

L'OGGETTO DI PROBLEMI DI CLASSIFICAZIONE È QUELLO DI RICEVERE UN VETTORE \bar{x} IN INPUT E ASSEGNARE AD esso SPECIFICA TRA K CLASSI.

UN'UNA CLASSE È DENOMINATA C_u PER $u = 1, \dots, K$.

NEL CASO PIÙ SEMPLICE SI HA CLASSIFICAZIONE DI UNA SINGOLA CLASSE, DOVE OGNI UNICO VALORE È APPARTENENTE ESATTAMENTE A UNA SPECIFICA CLASSE.

PUNTI SI VA A SOTTODERIVARE IL SPazio DELLE VARIABILI IN INGRESSO IN REGIONI DI DECISIONE DOVE I LIMITI DI QUESTE REGIONI SONO DEFINITI SUPERFICI DI DECISIONE. NEL CASO DI CLASSIFICATORI LINEARI SI HA CHE QUESTE SUPERFICI DI DECISIONE SONO FUNZIONI LINEARI DEL VETTORE DI INPUT \bar{x} .

INFINE, QUANDO RISULTANO A SEPARARE PERFETTAMENTE LE CLASSI TRAMITE QUESTE SUPERFICI DI DECISIONE LINEARI, SI DICE CHE I DATI SONO LINIARMENTE SEPARABILI.

LA PIÙ SEMPLICE FUNZIONE DISCRIMINANTE PER LA DISTINZIONE DI DUE CLASSI, È LA SEGUENTE FUNZIONE:

$$f(\bar{x}) = \bar{w}^T \bar{x} + w_0$$

Dove \bar{w} È IL VETTORE DEI PESI E w_0 È UN COSTANTE.

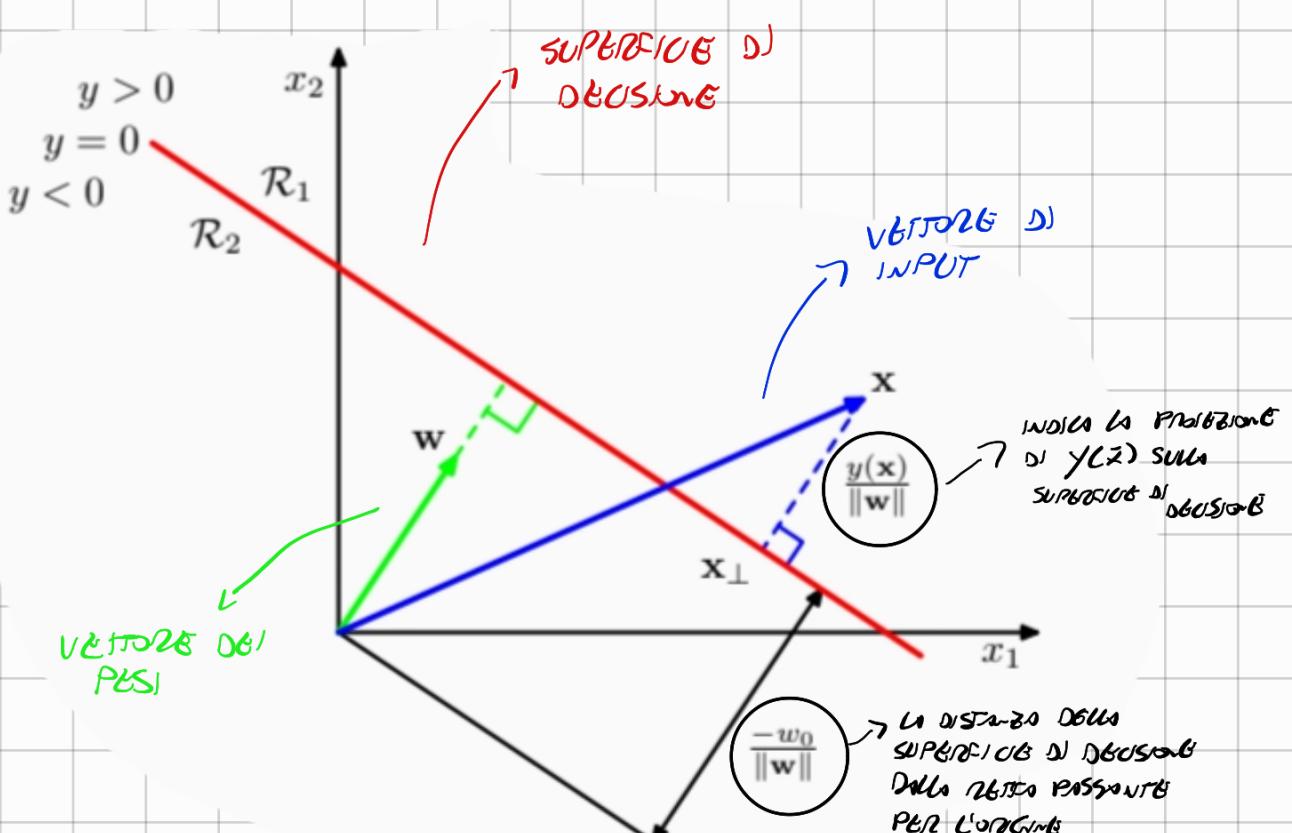
che rappresenta il **BIAS**, definiamo quindi le **REGOLE DI DECISIONE** di appartenenza ad una certa classe:

$$\text{class}(x) = \begin{cases} C_1 & \text{se } \bar{w}^T x + w_0 > 0 \\ C_2 & \text{se } \bar{w}^T x + w_0 \leq 0 \end{cases}$$

E definiamo la **DISTANZA normale dall'origine alla superficie di decisione** in questo modo:

$$\frac{\bar{w}^T x}{\|\bar{w}\|} = \frac{-w_0}{\|\bar{w}\|}$$

GRADIMENTO:

