiamo:

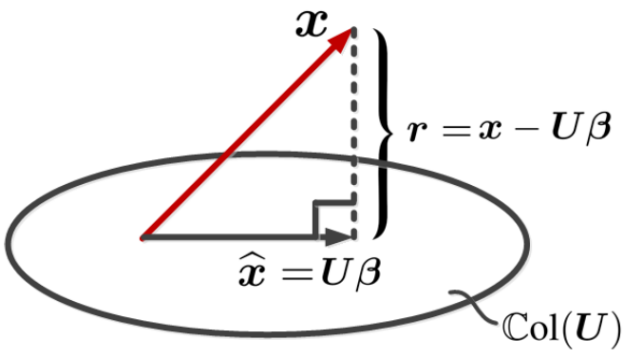
$$\mathbb{R}^{d \times d} \ni \boldsymbol{\Pi} := \boldsymbol{U}(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T \tag{4}$$

questa è la matrice di proiezione che manda $\mathbf{x}$ in $\mathbb{C}ol(\mathbf{U})$. Se i vettori $\{\mathbf{u}_1, ..., \mathbf{u}_p\}$ sono ortonormali allora $\mathbf{U}$ è ortogonale ($\mathbf{U}^T = \mathbf{U}^{-1}, \mathbf{U}^T\mathbf{U} = \mathbf{I}$) e la (4) è semplificata :

$$\boldsymbol{\Pi} = \mathbf{U}\mathbf{U}^T \tag{5}$$

e perciò:

$$\hat{\boldsymbol{x}} = \boldsymbol{\Pi x} = \boldsymbol{U}\boldsymbol{U}^T\mathbf{x} \tag{6}$$



Proiezione in un sottospazio

La proiezione di un vettore $\mathbf{x} \in \mathbb{R}^d$ nello spazio formato dallo span delle colonne di $\mathbf{U} \in \mathbb{R}^{d \times p}$ è definita come:

$$\mathbb{R}^p \ni \tilde{\boldsymbol{x}} := \mathbf{U}^T\mathbf{x} \tag{7}$$

$$\mathbb{R}^d \ni \hat{\boldsymbol{x}} := \ \mathbf{U}\mathbf{U}^T\mathbf{x} = \boldsymbol{U}\tilde{\boldsymbol{x}} \tag{8}$$

dove $\tilde{\boldsymbol{x}}$ è la proiezione di $\mathbf{x}$, mentre $\hat{\boldsymbol{x}}$ la sua ricostruzione

Se abbiamo n dati (punti) $\{\mathbf{x}\}_{i=1}^n$ possiamo ragrupparli in un'unica matrice $\mathbf{X} \in \mathbb{R}^{d \times n}$, lo stesso vale per la proiezione e la ricostruzione rispettivamente :

$$\mathbb{R}^{p \times n} \ni \tilde{\boldsymbol{X}} := \mathbf{U}^T\mathbf{X}, \tag{9}$$

$$\mathbb{R}^{d \times n} \ni \hat{\boldsymbol{X}} := \ \mathbf{U}\mathbf{U}^T\mathbf{X} = \boldsymbol{U}\tilde{\boldsymbol{X}} \tag{10}$$

Proiezione in un sottospazio 1-dimensionale

Considerando i nostri campioni $\{\mathbf{x}_i\}_{i=1}^n$, la media è :

$$\mathbb{R}^d \ni \boldsymbol{\mu}_x := \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i \tag{11}$$

E la matrice dei campioni centrati è:

$$\mathbb{R}^{d \times n} \ni \breve{\boldsymbol{X}} := \mathbf{X} - \boldsymbol{\mu}_x = \boldsymbol{X}\boldsymbol{H} \tag{12}$$

dove $\breve{\boldsymbol{X}} = [\breve{\boldsymbol{x}}_1, ..., \breve{\boldsymbol{x}}_n] \in \mathbb{R}^{d\times n}$ in cui $\mathbb{R}^d \ni \breve{\boldsymbol{x}}_i := \boldsymbol{x}_i - \boldsymbol{\mu}_x$;

$\mathbb{R}^{n\times n} \ni H := \mathbf{I} - (\frac{1}{n})\mathbf{1}\mathbf{1}^T$.

Se nella (7) sostituiamo 1 a p, stiamo proiettando $\mathbf{x}$ in un solo vettore $\mathbf{u}$ e andando poi a ricostruirla otteniamo, se il dato è centrato,

$\hat{\boldsymbol{x}} = \boldsymbol{u}\boldsymbol{u}^T\breve{\boldsymbol{x}}$

e la norma *l2* di questo vettore ricostruito è:

$$\begin{aligned} \|\hat{\boldsymbol{x}}\|_2^2 = \|\boldsymbol{u}\boldsymbol{u}^{\boldsymbol{T}}\breve{\boldsymbol{x}}\|_2^2 = (\boldsymbol{u}\boldsymbol{u}^{\boldsymbol{T}}\breve{\boldsymbol{x}})^T(\boldsymbol{u}\boldsymbol{u}^{\boldsymbol{T}}\breve{\boldsymbol{x}}) \\ = \breve{\boldsymbol{x}}^{\boldsymbol{T}}\boldsymbol{u}\boldsymbol{u}^{\boldsymbol{T}}\boldsymbol{u}\boldsymbol{u}^{\boldsymbol{T}}\breve{\boldsymbol{x}} =_{(a)} \breve{\boldsymbol{x}}^{\boldsymbol{T}}\boldsymbol{u}\boldsymbol{u}^{\boldsymbol{T}}\breve{\boldsymbol{x}} =_{(b)} \boldsymbol{u}^{\boldsymbol{T}}\breve{\boldsymbol{x}}\breve{\boldsymbol{x}}^{\boldsymbol{T}}\boldsymbol{u} \end{aligned} \tag{13}$$

(a) perchè $\mathbf{u}$ è un vettore normalizzato (di norma 1) e (b) perchè è un prodotto scalare in $l2(\mathbb{R})$ e di conseguenza commutativo.

Supponiamo di avere n dati $\{\mathbf{x}_i\}_{i=1}^n$, le somma delle norme *l2* delle loro proiezioni (ricostruite) è:

$$\sum_{i=1}^{n} \|\hat{\boldsymbol{x}}_i\|_2^2 =_{(13)} \sum_{i=1}^{n} \boldsymbol{u}^T\breve{\boldsymbol{x}}_{\boldsymbol{i}}\breve{\boldsymbol{x}}_{\boldsymbol{i}}{}^T\boldsymbol{u} = \boldsymbol{u}^T(\sum_{i=1}^{n} \breve{\boldsymbol{x}}_{\boldsymbol{i}}\breve{\boldsymbol{x}}_{\boldsymbol{i}}{}^T)\boldsymbol{u} \tag{14}$$

Se prendiamo $\breve{\boldsymbol{X}} = [\breve{\boldsymbol{x}}_1, ..., \breve{\boldsymbol{x}}_n] \in \mathbb{R}^{d\times n}$, abbiamo :

$$\begin{aligned} \mathbb{R}^{d\times d} \ni \boldsymbol{S} := \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{x}})(\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{x}})^T \\ = \sum_{i=1}^{n} \breve{\boldsymbol{x}}_i\breve{\boldsymbol{x}}_i^T = \breve{\boldsymbol{X}}\breve{\boldsymbol{X}}^T =_{(12)} \boldsymbol{X}\boldsymbol{H}\boldsymbol{H}\boldsymbol{X}^T \end{aligned} \tag{15}$$

$\mathbf{S}$ è la matrice di varianza covarianza campionaria e se i dati sono già centrati abbiamo $\mathbf{S} = \mathbf{X}\mathbf{X}^T$.