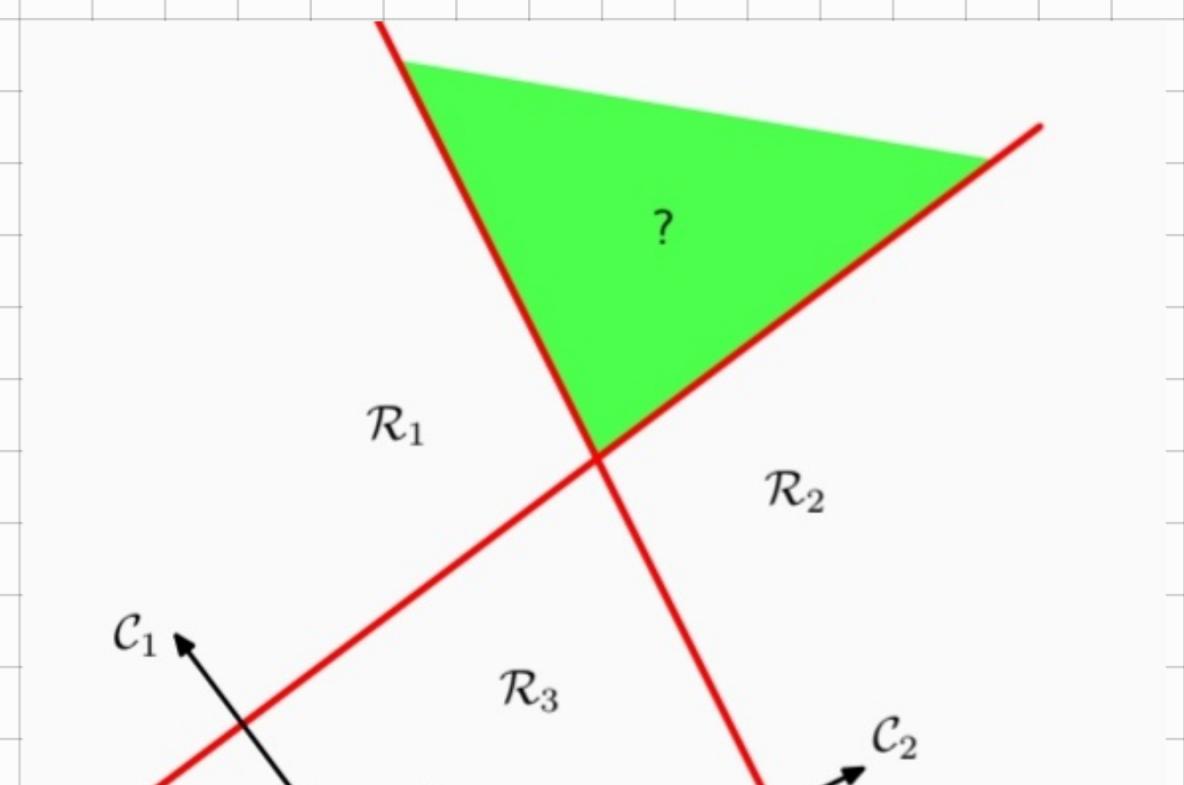


AL FINE DI UTILIZZARLE CON IPIAZIONE PIÙ SEMPLICI E
CONNETTI, SI USANO INTRACCIARE CON DUMMY DIMENSION $x=1$:

$$\begin{aligned} \tilde{w} &= [w_0, \bar{w}]^T \\ \tilde{x} &= [x_0 = 1, \bar{x}]^T \end{aligned} \quad \left\{ \Rightarrow y(\bar{x}) = \tilde{w}^T \tilde{x} \right.$$

CLASSIFICATORI MUOVI-CLASSE

NEL CASO IN CUI VOLIAMO CLASSIFICARE IN PIÙ CLASSE,
È POSSIBILE UTILIZZARE K-1 CLASSIFICATORI, OSSIA
DEI QUALI NSOLVE UN PROBLEMA DI CLASSIFICAZIONE A DUE
CLASSE SEPARANDO PUINDI LA CLASSE C_k DA TUTTI I PLURI
CHE NON LE APPARTENGONO



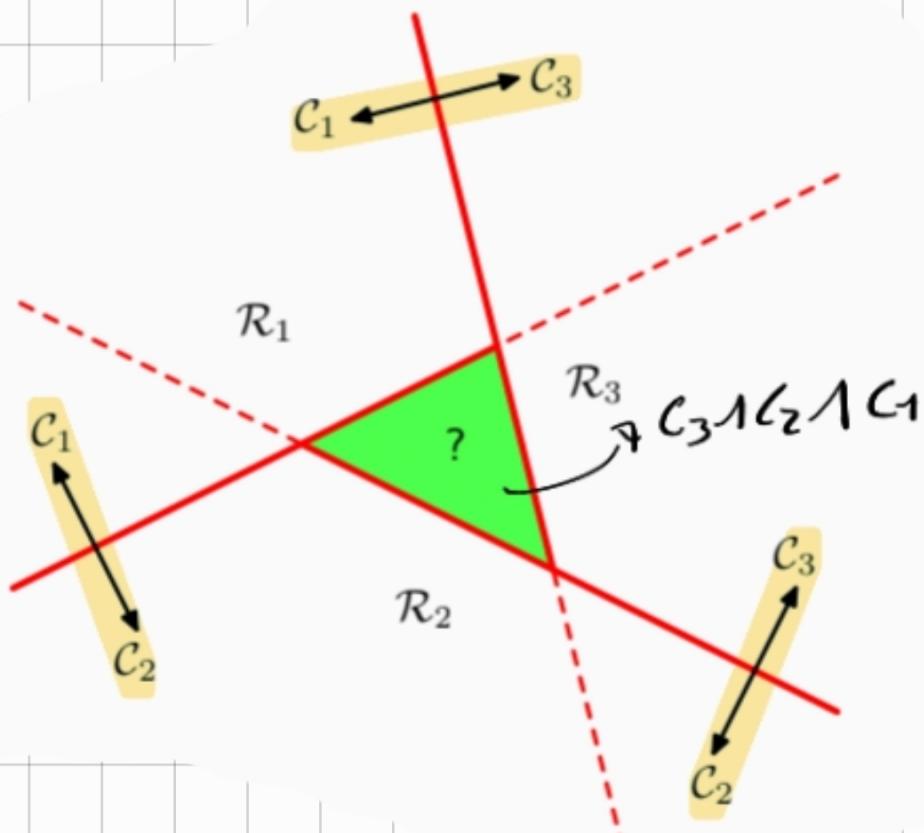
not C_1

not C_2

Sono definiti **CLASSIFICATORI** uno vs tutti. Gli si vorrà
possibile suddividere lo spazio utile secondo

$$U(U-1)/2$$

discriminanti binari, avendo una per ciascuna coppia di
classi;



Tuttavia è possibile semplificare tutti questi
procedimenti riducendo a definite **CLASSIFICATORI**
che riesce quindi a discriminare **UN'ONDE classe**!

$$Y_u(\bar{x}) = \tilde{W}_u' \bar{x} + W_{u0}$$

IL QUALE POSSIAMO RAPPRESENTARE UTILIZZANDO UN
PRODOTTO DI MATE:

$$Y_u(\bar{x}) = \tilde{W}' \bar{x} \Rightarrow \begin{bmatrix} Y_1(\bar{x}) \\ \vdots \\ Y_u(\bar{x}) \end{bmatrix} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ \tilde{w}_1 & \dots & \tilde{w}_u \\ 1 & 1 & \dots & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$$

È PUNTO, IN QUESTO CASO, SI VUOLE ASSOCIARE IL
VALORE SPECIFICA CLASSE C_u SE:

REGOLE DI DECISIONE
GENERALI

$$Y_u(\bar{x}) \geq Y_j(\bar{x})$$

$$\text{PEN } J+u$$

MINIMI PUNTI PER LA CLASSIFICAZIONE

SI UTILIZZANO I MINIMI PUNTI NELLA REGRESSIONE AL
FINO DI OTTENERE UNA FORMA CHIUSA SIMPLICE, PER
PUNTO NUMERO LA CLASSIFICAZIONE:

CATEGORIE CLASSE E' DESCRITA DAL PROPRIO MODELLO
UNICO:

$$y_u(\bar{x}) = \bar{w}_u^T \bar{x} + w_{u0}$$

OPPURE:

$$y_u(\bar{x}) = \tilde{w}^T \tilde{x}$$

MATRICE \tilde{W} CONTIENE TUTTI
I CLASSIFICATORI DELLE K CLASSE!

\tilde{x}	y
$\bar{x}_1 \in C_1$	$[1 \ 0 \ 0]$
$\bar{x}_2 \in C_2$	$[0 \ 1 \ 0]$
$\bar{x}_3 \in C_3$	$[0 \ 0 \ 1]$

PENSIAMO POSSIAMO RISOLVERE PER \tilde{W} MINDO A
MINIMIZZARE LA SOMMA DEGLI ERRORE QUADRATICO. SE SI
CONSIDERA UN CERTO TRAINING SET

$$\{\bar{x}_n, \bar{t}_n\}$$

$$n \in \{1, \dots, N\}$$

E definissons une matrice \tilde{T} dove l' n -esima riga
è composta dai valori di t_m^T insieme alla matrice
corrispondente \tilde{X} dove l' m -esima riga è data da \tilde{X}_m^T
possiamo allora scrivere la funzione che descrive
l'errore come:

$$E(\tilde{W}) = \frac{1}{2} \text{Tr} \left\{ (\tilde{X}\tilde{W} - T)^T (\tilde{X}\tilde{W} - T) \right\}$$

Che si risolve impostando il $\nabla(\tilde{W}) = 0$:

$$\tilde{W} = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T T = \tilde{X}^+ T$$

→ PSEUDO-INVERSA DI
PENROSE-MOULÉ (?)

E quindi abbiamo ottenuto una forma normale per
la classificazione a n -classi:

$$y(\tilde{x}) = \tilde{W}^T \tilde{x} = T^T (\tilde{X}^+)^T \tilde{x}$$

!

DISCRIMINANTE LINEARE DI Fisher

Possiamo affermare che la classificazione lineare tramite
l'utilizzo di discriminanti consiste nella riduzione della
dimensionalità dei dati ad un vettore uno-dimensionale.

SE ANDIAMO QUINDI A CALCOLARE LA MEDIA DI OGNI CLASSE NELLO SPazio DEGLI FEATURES (**CENTROIDI**) OTTIENEMO, NEL CASO BIPARTITO:

$$\bar{m}_1 = \frac{1}{N_1} \sum_{n=1}^{N_1} \bar{x}_{1,n} ; \quad \bar{m}_2 = \frac{1}{N_2} \sum_{n=1}^{N_2} \bar{x}_{2,n}$$

ANDIAMO QUINDI A CALCOLARE \bar{w} IN MODO TALE CHE ANDIAMO A PROIECTARE I CENTROIDI SU \bar{w} USANDO A MASSIMIZZARE LA DISTANZA!

$$\bar{w}^T \bar{m}_2 - \bar{w}^T \bar{m}_1 = \bar{w}^T (\bar{m}_2 - \bar{m}_1)$$