Combiando la (15) con la (14) arriviamo a :

$$\sum_{i=1}^{n} \|\hat{\boldsymbol{x}}_i\|_2^2 = \boldsymbol{u}^T\boldsymbol{S}\boldsymbol{u} \tag{16}$$

Da notare che questa rappresenta anche la varianza dei dati nel sottospazio della PCA

Questa può essere interpretata sia come la misura in norma *l2* al quadrato della ricostruzione del campionamento proiettato, sia come la varianza di proiezione.

$$\|\hat{\boldsymbol{X}}\|_F^2 = \boldsymbol{u}^T\boldsymbol{S}\boldsymbol{u} \tag{17}$$

Proiezione in un sottospazio multidimensionale

In questo caso p>1, il sottospazio in cui vogliamo proiettare i dati è generato da $\{\boldsymbol{u}_j\}_{j=1}^{p}$. Se assumiamo i dati centrati, la ricostruzione sarà

$\hat{\boldsymbol{X}} = \boldsymbol{U}\boldsymbol{U}^T\check{\boldsymbol{X}}$

e

$$\begin{aligned}\|\hat{\boldsymbol{X}}\|_F^2 = \|\mathbf{U}\mathbf{U}^T\check{\boldsymbol{X}}\|_F^2 = \mathbf{tr}((\mathbf{U}\mathbf{U}^T\check{\boldsymbol{X}})^T(\mathbf{U}\mathbf{U}^T\check{\boldsymbol{X}})) = \\ \mathbf{tr}(\check{\boldsymbol{X}}^T\mathbf{U}\mathbf{U}^T\mathbf{U}\mathbf{U}^T\check{\boldsymbol{X}}) =_{(a)} \mathbf{tr}(\check{\boldsymbol{X}}^T\mathbf{U}\mathbf{U}^T\check{\boldsymbol{X}}) = \\ \mathbf{tr}(\mathbf{U}^T\check{\boldsymbol{X}}\check{\boldsymbol{X}}^T\mathbf{U}) =_{(15)} tr(\boldsymbol{U}^T\boldsymbol{S}\boldsymbol{U})\end{aligned} \tag{18}$$

FDA: Fisher Discriminant Analysis

Sottospazio 1 dimensionale

2 classi

Trattiamo prima il caso di due sole classi per fissare meglio le idee e generalizziamolo in seguito al caso di più classi. Assumiamo di avere quindi due classi di dati $\{\mathbf{x}_i^{(1)}\}_{i=1}^{n_1}$ e $\{\mathbf{x}_i^{(2)}\}_{i=1}^{n_2}$, con $\mathbf{x}_i^{(j)}$che denota l'i-esima istanza della j-esima classe. Se i campioni sono proiettati in un sottospazio 1-dimensionale (generato da vettore $\mathbf{u}$) da $\mathbf{u}^T\mathbf{x}_i^{(j)}$, la media e la varianza campionarie proiettate sono $\mathbf{u}^T\boldsymbol{\mu}_j$ e $\mathbf{u}^T\mathbf{S}\mathbf{u}$

$$\mathbb{R}^d \ni \boldsymbol{\mu}_j := \frac{1}{n_j}\sum \mathbf{x}_i^{(j)} \tag{19}$$

dopo la proiezione la distanza tra le medie delle classi è:

$$\begin{aligned}\mathbb{R} \ni d_B := (\mathbf{u}^T\boldsymbol{\mu_1} - \mathbf{u}^T\boldsymbol{\mu_2})^T(\mathbf{u}^T\boldsymbol{\mu_1} - \mathbf{u}^T\boldsymbol{\mu_2}) \\ = (\boldsymbol{\mu_1} - \boldsymbol{\mu_2}^T)\mathbf{u}\mathbf{u}^T(\boldsymbol{\mu_1} - \boldsymbol{\mu_2}) \\ =_{(a)} \mathbf{tr}((\boldsymbol{\mu_1} - \boldsymbol{\mu_2})^T\mathbf{u}\mathbf{u}^T(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})) \\ =_{(b)} \mathbf{tr}(\mathbf{u}^T(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})^T\mathbf{u}) \\ =_{(c)} \mathbf{u}^T(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})^T\mathbf{u} =_{(d)} \mathbf{u}^T\boldsymbol{S_B}\boldsymbol{u}\end{aligned} \tag{20}$$

(a) perchè $(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})^T\mathbf{u}\mathbf{u}^T(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})$ è uno scalare, (b) per la proprietà ciclica delle tracce , (c) perchè $\mathbf{u}^T(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})^T\mathbf{u}$ è uno scalare e (d) è perchè definiamo:

$$\mathbb{R}^{d\times d} \ni \boldsymbol{S_B} := (\boldsymbol{\mu_1} - \boldsymbol{\mu_2})(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})^T \tag{21}$$

matrice di between-scatter, cioè di varianza covarianza TRA le classi. L'equazione (20) può anche essere interpretata in accordo con (17) : d_B rappresenta la varianza di proiezione delle medie delle classi o la norma al quadrato della ricostruzione delle medie delle classi. La varianza di proiezione per la j-esima classe è $\mathbf{u}^T \boldsymbol{S_j} \boldsymbol{u}$. Se sommiamo le varianze di proiezione delle due classi abbiamo

$$\mathbb{R} \ni d_W := \mathbf{u}^T \boldsymbol{S_1} \boldsymbol{u} + \mathbf{u}^T \boldsymbol{S_2} \boldsymbol{u} = \mathbf{u}^T (\boldsymbol{S_1} + \boldsymbol{S_2}) \mathbf{u} = \mathbf{u}^T \boldsymbol{S_W} \boldsymbol{u} \tag{22}$$

dove:

$$\mathbb{R}^{d \times d} \ni \boldsymbol{S_W} := \boldsymbol{S_1} + \boldsymbol{S_2} \tag{23}$$

within-scatter matrix è la matrice di varianza covarianza INTRAclasse.