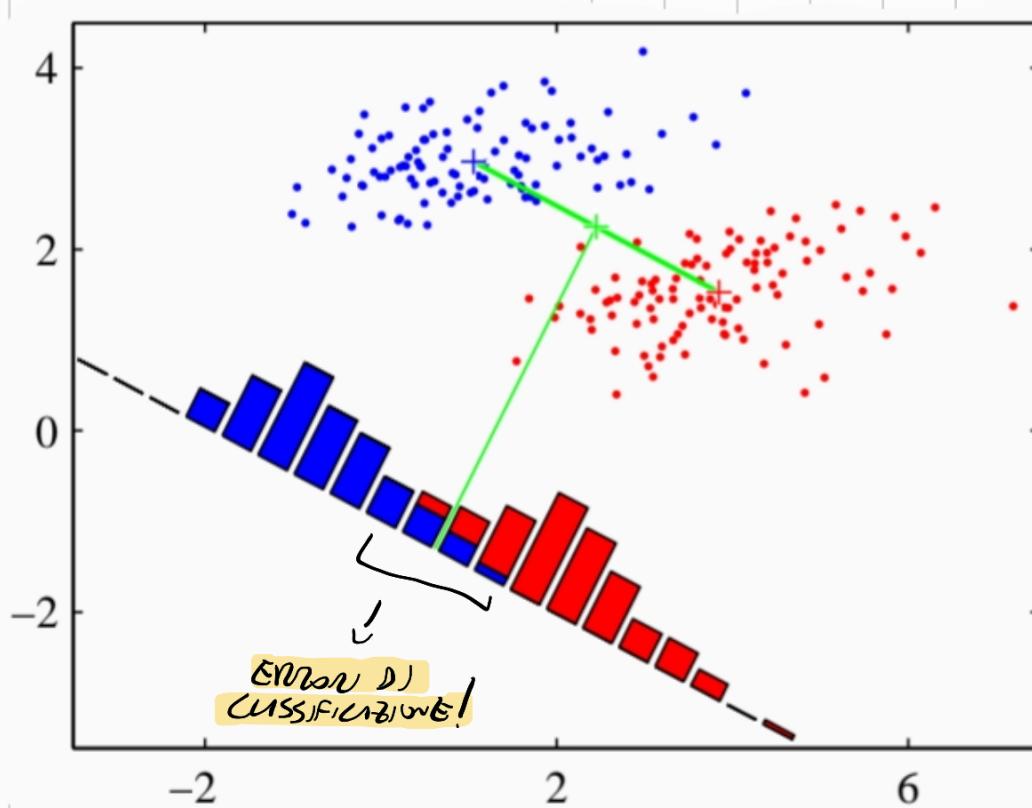
PRESA QUESTA FORMULAZIONE DIVENTEREBBE LA SOLUZIONE OTTIMALE
NUOVO A MASSIMIZZARE LA PROIEZIONE $\bar{w}^T (\bar{m}_2 - \bar{m}_1)$, MA QUESTO
PREVEDEREbbe IL FAITO CHE POTREMMO DIVIDERE PIAZZI SÌ
VOLTE IL VERSO DEI PARAMETRI \bar{w} E PUSSER DIVERSAMENTE NON È
UN LAVORAMENTO CORRETTO AL FINE DEL RISULTATO.

PER QUESTO SI INTRODUCCE IL VERSO $\|\bar{w}\|_2 = 1$ OTTENENDO CHE:

$$\bar{w} \propto (\bar{m}_2 - \bar{m}_1)$$

DIVERO CHE IL VETTORE \bar{w} È **PARALLELO** AL VERSO CHE CONCERNITO I DUE CENTROIDI OTTENUTI DALLE DUE CLASSI!

MOLTI PROBLEMI SONO, CHE HANNO DI SEVERE CONSIDERAZIONI
SOTTOVIA LI MIGLIOR DATASET, SONO TUTTAVIA MOLTO
CONFERITO:



QUESTO È DOVUTO AL Fatto CHE ENTRAMBE LE CLASSI HANNO
MOLTO DI COMUNI MA NON DISTINTI!

COSÌ SONO INVISIBILI!

L'IDEA DI FISHER È QUINDI QUESTA DI:



MAXIMIZZARE LA VARIANZA INTER-CLASSE \rightarrow DISTINGUERE LE CLASSI



MINIMIZZARE LA VARIANZA INTRACLASSE \rightarrow RIACCORRIRE I PUNTI ALLA CLASSE

POSSIAMO QUINDI DEFINIRE LA VARIANZA INTRACLASSE, DETTA
ANCHE COMPATTEZZA:

$$S_K^2 = \sum_{n=1}^{N_K} (\bar{w}^T \bar{x}_{K,n} - w^T m_K)^2$$

E quindi il **CITERIO DI FISHER** è il rapporto tra le due varianze:

$$J(\bar{w}) = \frac{(\bar{w}^T \bar{m}_2 - \bar{w}^T \bar{m}_1)^2}{S_1^2 + S_2^2}$$

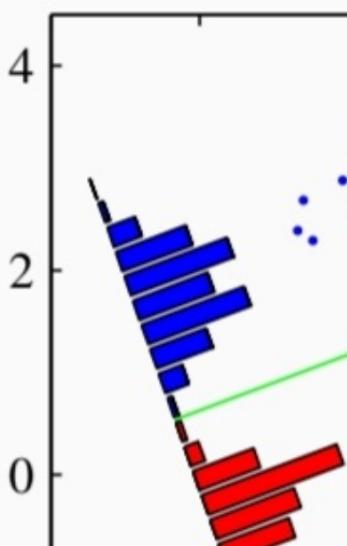
↑ inter-class
↓ intra-class

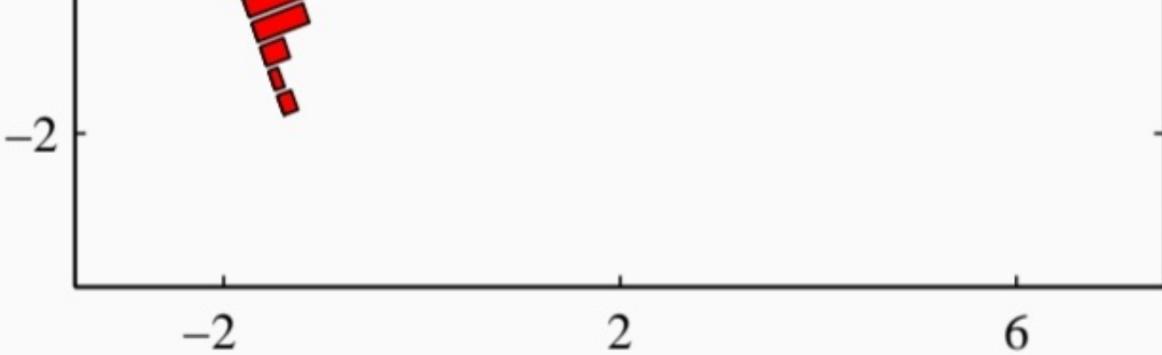
E, utilizzando la covarianza:

$$J(\bar{w}) = \frac{\bar{w}^T \bar{S}_B \bar{w}}{\bar{w}^T \bar{S}_W \bar{w}}$$

↑ covarianza inter-classe
↓ covarianza intra-classe

E questo porta a notevole aumento delle pressioni:





QUINDI, PER QUANTO RIGUARDA LA GEOMETRIA DELLA CLASSIFICAZIONE LINEARE SI HA CHE:

- SI PROGETTA I DATI DI INPUT SUL VETTORE DEI PESI \bar{w}
- IL VETTORE \bar{w} È PURO PERPENDICOLARE ALLA SUPERFICIE DI DECISIONE
- QUASI IL DISCRIMINANTE LINEARE HA RIDOTTO IL PROBLEMA MULTIDIMENSIONALE A UN PROBLEMA UNIDIMENSIONALE!

APPROCCIO PROBABILISTICO ALLA CLASSIFICAZIONE

ABBIAMO FINO AD ORA TRATTO UN MODELLO PER LA CLASSIFICAZIONE MA, COME NEL CASO DELLA REGRESSIONE CON LEAST SQUARES, NOI SIAMO ANCHE IN CASO DI MISURE DI BEGLIE SUI RISULTATI DEL MODELLO OTENUTI. L'OGGETTO È PURÒ QUESTO DI TRASFERIRE UN COLLEGAMENTO TRA LA VISONE GEOMETRICA E QUESTA PROBABILISTICA.

NEL CASO DI CLASSIFICATORI BINARI ($K=2$) POSSIAMO
CALCOLARE IL PROBABILITÀ PER UNA CERTA CLASSE C_1 :

$$P(C_1|\bar{x}) = \frac{P(\bar{x}|C_1)P(C_1)}{P(\bar{x}|C_1)P(C_1) + P(x|C_2)P(C_2)}$$

↳ PER LA REGOLA DI BAYES

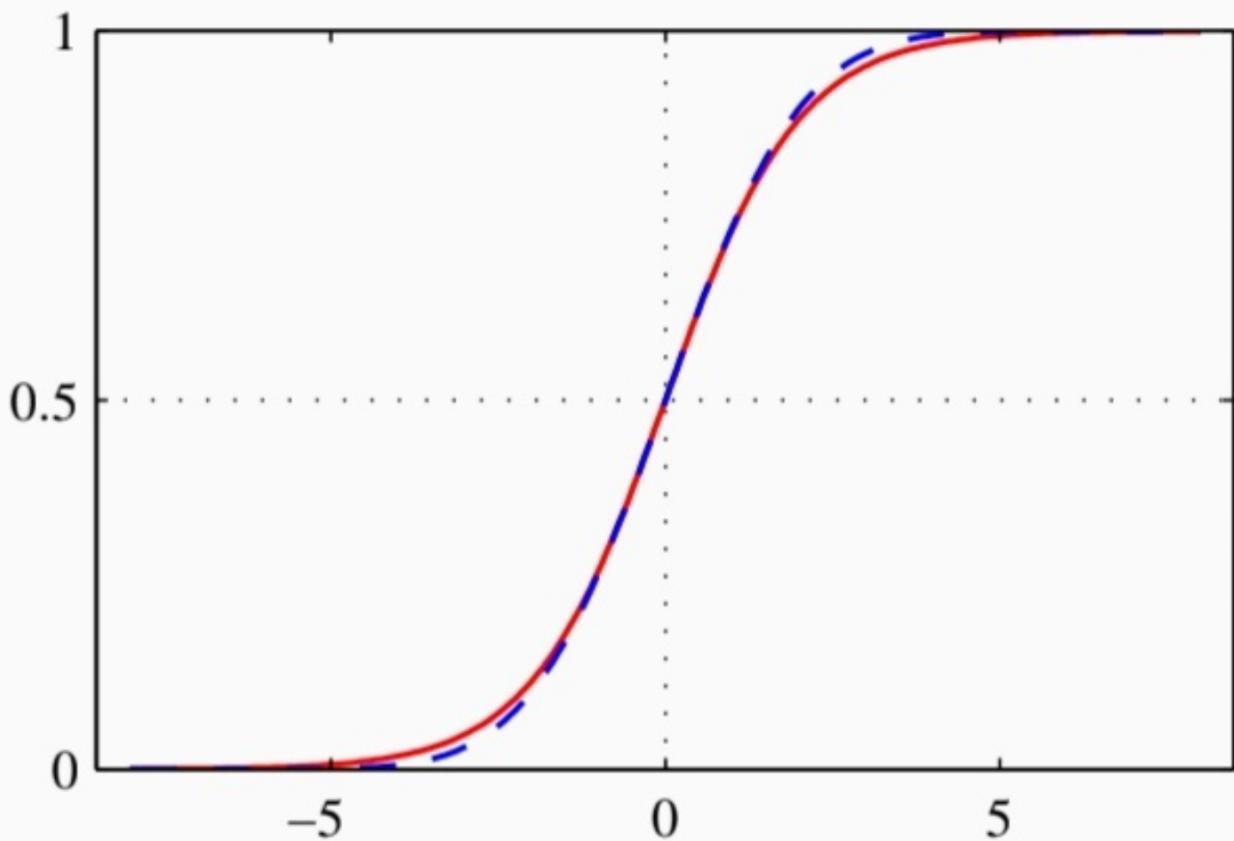
Dove possiamo scrivere che:

$$P(C_1|\bar{x}) = \frac{1}{1 + e^{-\lambda(\bar{x})}} = \sigma(\lambda(\bar{x}))$$

Dove $\lambda(\bar{x}) = \ln \frac{P(\bar{x}|C_1)P(C_1)}{P(\bar{x}|C_2)P(C_2)}$

La funzione $\sigma(\cdot)$ è la **SIGMOIDE LOGISTICA** che, in
pratica, permette di mappare un valore reale in un
INTERVALLO COMPRESO TRA ZERO E UNO

Quindi, nel caso della classificazione, il suo utilizzo è utile
nel caso in cui si voglia calcolare la probabilità
che una certa istanza appartenga ad una certa classe



PUNTO IL SIGMOIDE VIENE UTILIZZATO NEL CASO DI
CLASSIFICAZIONE BINARIA.

NEL CASO IN CUI $K > 2$ SI HA:

$$P(C_k | \bar{x}) = \frac{P(\bar{x} | C_k) P(C_k)}{\sum_j P(\bar{x} | C_j) P(C_j)}$$

$$= \frac{e^{\alpha_k}}{\sum_j e^{\alpha_j}}$$

SI UTILIZZA LA FUNZIONE **SOFTMAX** CHE È LA GENERALIZZAZIONE
DELLA FUNZIONE SIGMOIDE NEL CASO DI PIÙ CLASSI.