Più classi

Ora aumentiamo il numero di classi, portandolo da 2 ad un generico c ed estendiamo i concetti visti fino ad ora, la matrice di varianza tra le classi è definita come:

$$\mathbb{R}^{d \times d} \ni \boldsymbol{S_B} := \sum_{j=1}^{c} (\boldsymbol{\mu_j} - \boldsymbol{\mu})(\boldsymbol{\mu_j} - \boldsymbol{\mu})^T \tag{24}$$

$$\mathbb{R}^d \ni \boldsymbol{\mu} := \frac{1}{\sum_{k=1}^{c} n_k} \sum_{j=1}^{c} n_j \boldsymbol{\mu_j} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \tag{25}$$

e la matrice di covarianza intraclasse diventa:

$$\mathbb{R}^{d \times d} \ni \boldsymbol{S_W} := \sum_{j=1}^{c} \boldsymbol{S_j} =_{(15)} \sum_{j=1}^{c} \sum_{i=1}^{n_j} (\boldsymbol{x}_i^{(j)} - \boldsymbol{\mu}_j)(\boldsymbol{x}_i^{(j)} - \boldsymbol{\mu}_j)^T \tag{26}$$

$$\mathbb{R} \ni d_B := \mathbf{u}^T \boldsymbol{S_B} \boldsymbol{u} \tag{27}$$

$$\mathbb{R} \ni d_W := \mathbf{u}^T \boldsymbol{S_W} \boldsymbol{u} \tag{28}$$

matrice di varianza covarianza

Un'altra variante molto interessante per trovare $\boldsymbol{S_W}$ *e* $\boldsymbol{S_B}$ è la seguente: $\boldsymbol{S_W}$ è definita come in (26) . Viene ora introdotta la matrice di varianza

covarianza totale (dell'intero universo campionario):

$$\mathbb{R}^{d\times d} \ni \boldsymbol{S_T} := \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T \tag{29}$$

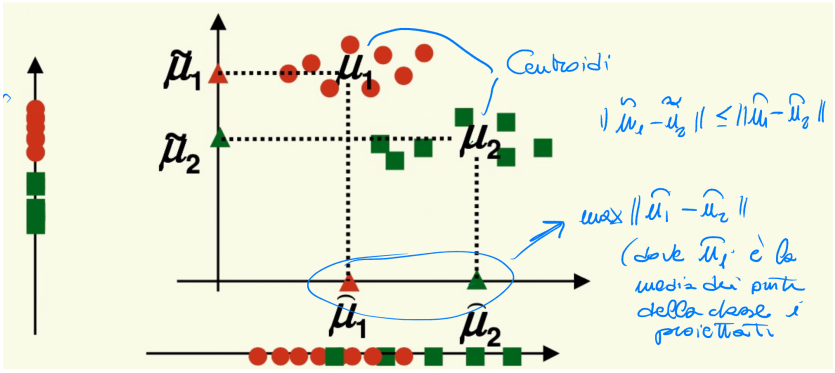
e

$$\mathbb{R}^{d\times d} \ni \boldsymbol{S_T} := \boldsymbol{S_B} + \boldsymbol{S_W} \tag{30}$$

Di conseguenza $\boldsymbol{S_B}$, ricavata da (30) ne esce leggermente diversa ma ci permette di capire i ruoli della matrice di varianza covarianza intraclasse e tra le classi attraverso un analogia con la fisica. Supponiamo i campioni delle varie classi distributi a formare una specie di ellissoide o anche a un livello più astratto corpi di materia. La matrice di varianza covarianza totale rappresenta l'inerzia totale del sistema ed è quindi possibile scomporla nella somma delle inerzie di ogni ellissoide (o corpo di materia) ($\boldsymbol{S_W}$) + l'inerzia tra i vari corpi ($\boldsymbol{S_B}$)
Una volta evidenziata questa interpretazione tornerei alle prime definizioni che ci porteranno in seguito ad evidenziare come entrambe le matrici hanno un ruolo importante nella ricerca delle funzioni discriminanti, ma l'intero metodo è comunque ricavabile anche a partire da queste seconde definizioni delle matrici di varianza covarianza.

Sottospazio di Fisher

Nella FDA vogliamo massimizzare la varianza di proiezione delle medie delle classi e minimizzare invece la varianza di proiezione 'intraclasse' (tra le istanze di una stessa classe). Questo è infatti il modo per evidenziare maggiormente la separazione delle classi.



Vogliamo quindi massimizzare d_B e minimizzare d_W. Così facendo le istanze di una stessa classe si avvicinano una con l'altra compattandosi mentre le classi si allontanano una dall'altra.

I due problemi di ottimizzazione sono:

$$max_{\mathbf{u}} d_B(\mathbf{u}) \tag{31}$$

$$min_{\mathbf{u}} d_W(\mathbf{u}) \tag{32}$$

Uniamo ora (31) e (32) in un unico problema di ottimizzazione :

$$max_{\mathbf{u}} f(\mathbf{u}) := \frac{d_B(\mathbf{u})}{d_W(\mathbf{u})} = \frac{\mathbf{u}^T \boldsymbol{S_B} \boldsymbol{u}}{\mathbf{u}^T \boldsymbol{S_W} \boldsymbol{u}} \tag{33}$$

Dove f($\mathbf{u}$) si riferisce al criterio di Fisher

Lemma: Quoziente di Rayleight-Ritz generalizzato

Sia

$$\mathbb{R} \ni R(\mathbf{A}, \mathbf{B}; \mathbf{x}) := \frac{\mathbf{x}^T \mathbf{A} \mathbf{x}}{\mathbf{x}^T \mathbf{B} \mathbf{x}} \tag{34}$$

il quoziente generalizzato di Rayleight-Ritz, con $\mathbf{A}$ e $\mathbf{B}$ simmetriche e definite positive;

Allora abbiamo che :

$$\begin{aligned} &minimize/maximize_{\mathbf{x}}\ R(\mathbf{A}, \mathbf{B}; \mathbf{x}) \equiv \\ &minimize/maximize_{\mathbf{x}}\ \mathbf{x}^T \mathbf{A} \mathbf{x} \\ &S.T\ \mathbf{x}^T \mathbf{B} \mathbf{x} = 1 \end{aligned} \tag{35}$$

In accordo con (35) la (33) è equivalente al problema:

$$\begin{aligned} max_{\mathbf{u}} &= \mathbf{u}^T \boldsymbol{S_B} \boldsymbol{u} \\ S.T.\ &\mathbf{u}^T \boldsymbol{S_W} \boldsymbol{u} = 1 \end{aligned} \tag{36}$$

La lagrangiana di tale problema è:

$\mathcal{L} = \mathbf{u}^T \boldsymbol{S_B} \boldsymbol{u} - \lambda(\mathbf{u}^T \boldsymbol{S_W} \boldsymbol{u} - 1)$; λ è il moltiplicatore di Lagrange e ponendo la derivata parziale rispetto a $\mathbf{u}$ a 0 otteniamo :

$$\begin{aligned} \mathbb{R}^d \ni \frac{\partial \mathcal{L}}{\partial \mathbf{u}} &= 2\boldsymbol{S_B} \boldsymbol{u} - 2\lambda \boldsymbol{S_W} \boldsymbol{u} = \mathbf{0} \\ &\Rightarrow \boldsymbol{S_B} \boldsymbol{u} = \lambda \boldsymbol{S_W} \boldsymbol{u} \end{aligned} \tag{37}$$

Questo è un problema agli autovettori e autovalori ($\boldsymbol{S_B}, \boldsymbol{S_W}$). La soluzione $\mathbf{u}$ è l'autovettore corrispondente al più grande autovalore λ . $\mathbf{u}$ è la

direzione di Fisher o la *prima funzione discriminante lineare* . La proiezione e la ricostruzione dei dati viene effettuata con (9) e (10) (dove $\mathbf{U} \in \mathbb{R}^{d \times p}$ è in questo caso il vettore $\mathbf{u} \in R^d$ in quanto ora $p = 1$)

$$\begin{aligned} \boldsymbol{S_B u} = \lambda \boldsymbol{S_W u} \Rightarrow \boldsymbol{S_W^{-1} S_B u} = \lambda \mathbf{u} \\ \Rightarrow \mathbf{u} = \mathbf{eig}(\boldsymbol{S_W^{-1} S_B}) \end{aligned} \tag{38}$$

dove con $\mathbf{eig}(.)$ denotiamo l'autovettore della matrice corrispondente all'autovalore maggiore.

Sottospazio Multidimensionale

Vogliamo ora generalizzare quanto visto fino ad ora (caso il sottospazio dimensionale di Fisher sia un vettore) a sottospazio multidimensionale. Ora il sottospazio di Fisher sarà quindi generato dallo span di più direzioni di Fisher $\{\mathbf{u}_j\}_{j=1}^{p}$, $\mathbf{u}_j \in \mathbb{R}^d$. Estendiamo i problemi e le definizioni visti in precedenzaa:

$$\mathbb{R} \ni d_B := \mathbf{tr}(\mathbf{U}^T \boldsymbol{S_B U}) \tag{39}$$

$$\mathbb{R} \ni d_W := \mathbf{tr}(\mathbf{U}^T \boldsymbol{S_W U}) \tag{40}$$

($\mathbb{R}^{d \times p} \ni \mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_p]$)
(33) diventa :

$$max_{\mathbf{U}} f(\mathbf{U}) := \frac{d_B(\mathbf{U})}{d_W(\mathbf{U})} = \frac{\mathbf{tr}(\mathbf{U}^T \boldsymbol{S_B U})}{\mathbf{tr}(\mathbf{U}^T \boldsymbol{S_W U})} \tag{41}$$

e (36):

$$\begin{aligned} max_{\mathbf{U}} = \mathbf{tr}(\mathbf{U}^T \boldsymbol{S_B U}) \\ S.T.\ \mathbf{U}^T \boldsymbol{S_W U} = \mathbf{I} \end{aligned} \tag{42}$$

La Lagrangiana è $\mathcal{L} = \mathbf{tr}(\mathbf{U}^T \boldsymbol{S_B U}) - \mathbf{tr}(\Lambda^T(\mathbf{U}^T \boldsymbol{S_W U} - \mathbf{I}))$ dove $\Lambda \in \mathbb{R}^{d \times d}$ è una matrice diagonale contenente i moltiplicatori di Lagrange . Ponendo la derivata parziale a 0 otteniamo:

$$\begin{aligned} \mathbb{R}^{d \times p} \ni \frac{\partial \mathcal{L}}{\partial \mathbf{U}} = 2\boldsymbol{S_B U} - 2\boldsymbol{S_W U \Lambda} = \mathbf{0} \\ \Rightarrow \boldsymbol{S_B U} = \boldsymbol{S_W U \Lambda} \end{aligned} \tag{43}$$

Problema agli autovalori $(\boldsymbol{S_B}, \boldsymbol{S_W})$.

Le colonne di $\mathbf{U}$ sono gli autovettori ordinati in ordine decrescente rispetto agli autovalori corrispondenti e definiamo Λ come una matrice diagonale avente come elementi diagonali gli autovalori ordinati sempre in ordine decrescente. Le colonne di $\mathbf{U}$ sono le direzioni di Fisher. Per la proiezione e ricostruzione dei dati richiamiamo (9) e (10) .

La soluzione del problema è:

$$\begin{aligned} \boldsymbol{S_B U} = \boldsymbol{S_W U \Lambda} \Rightarrow \boldsymbol{S_W^{-1} S_B U} = \mathbf{U\Lambda} \\ \Rightarrow \mathbf{U} = \mathbf{eig}(\boldsymbol{S_W^{-1} S_B}) \end{aligned} \tag{44}$$

dove con $\mathbf{eig}(.)$ denotiamo gli autovettori della matrice, disposti sulle colonne, corrispondenti agli autovalori maggiori.

Dimensionalità del sottospazio di Fisher

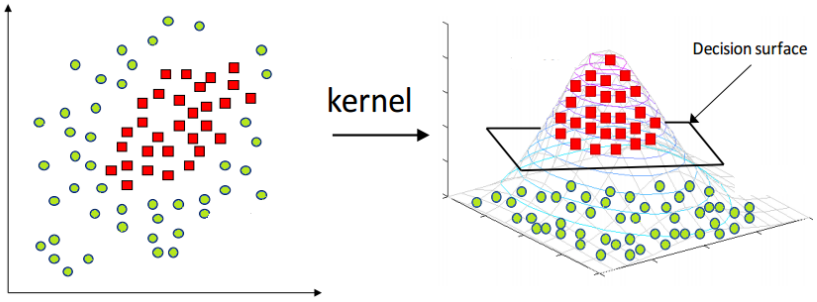
in generale il rango di una matrice di covarianza di dati d-dimensionali calcolata su n campioni è al massimo $\min\{d, n-1\}$. d perchè la matrice è di dimensione $d \times d$, n perchè iteriamo su n dati per calcolarla e -1 perchè ci sottraiamo la media nel calcolarla. Il rango di $\boldsymbol{S_W}$ è perciò $\min\{d, n-1\}$ mentre per $\boldsymbol{S_B}$ dalla definizione ci accorgiamo che è $\min\{d, c-1\}$ Il rango di $\boldsymbol{S_W}^{-1}\boldsymbol{S_B}$ è $c-1$ in quanto solitamente abbiamo $c < d, n$ Possiamo ottenere perciò al massimo $c-1$ autovettori relativi ad autovalori non nulli e quindi è questa la massima dimensione del sottospazio di Fisher. In realtà ci si ferma quando l'autovalore associato diventa trascurabile.