POSSIAMO CONSIDERARE CHE LE VARIABILI INCLUSO DELL'CLASSI SIANO DISTRIBUITE COME GAUSSIANE CON IDENTICO MATEMATICO DI COVARIANZA, ANERO:

$$P(\bar{x} | C_k) = \frac{1}{(2\pi)^{D/2} |\Sigma|^{\frac{1}{2}}} \cdot e^{\left\{ -\frac{1}{2} (\bar{x} - \mu_k)^T \Sigma^{-1} (\bar{x} - \mu_k) \right\}}$$

MATEMATICO
 COVARIANZA
 ↓
 VALOR MEDIO DELLA
 CLASSE K

E, considerando solo le due prime classi ottieniamo:

$$P(C_1 | \bar{x}) = \delta(\bar{w}^T \bar{x} + w_0)$$

$$\text{con } w_0 = -\frac{1}{2} \mu_1^T \Sigma^{-1} \mu_1 + \frac{1}{2} \mu_2^T \Sigma^{-1} \mu_2 + \ln \frac{P(C_1)}{P(C_2)}$$

$$E \quad \bar{w} = \Sigma^{-1} (\mu_1 - \mu_2)$$

Oltre la posteriore è quindi una sigmoidale inversa su cui la **FUNZIONE LINEARE**

Dim:

$$\text{sia } \Delta(\bar{x}) = \ln \frac{e^{-\frac{1}{2} f(\bar{x}, \mu_1)}}{e^{-\frac{1}{2} f(\bar{x}, \mu_2)}} \cdot \frac{P(C_1)}{P(C_2)}, \text{ allora:}$$

$$e^{-\frac{1}{2} \{ (\bar{x}, \mu_1) + \frac{1}{2} \{ (\bar{x}, \mu_2) + \ln \frac{P(C_1)}{P(C_2)} \}}$$

$$\mathcal{L}(\bar{x}) = -\frac{1}{2} \{ (\bar{x}, \mu_1) + \frac{1}{2} \{ (\bar{x}, \mu_2) + \ln \frac{P(C_1)}{P(C_2)}$$

$$= \frac{1}{2} [(\bar{x} - \mu_2)^T \Sigma^{-1} (\bar{x} - \mu_2) - (\bar{x} - \mu_1)^T \Sigma^{-1} (\bar{x} - \mu_1)] + \ln \frac{\cdot}{\cdot}$$

$$= \frac{1}{2} [(\bar{x} - \mu_2)^T (\Sigma^{-1} \bar{x} - \Sigma^{-1} \mu_2) - (\bar{x} - \mu_1)^T (\Sigma^{-1} \bar{x} - \Sigma^{-1} \mu_1)] + \ln \frac{\cdot}{\cdot}$$

$$= \cancel{\frac{1}{2} \bar{x}^T \Sigma^{-1} \bar{x} - \mu_2^T \Sigma^{-1} \bar{x} + \frac{1}{2} \mu_2^T \Sigma^{-1} \mu_2} - \cancel{\frac{1}{2} \bar{x}^T \Sigma^{-1} \bar{x} + \mu_1^T \Sigma^{-1} \bar{x} - \frac{1}{2} \mu_1^T \Sigma^{-1} \mu_1} + \ln \frac{\cdot}{\cdot}$$

$$= (\mu_1^T - \mu_2^T) \Sigma^{-1} \bar{x} + \frac{1}{2} \mu_2^T \Sigma^{-1} \mu_2 - \frac{1}{2} \mu_1^T \Sigma^{-1} \mu_1 + \ln \frac{\cdot}{\cdot}$$

\bar{w} w_0

$$\Rightarrow \bar{w} = \Sigma^{-1} (\mu_1 - \mu_2)$$

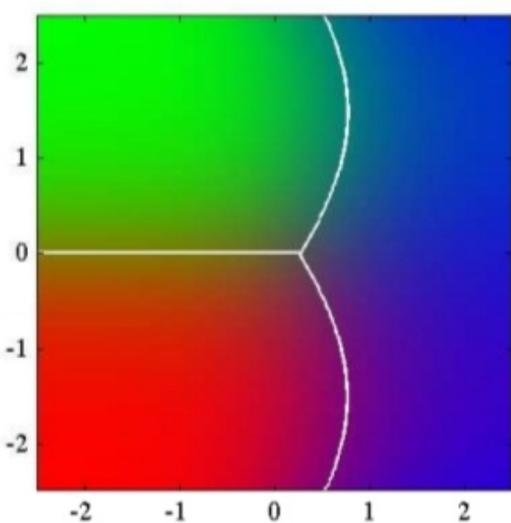
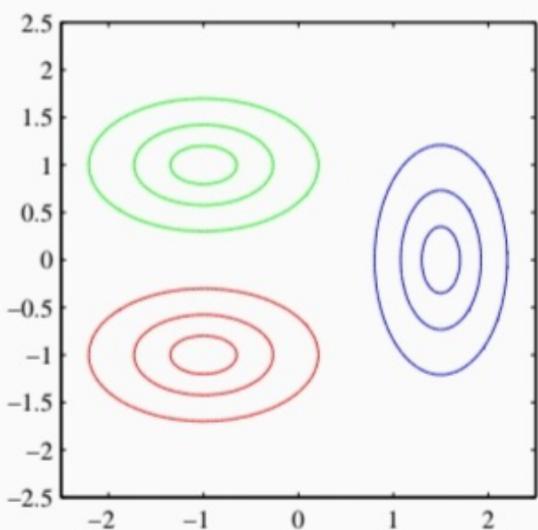
SE INVECE DI CONSIDERARE SOLO UNA CLASSE POSSIAMO CONSIDERARE IL CASO GENERICO DI **M CLASSI** POSSIAMO DIVIDERE USARE IL SOFTMAX INVECE CHE IL SIGMOIDE:

$$\mathcal{L}_W(\bar{x}) = \bar{w}_1^T \bar{x} + w_{k+1} \text{NOVA } \bar{w}_{k+1} = \Sigma^{-1} \mu_k$$

$$W_{HO} = \frac{1}{2} \mu_u^T \Sigma^{-1} \mu_u + \ln p(C_u)$$

E PENSI IL VERSO E CHE LE DECISION BOUNDARIES SONO OTTENUTE DONGE LE POSTERIORI SONO COSTANTI

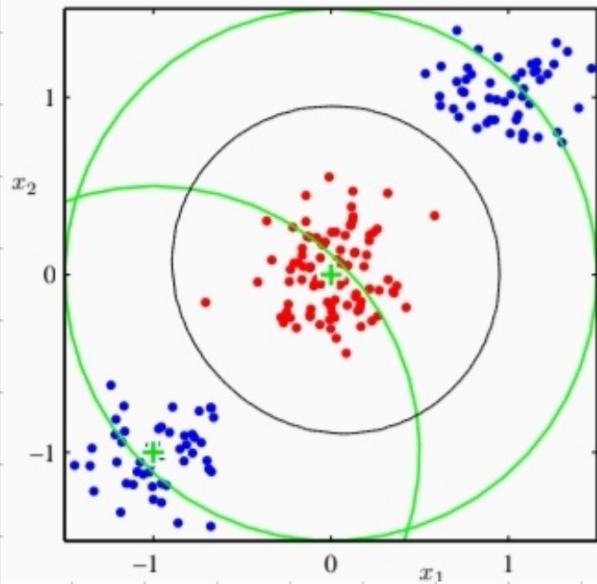
E, INFINE, SE POSSIAMO IL VECCHIO CHE TUTTE LE VARIABILI DI CUI AVEMMO STOCCHE POSSERE UMANO, OTTERIAMO CHE I TERMINI PROPORTIVI SONO SEMPRE PIÙ E PIÙ DISTANZIATI IN CLASSIFICATORI BAYESIANO QUADRATICI



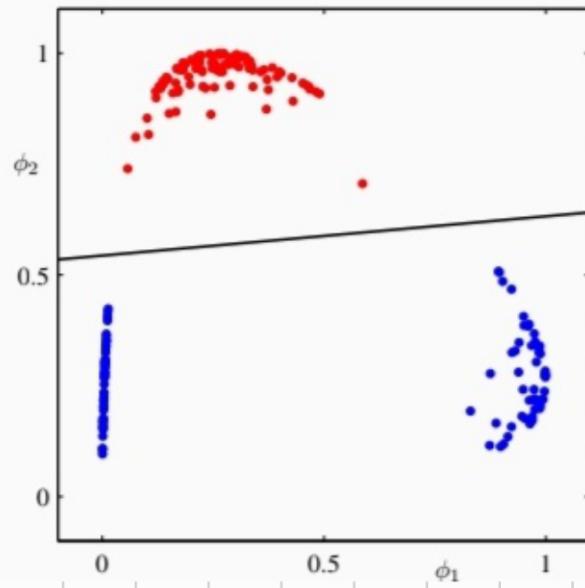
RASSUMENTANDO, ABBIAMO VISTO CHE LA POSTERIORI IN UN PROBLEMA DI CLASSIFICAZIONE BINARIA PUÒ ESSERE SCRITTA COME UNA SIGMOIDE DI UNA FUNZIONE LINEARE, MENTRE NEL CASO MULTICLASSIFICO POSSIAMO RISCONTRARE CONTINUE SOFTMAX DI UNA FUNZIONE LINEARE DI \bar{x}

FINO AD ORA I CLASSIFICATORI SONO STATI SVILUPPATI
MOLTO A LAVORATO DIRETTAMENTE SUL INPUT ORIGINALE
DEL PROBLEMA. TUTTI SONO CONSIDERATI SUL VETTORE
DELE FUNZIONI BASE $\phi(x)$.

QUESTO CI PERMETTE DI COSTRUIRE DELLE BORDURE
CHE SONO LINEARI NELLO SPAZIO DEI PARAMETRI MA
NON LINEARI NELLO SPAZIO ORIGINALE.



↳ UN VENTAGLIO
SEPARABILE



↳ UN VENTAGLIO SEPARABILE
MOLTO A CONSIDERARE UNA
BASE NON LINEARE

QUA SI POSTERIORI CON K=2 DIVENTA:

$$p(c_i | \phi) = \sigma(\bar{w}^\top \phi)$$

DEFINITO MODELLO DI REGRESSIONE LOGISTICA

UTILIZZANDO LA MAXIMUM LIKELIHOOD E' POSSIBILE DETERMINARE

IL PROBLEMA DI QUESTO MASSIMO E PER TUTTO QUESTO SI DEVE PENSARE NELL'ARTE DI DETERMINARE DEL SISTEMA.

$$\frac{d}{d\alpha} \delta(\alpha) = \delta(\alpha) (1 - \delta(\alpha))$$

PENSI PER IL DATASET $D = \{\phi_n, t_n\}$ con $t_n \in \{0, 1\}$ E

$\phi_n = \phi(x)$ UN UNKNOWN E DICO SO:

$$P(\bar{t} | \bar{w}) = \prod_{n=0}^N y_n^{t_n} (1 - y_n)^{1-t_n}$$

Dove $\bar{t} = (t_1, t_2, \dots, t_N)^T$ E $y_n = p(C_1 | \phi_n) = \delta(\bar{w}^T \phi_n)$

E PER LA FUNZIONE DI ERRORE SI UTILIZZA LA LOG-VEROLOGIA:

$$E(\bar{w}) = -\ln P(\bar{t} | \bar{w})$$

POSSIAMO INOLTRE OTTENERE UNA CONVENIENTE FORMA DI LIKELIHOOD E I POSI W:

$$\nabla E(\bar{w}) = \sum_{n=1}^N (y_n - t_n) \phi_n$$

ES $\hat{\epsilon}$ PER PUSS CHE IL MODELLO È ANNUO REGRESSIONE
 CONSTANTE, POICHÉ QUESTA FORMULA È LA STESSA USATA
 NEGLI REGRESSIONI UNIVARI CON FUNZIONE BASE ϕ , DANNO
 IN PUSS CASO L'OBIETTIVO È REGREDIRE I FUORTI
 $t_{m\ell}(0,1) \approx \phi(\bar{x})$

RUTTAVIA, NEL CASO IN CUI C_1 E C_2 SONO LINEARMENTE
 SEPARABILI SI HA CHE LA SOLUZIONE DELLA MAXIMUM
 LIKELIHOOD È SOGGETTA AD OVERFITTING!