Interpretazione dell' FDA: metafora della vista debole

Questa può essere una metafora interessante per meglio capire il funzionamento del metodo di Fisher. Consideriamo un uomo con 2 problemi di vista : è daltonico e la sua vista è anche molto debole. Supponiamo ci siano due classi di palline: rosse e blu. L'uomo vuole poter distinguere le palline di una classe da quelle dell'altra (ricordiamo che i daltonici distinguono il rosso dal blu solo quando ben separati, non riescono a distinguerli quando sono vicini, faticano con le sfumature). Per risolvere il problema del suo daltonismo separiamo le classi in due set distinti (rosso e blu). Per farlo aumentiamo la distanza tra le palline con colori diverse. Questo equivale ad aumentare la varianza tra le classi ($\boldsymbol{S_B}$). A causa della sua debole vista però per lui è ancora tutto blu, anche se le palline sono quasi separate. Avviciniamo le palline

di una stessa classe le une alle altre e così facendo l'uomo vede ciascuna delle due classi come due palle sfocate e riesce finalmente a distinguerle. Quest'ultimo passaggio equivale a diminuire la varianza intraclasse per ognuna delle due ($\boldsymbol{S_w}$).

Kernel FDA



In alcuni casi si potrebbe non trovare una proiezione in un sottospazio che separi le classi in maniera opportuna, la soluzione è spostare i dati in uno spazio a dimensione molto maggiore e separare le classi con le stesse tecniche ed idee viste fino ad ora ma con più opportunità di trovare sottospazi su cui proiettare i dati, avendo aumentato di molto lo spazio di partenza, e riuscire a separe meglio le classi. Il modo migliore per fare ciò è con l'uso dei Kernels che sfruttano le proprietà degli spazi di Hilbert e il ruolo dei prodotti scalare nelle proiezioni. Riusciremo in questo modo a costruire il nostro classificatore che assegni un dato x alla sua classe di appartenenza.

Kernel e spazi di Hilbert

Sia X il nostro spazio dei dati, allora uno spazio di Hilbert G di funzioni $g : X \to \mathbb{R}$ con prodotto scalare $< ., . >_G$ è chiamato *reproducing kernel Hilbert space* (RKHS) con *reproducing Kernel* $k : X \times X \to \mathbb{R}$ se:

1. $\forall \mathbf{x} \in X, k_x := k(\mathbf{x}, .) \in G$
2. $k(\mathbf{x}, \mathbf{x}) < \infty \ \forall \mathbf{x} \in X$
3. $\forall \mathbf{x} \in X \ e \ g \in G, g(\mathbf{x}) =< g, k_x >_G$ (proprietà di reproducing)

k_x è il rappresentante di valutazione. Il reproducing kernel di uno spazio di Hilbert è unico, è simmettrico ($k(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}', \mathbf{x}')$) e semidefinito positivo:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i k(\mathbf{x}_i,\mathbf{x}_j)\alpha_j \geq 0 \tag{45}$$

Teorema di Moore- Aronszajn
Dato uno spazio non vuoto X e una funzione simmetrica e semidefinita positiva k: $X\times X\to\mathbb{R}$ esiste un RKHS G di funzioni $g: X\to\mathbb{R}$ con reproducing kernel k. G è unico.

Grazie a questo teorema per definire un RKHS su X ci basta specificare il reproducing Kernel $k: X\times X\to\mathbb{R}$ che deve essere finito, simmetrico e semidefinito positivo. Sia $\Phi: \mathbf{x}\to H$ una funzione che mappa i dati in uno spazio di Hilbert, e $\Phi(\mathbf{x})\in\mathbb{R}^t, \mathbf{x}\in\mathbb{R}^d$ con t » d , Per definire l' RKHS ci basta ora definire k nel modo seguente :

$$k(\mathbf{x}_1,\mathbf{x}_2) :=< \Phi(\mathbf{x}_1),\Phi(\mathbf{x}_2) >=_{(a)} \Phi(\mathbf{x}_1)^T\Phi(\mathbf{x}_2) \tag{46}$$

(a) in quanto $< . , . >$ è il prodotto scalare euclideo, il prodotto scalare dello spazio di Hilbert $l^2(\mathbb{R})$ esso è finito, simmetrico e dobbiamo verificare sia semidefinito positivo: sia $\mathbf{\Phi}$ la matrice che ha per righe $\phi(\mathbf{x}_1)^T,...,\phi(\mathbf{x}_n)^T$ e $\boldsymbol{\alpha}=[\alpha_1,...,\alpha_n]^T\in\mathbb{R}^n$ allora :

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i k(\mathbf{x}_i,\mathbf{x}_j)\alpha_j = \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\phi(\mathbf{x}_i)\phi(\mathbf{x}_j)\alpha_j = \boldsymbol{\alpha}^T\mathbf{\Phi}\mathbf{\Phi}^T\boldsymbol{\alpha} =$$

$= \|\mathbf{\Phi}^t\boldsymbol{\alpha}\|_2^2 \geq 0 \Rightarrow k$ semidefinito positivo

Possiamo calcolare il kernel di due matrici $\mathbf{X}_1\in\mathbb{R}^{d\times n_1}$ *e* $\mathbf{X}_2\in\mathbb{R}^{d\times n_2}$ (*kernel matrix*):

$$\mathbb{R}^{n_1\times n_2}\ni\mathbf{K}(\mathbf{X}_1,\mathbf{X}_2) := \mathbf{\Phi}(\mathbf{X}_1)^T\mathbf{\Phi}(\mathbf{X}_2) \tag{47}$$

dove $\mathbf{\Phi(X_1)} := [\phi(\mathbf{x}_1,...,\phi(\mathbf{x}_n)]\in\mathbb{R}^{t\times n_1}$ e allo stesso modo è definita $\mathbf{\Phi(X_2)}$. Il kernel Matrix del nostro dataset sarà:

$$\mathbb{R}^{n\times n}\ni\mathbf{K}_x := \mathbf{K}(\mathbf{X},\mathbf{X}) := \mathbf{\Phi}(\mathbf{X})^T\mathbf{\Phi}(\mathbf{X}) \tag{48}$$

Esistono diversi tipi di kernels, I piu conosciuti sono i seguenti:

$$Linear: k(\mathbf{x}_1,\mathbf{x}_2) = \mathbf{x}_1^T\mathbf{x}_2 + c_1 \tag{49}$$

$$Polynomial: k(\mathbf{x}_1,\mathbf{x}_2) = (c_1\mathbf{x}_1^T\mathbf{x}_2 + c_2)^{c_3} \tag{50}$$

$$Gaussian(RBF\ radial\ basis\ function): k(\mathbf{x}_1,\mathbf{x}_2) = exp(-\frac{\|\mathbf{x}_1-\mathbf{x}_2\|_2^2}{2\sigma^2}) \tag{51}$$

Sottospazio 1 dimensionale

classificazione in 2 classi

L'equazione (21) nel nostro spazio delle funzioni diventa:

$$\mathbb{R}^{t\times t} \ni \boldsymbol{\Phi}(\mathbf{S}_B) := (\phi(\boldsymbol{\mu_1}) - \phi(\boldsymbol{\mu_2}))(\phi(\boldsymbol{\mu_1}) - \phi(\boldsymbol{\mu_2}))^T \tag{52}$$

dove la media della j- esima classe nello spazio di Hilbert delle funzioni è :

$$\mathbb{R}^t \ni \phi(\boldsymbol{\mu_j}) := \frac{1}{n_j}\sum_{i=1}^{n_j}\phi(\mathbf{x}_i^{(j)}) \tag{53}$$

Ogni soluzione (direzione) $\phi(\mathbf{u}) \in H$ è combinazione lineare di tutti i vettori mappati in H, cioè: $\boldsymbol{\Phi}(\mathbf{X}) = [\phi(\mathbf{x}_1), ..., \phi(\mathbf{x}_n)] \in \mathbb{R}^{t\times n}$ (dove solitamente t » d , H denota lo spazio di Hilbert : spazio delle funzioni).

$$\mathbb{R}^t \ni \phi(\mathbf{u}) = \sum_{i=1}^{n}\theta_i\phi(\mathbf{x}_i) = \boldsymbol{\Phi}(\mathbf{X})\boldsymbol{\theta} \tag{54}$$

dove $\mathbb{R}^n \ni \boldsymbol{\theta} := [\theta_1, ..., \theta_n]^T$ è il vettore sconosciuto dei coefficienti e $\phi(\mathbf{u}) \in \mathbb{R}^t$ la direzione di Fishe nello spazio delle funzioni. Le direzioni possono essere compattate in una matrice : $\mathbb{R}^{t\times p} \ni \ \boldsymbol{\Phi}(\mathbf{U}) := [\phi(\mathbf{u}_1), ..., \phi(\mathbf{u}_p)]$:

$$\mathbb{R}^{t\times p} \ni \boldsymbol{\Phi}(\mathbf{U}) = \boldsymbol{\Phi}(\mathbf{X})\boldsymbol{\Theta} \tag{55}$$

dove $\boldsymbol{\Theta} := [\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_p] \in \mathbb{R}^{n\times p}$. d_B diventa :

$$\begin{gathered}\mathbb{R} \ni d_B := \phi(\mathbf{u})^T\boldsymbol{\Phi}(\mathbf{S}_B)\phi(\mathbf{u}) =_{(a)} \\ \boldsymbol{\theta}^T\boldsymbol{\Phi}(\mathbf{X})^T(\phi(\boldsymbol{\mu_1}) - \phi(\boldsymbol{\mu_2}))(\phi(\boldsymbol{\mu_1}) - \phi(\boldsymbol{\mu_2}))^T\boldsymbol{\Phi}(\mathbf{X})\boldsymbol{\theta}\end{gathered} \tag{56}$$

(a) per (52) e (54)

Per la j - esima classe (nel caso in esame ora $j \in \{1, 2\}$) abbiamo :

$$\begin{gathered}\boldsymbol{\theta}^T\boldsymbol{\Phi}(\mathbf{X})^T\phi(\boldsymbol{\mu}_j) =_{(54)} \sum_{i=1}^{n}\theta_i\phi(\mathbf{x}_i)^T\phi(\boldsymbol{\mu}_j) =_{(53)} \\ \frac{1}{n_j}\sum_{i=1}^{n}\sum_{k=1}^{n_j}\theta_i\phi(\mathbf{x}_i)^T\phi(\mathbf{x}_k^{(j)}) =_{(46)} \\ \frac{1}{n_j}\sum_{i=1}^{n}\sum_{k=1}^{n_j}\theta_i k(\mathbf{x}_i, \mathbf{x}_k^{(j)}) = \boldsymbol{\theta}^T\mathbf{m}_j\end{gathered} \tag{57}$$

dove $\mathbf{m}_j \in \mathbb{R}^n$ la cui i-esima entrata è :

$$\mathbf{m}_j(i) := \frac{1}{n_j} \sum_{k=1}^{n_j} k(\mathbf{x}_i, \mathbf{x}_k^{(j)}) \tag{58}$$

La (56) diventa :

$$d_B =_{(57)} \boldsymbol{\theta}(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T\boldsymbol{\theta} = \boldsymbol{\theta}^T\mathbf{M}\boldsymbol{\theta} \tag{59}$$

con:

$$\mathbb{R}^{n\times n} \ni \mathbf{M} := (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T \tag{60}$$

$\mathbf{M}$ è la matrice di covarianza tra le classi nel kernel FDA e otteniamo la seguente :

$$d_B = \phi(\mathbf{u})^T\boldsymbol{\Phi}(\mathbf{S}_b)\phi(\mathbf{u}) = \boldsymbol{\theta}^T\mathbf{M}\boldsymbol{\theta} \tag{61}$$

la matrice di covarianza intraclasse $\mathbf{S}_W$ in (23) nello spazio delle funzioni è:

$$\mathbb{R}^{t\times t} \ni \boldsymbol{\Phi}(\mathbf{S}_W) := \sum_{j=1}^{c}\sum_{i=1}^{n_j}(\phi(\mathbf{x}_i^{(j))} - \phi(\boldsymbol{\mu}_j))(\phi(\mathbf{x}_i^{(j)}) - \phi(\boldsymbol{\mu}_j))^T \tag{62}$$

d_W nello spazio delle funzioni è :

$$d_W := \phi(\mathbf{u})^T \mathbf{\Phi}(\mathbf{S}_W)\phi(\mathbf{u}) =_{(a)}$$

$$(\sum_{l=1}^{n} \theta_l \phi(\mathbf{x}_l)^T)(\sum_{j=1}^{c}\sum_{i=1}^{n_j}(\phi(\mathbf{x}_i^{(j)}) - \phi(\boldsymbol{\mu}_j))(\phi(\mathbf{x}_i^{(j)}) - \phi(\boldsymbol{\mu}_j))^T)(\sum_{k=1}^{n}\theta_k\phi(\mathbf{x}_k)) =$$

$$\sum_{j=1}^{c}\sum_{l=1}^{n}\sum_{i=1}^{n_j}\sum_{k=1}^{n}(\theta_l\phi(\mathbf{x}_l)^T(\phi(\mathbf{x}_i^{(j)}) - \phi(\boldsymbol{\mu}_j))(\phi(\mathbf{x}_i^{(j)}) - \phi(\boldsymbol{\mu}_j))^T\theta_k\phi(\mathbf{x}_k)) =_{(53)}$$

$$\sum_{j=1}^{c}\sum_{l=1}^{n}\sum_{i=1}^{n_j}\sum_{k=1}^{n}(\theta_l\phi(\mathbf{x}_l)^T(\phi(\mathbf{x}_i^{(j)}) - \frac{1}{n_j}\sum_{e=1}^{n_j}\phi(\mathbf{x}_e^{(j)}))(\phi(\mathbf{x}_i^{(j)}) - \frac{1}{n_j}\sum_{z=1}^{n_j}\phi(\mathbf{x}_z^{(j)}))^T\theta_k\phi(\mathbf{x}_k)) =_{(46)}$$

$$\sum_{j=1}^{c}\sum_{l=1}^{n}\sum_{i=1}^{n_j}\sum_{k=1}^{n}(\theta_l k(\mathbf{x}_l, \mathbf{x}_i^{(j)}) - \frac{1}{n_j}\sum_{e=1}^{n_j}\theta_l k(\mathbf{x}_l, \mathbf{x}_e^{(j)}))(\theta_k k(\mathbf{x}_k, \mathbf{x}_i^{(j)}) - \frac{1}{n_j}\sum_{z=1}^{n_j}\theta_k k(\mathbf{x}_k, \mathbf{x}_z^{(j)})) =$$

$$\sum_{j=1}^{c}\sum_{l=1}^{n}\sum_{i=1}^{n_j}\sum_{k=1}^{n}(\theta_l\theta_k k(\mathbf{x}_l, \mathbf{x}_i^{(j)})(k(\mathbf{x}_k, \mathbf{x}_i^{(j)}) - \frac{2\theta_l\theta_k}{n_j}\sum_{z=1}^{n_j}k(\mathbf{x}_l, \mathbf{x}_i^{(j)})(k(\mathbf{x}_k, \mathbf{x}_z^{(j)})$$

$$+\frac{\theta_l\theta_k}{n_j^2}\sum_{e=1}^{n_j}\sum_{z=1}^{n_j}k(\mathbf{x}_l, \mathbf{x}_e^{(j)})(k(\mathbf{x}_k, \mathbf{x}_z^{(j)}) =$$

$$\sum_{j=1}^{c}\sum_{l=1}^{n}\sum_{i=1}^{n_j}\sum_{k=1}^{n}(\theta_l\theta_k k(\mathbf{x}_l, \mathbf{x}_i^{(j)})(k(\mathbf{x}_k, \mathbf{x}_i^{(j)}) - \frac{\theta_l\theta_k}{n_j}\sum_{z=1}^{n_j}k(\mathbf{x}_l, \mathbf{x}_i^{(j)})(k(\mathbf{x}_k, \mathbf{x}_z^{(j)})) =$$

$$\sum_{j=1}^{c}(\sum_{l=1}^{n}\sum_{i=1}^{n_j}\sum_{k=1}^{n}(\theta_l\theta_k k(\mathbf{x}_l, \mathbf{x}_i^{(j)})(k(\mathbf{x}_k, \mathbf{x}_i^{(j)}))$$

$$-\sum_{l=1}^{n}\sum_{i=1}^{n_j}\sum_{k=1}^{n}(\frac{\theta_l\theta_k}{n_j}\sum_{z=1}^{n_j}k(\mathbf{x}_l, \mathbf{x}_i^{(j)})(k(\mathbf{x}_k, \mathbf{x}_z^{(j)}))) =_{(b)}$$

$$\sum_{j=1}^{c}(\boldsymbol{\theta}^T\mathbf{K}_j\mathbf{K}_j^T\boldsymbol{\theta} - \boldsymbol{\theta}^T\mathbf{K}_j\frac{1}{n_j}\mathbf{1}\mathbf{1}^T\mathbf{K}_j^T\boldsymbol{\theta}) = \sum_{j=1}^{c}\boldsymbol{\theta}^T\mathbf{K}_j(\mathbf{I} - \frac{1}{n_j}\mathbf{1}\mathbf{1}^T)\mathbf{K}_j^T\boldsymbol{\theta} =_{(c)}$$

$$\sum_{j=1}^{c}\boldsymbol{\theta}^T\mathbf{K}_j\mathbf{H}_j\mathbf{K}_j^T\boldsymbol{\theta} = \boldsymbol{\theta}^T(\sum_{j=1}^{c}\mathbf{K}_j\mathbf{H}_j\mathbf{K}_j^T)\boldsymbol{\theta} \tag{63}$$

(a) per la (62) e (54) , (b) perchè $\mathbf{K}_j \in \mathbb{R}^{n\times n_j}$è la matrice del kernel di tutti i dati della classe j, in cui l'elemento in posizione (a,b) è :

$$\mathbf{K}_{j(a,b)} := k(\mathbf{x}_a, x_b^{(j)}) \tag{64}$$

e (d) perchè :

$$\mathbb{R}^{n_j \times n_j} \ni \mathbf{H}_j := \mathbf{I} - \frac{1}{n_j}\mathbf{1}\mathbf{1}^T \tag{65}$$

è la matrice che centra i dati, definiamo:

$$\mathbb{R}^{n,n} \ni \mathbf{N} := \sum_{j=1}^{c} \mathbf{K}_j \mathbf{H}_j \mathbf{K}_j^T \tag{66}$$

come la matrice di covarinza intraclasse nel kernel FDA, perciò avremo :

$$d_W = \phi(\mathbf{u})^T \mathbf{\Phi}(\mathbf{S}_W)\phi(\mathbf{u}) = \boldsymbol{\theta}^T \mathbf{N}\boldsymbol{\theta} \tag{67}$$

il criterio di Fisher usando il kernel è:

$$f(\boldsymbol{\theta}) := \frac{d_B(\boldsymbol{\theta})}{d_W(\boldsymbol{\theta})} = \frac{\boldsymbol{\theta}^T \boldsymbol{M}\boldsymbol{\theta}}{\boldsymbol{\theta}^T \boldsymbol{N}\boldsymbol{\theta}} \tag{68}$$

$\boldsymbol{\theta} \in \mathbb{R}^n$ è la direzione di Fisher ottenuta con il Kernel. la soluzione di massimo di (68) è :

$$\boldsymbol{M}\boldsymbol{\theta} = \lambda \boldsymbol{N}\boldsymbol{\theta} \tag{69}$$

Questo è di nuovo un problema agli autovalori ($\mathbf{M}$, $\mathbf{N}$), $\boldsymbol{\theta}$ è l'autovettore corrispondente al maggiore autovalore λ :

$$\boldsymbol{\theta} = \mathbf{eig}(\mathbf{N}^{-1}\mathbf{M}) \tag{70}$$

Nelle formule di ricostruzione $\mathbf{\Phi}(\mathbf{X})$ non è sempre ottenibile e nella KFDA non si può eseguire la ricostruzione dei dati. Per l'intero set dei dati le proiezioni sono:

$$\mathbb{R}^{1 \times n} \ni \mathbf{\Phi}(\tilde{\boldsymbol{X}}) = \boldsymbol{\theta}^T \mathbf{K}(\mathbf{X}, \mathbf{X}) \tag{71}$$

classificazione in più classi

Nel caso in cui le classi sono più di 2 usiamo lo stesso identico procedimento, ci basta ridefinire alcuni concetti al caso multi classe.

$$\mathbb{R}^{t \times t} \ni \mathbf{\Phi}(\mathbf{S}_B) := \sum_{j=1}^{c} (\phi(\boldsymbol{\mu_j}) - \phi(\boldsymbol{\mu}))(\phi(\boldsymbol{\mu_j}) - \phi(\boldsymbol{\mu}))^T \tag{72}$$

$$\mathbb{R}^t \ni \phi(\boldsymbol{\mu}) := \frac{1}{\sum_{k=1}^{c} n_k} \sum_{j=1}^{c} n_j \phi(\boldsymbol{\mu}_j) = \frac{1}{n} \sum_{i=1}^{n} \phi(\mathbf{x}_i) \tag{73}$$

$$\mathbb{R} \ni d_B := \phi(\mathbf{u})^T \mathbf{\Phi}(\mathbf{S}_B)\phi(\mathbf{u}) =_{(a)}$$
$$\sum_{j=1}^{c} \boldsymbol{\theta}^T \mathbf{\Phi}(\mathbf{X})^T(\phi(\boldsymbol{\mu_j}) - \phi(\boldsymbol{\mu}))(\phi(\boldsymbol{\mu_j}) - \phi(\boldsymbol{\mu}))^T \mathbf{\Phi}(\mathbf{X})\boldsymbol{\theta} \tag{74}$$

(a) per (72) e (54) . Abbiamo poi:

$$\begin{aligned} \boldsymbol{\theta}^T \mathbf{\Phi}(\mathbf{X})^T \phi(\boldsymbol{\mu}) =_{(54)} \sum_{i=1}^{n} \theta_i \phi(\mathbf{x}_i)^T \phi(\boldsymbol{\mu}) =_{(73)} \\ \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{n} \theta_i \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_k) =_{(46)} \\ \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{n_j} \theta_i k(\mathbf{x}_i, \mathbf{x}_k) = \boldsymbol{\theta}^T \mathbf{m}_* \end{aligned} \tag{75}$$

dove $\mathbf{m}_* \in \mathbb{R}^n$ la cui i-esima entrata è :

$$\mathbf{m}_*(i) := \frac{1}{n} \sum_{k=1}^{n} k(\mathbf{x}_i, \mathbf{x}_k) \tag{76}$$

e d_B diventa :

$$d_B = \boldsymbol{\theta}^T \sum_{j=1}^{c} (\mathbf{m}_j - \mathbf{m}_*)(\mathbf{m}_j - \mathbf{m}_*)^T \boldsymbol{\theta} = \boldsymbol{\theta}^T \mathbf{M} \boldsymbol{\theta} \tag{77}$$

con $\mathbf{M}$ (matrice di covarianza tra le classi)

$$\mathbb{R}^{n \times n} \ni \mathbf{M} := \sum_{j=1}^{c} (\mathbf{m}_j - \mathbf{m}_*)(\mathbf{m}_j - \mathbf{m}_*)^T \tag{78}$$

Ci siamo riportati al problema agli autovalori ($\mathbf{M}$, $\mathbf{N}$) con queste piccole differenze

Sottospazio multidimensionale

Le matrici di covarianza intraclasse e tra le classi restano invariate ma il criterio di Fisher è leggermente differente

Innanzitutto:

$$d_B = \mathbf{tr}(\phi(\mathbf{U})^T \mathbf{\Phi}(\mathbf{S}_B)\phi(\mathbf{U})) = \mathbf{tr}(\mathbf{\Theta}^T \boldsymbol{M} \mathbf{\Theta}) \tag{79}$$

$$d_W = \mathbf{tr}(\phi(\mathbf{U})^T \mathbf{\Phi}(\mathbf{S}_W)\phi(\mathbf{U})) = \mathbf{tr}(\mathbf{\Theta}^T \boldsymbol{N} \mathbf{\Theta}) \tag{80}$$

con $\mathbb{R}^{n\times p} \ni \mathbf{\Theta} = [\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_p]$ il criterio di Fisher diventa:

$$f(\mathbf{\Theta}) := \frac{d_B(\mathbf{\Theta})}{d_W(\mathbf{\Theta})} = \frac{\mathbf{tr}(\mathbf{\Theta}^T \boldsymbol{M} \boldsymbol{\theta})}{\mathbf{tr}(\boldsymbol{\theta}^T \boldsymbol{N} \boldsymbol{\theta})} \tag{81}$$

dove le colonne di $\mathbf{\Theta}$ sono le direzioni di Kernel Fisher . La soluzione che massimizza il criterio è:

$$\boldsymbol{M}\mathbf{\Theta} = \boldsymbol{N}\mathbf{\Theta}\mathbf{\Lambda} \tag{82}$$

problema agli autovalori ($\mathbf{M}$, $\mathbf{N}$). Le colonne di $\mathbf{\Theta}$ sono gli autovettori ordinati relativi agli autovalori contenuti nella matrice diagonale $\mathbf{\Lambda}$ ordinati in ordine decrescente.

Le proiezioni sono:

$$\mathbb{R}^{p\times n} \ni \mathbf{\Phi}(\tilde{\boldsymbol{X}}) = \mathbf{\Theta}^T \mathbf{K}(\mathbf{X}, \mathbf{X}) \tag{83}$$

Assegnazione di x

Data una nuova $\mathbf{x}$ non appartenente al campionamento e di cui quindi non conosciamo la classe di appartenenza come la assegnamo utilizzando le direzioni o sottospazi di Fisher? Dato un sottospazio di Fisher t-dimensionale esso sarà associato alle prime t funzioni discriminanti lineari. Per q-esima funzione discriminante lineare si intende

$$\mathbb{R} \ni w_{(q)}(\mathbf{x}) := \mathbf{u}_{(q)}^T \mathbf{x} \tag{84}$$

dove $\mathbf{u}_{(q)}$ è l'autovettore associato al q-esimo autovalore, coincidente quindi con la q-esima direzione del sottospazio di Fisher. L'osservazione x è assegnata alla classe $j*$ tale che calcolato $w_{(q)}(x)$(cioè proiettando x sulla q-esima direzione $u_{(q)}$) per q=1,...,t abbiamo

$$\mathbb{R} \ni \sum_{q=1}^{t} |w_{(q)}(\mathbf{x}) - \hat{W}_{(q),j*}| = min_j \sum_{q=1}^{t} |w_{(q)}(\mathbf{x}) - \hat{W}_{(q),j}| \tag{85}$$

dove $\hat{W}_{(q),j}$ è la media di $W_{(q)} = \mathbf{u}_{(q)}^T \mathbf{X}$ nel campionamento della j-esima classe. Il mio punto lo assegno cioè alla classe per cui è minima la distanza da questo punto dal baricentro di quella classe nel nuovo sistema di riferimento.