QUINDI L'APPROCCIO BAYESIANO ALLA CLASSIFICAZIONE SI PUÒ
 MASSIMAMENTE NEI SEGUENTI PASSI:

- SI DEDE IL PROB $P(\bar{w})$
- SI MASSIMIZZA LA POSTERIORI RISULTANTE AL FINE DI
 OTTENERE UNA DISTRIBUZIONE DEI PARAMETRI $p(\bar{w}|\bar{t})$
- SI MCVA LA DISTRIBUZIONE PREDITTIVA $p(c_i|\bar{w}, \bar{t})$
 CHE POSSIAMO POI APPLICARE SU MANO DA $\phi(\bar{x})$

SE IL PRIMO PASSO RISOLVE SEMPRE, OLTRE DOBBERE
 INTRODUCERE POICHÉ LA POSTERIORI SU PARAMETRI NON È
 PIÙ GAUSSIANA.

POSSIAMO QUINDI ASSUMERE UNA PROBABILITÀ GAUSSIANA

$$p(\bar{w}) = N(\bar{w} | m_0, S_0)$$

POI SI HA AD APPROSSIMARE LA DISTRIBUZIONE DEI PARAMETRI
MENTRE L' APPROXIMAZIONE DI PLACE CHE UTILIZZA PROVA
UN' APPROXIMAZIONE GAUSSIANA:

$$q(\bar{w}) = N(\bar{w} | \bar{w}_{MAP}, \Sigma_w)$$

\rightarrow massima a posteriori

Dove i parametri riuniti dalla MAP STIMANO LA MEDIA E LA
COVARIANZA DARE CONSIDERANDO AD ESEGUIRE UNA MINIMIZZAZIONE DELL'
APPROXIMAZIONE AL SECONDO ORDINE DELLA VERA POSTERIORI

E PENSANDO CON QUESTA APPROXIMAZIONE POSSIAMO OTTENERE LA
DISTRIBUZIONE PREDITTIVA:

$$p(c_1 | \phi, \bar{t}) = \int p(c_1 | \phi, \bar{w}) p(\bar{w} | \bar{t}) d\bar{w}$$

$$= \int \delta(\bar{w}^\top \phi) q(\bar{w}) d\bar{w}$$

OVRORSI LA CONVOLZIONE DI UN SIGMOIDE CON UNA
GAUSSIANA

LINENR SUM

SI DEFINISCE UNA MAPPA BIENNEE UNA FUNZIONE

$\mathcal{R}: V \times V \rightarrow \mathbb{R}$ t.c:

$$\mathcal{R}(\lambda \bar{x} + \psi \bar{y}, \bar{z}) = \lambda \mathcal{R}(\bar{x}, \bar{z}) + \psi \mathcal{R}(\bar{y}, \bar{z})$$

$$\mathcal{R}(\bar{x}, \lambda \bar{y} + \psi \bar{z}) = \lambda \mathcal{R}(\bar{x}, \bar{y}) + \psi \mathcal{R}(\bar{x}, \bar{z})$$

DONG \mathcal{R} E' SIMMETRICA SE $\mathcal{R}(\bar{x}, \bar{y}) = \mathcal{R}(\bar{y}, \bar{x})$ E'

\mathcal{R} E' DEFINITA POSITIVA SE:

$$\mathcal{R}(\bar{x}, \bar{x}) \geq 0 \quad \forall x \in \mathcal{X} \Leftrightarrow \bar{x} = 0$$

SI DEFINISCE L'INNER PRODOTTO SU \mathcal{R} SIMMETRICA E DEFINITA POSITIVA IL PRODOTTO:

$$\boxed{\langle \bar{x}, \bar{y} \rangle} = \mathcal{R}(\bar{x}, \bar{y})$$

MAXIMUM MARGIN CLASSIFIERS

DI SEGUO CONSIDERIAMO LA CLASSIFICAZIONE COME UN

Processo PER CW:

. SI RAPPRESENTANO LE INFORMAZIONI IN \mathbb{R}^D

. SI PARTIZIONA \mathbb{R}^D IN MODO CHE I COMBINI CON LA STESSA ETICHETTA SI MOVINO NELLA STESSA PARTIZIONE

CONSIDERAMO QUINDI LA PARTIZIONE CHE SEPARA LO SPAZIO IN DUE TRAMITE UN IPERPANO

QUINDI SE CONSIDERAMO UNA FUNZIONE $f: \mathbb{R}^D \rightarrow \mathbb{R}$

$$f(\bar{x}, \bar{w}, b) = \langle \bar{w}, \bar{x} \rangle + b$$

POSSIAMO DEFINIRE L'IPERPANO PARTIZIONANTE USANDO f :

$$H = \left\{ \bar{x} \mid f(\bar{x}, \bar{w}, b) = \langle \bar{w}, \bar{x} \rangle + b = 0 \right\}$$

DOVE QUESTO IPERPANO H DEFINITO DA \bar{w} E b È
PERPENDICOLARE A \bar{w} , INFATTI!

CONSIDERARSI $\bar{x}_1, \bar{x}_2 \in H$:

$$\underbrace{\bar{x}_1}_{\text{O}} \quad \underbrace{\bar{x}_2}_{\text{O}} \quad \bar{w}$$

$$f(\bar{x}_1) - f(\bar{x}_2) = \langle \bar{w}, \bar{x}_1 \rangle + b - \langle \bar{w}, \bar{x}_2 \rangle - b$$

$$\Rightarrow \langle \bar{w}, \bar{x}_1 - \bar{x}_2 \rangle = 0 \Rightarrow H \perp \bar{W} \quad \checkmark$$

AL FINE DI CLASSIFICARE UN NUOVO CAMPIONE \bar{x} SI DEVE DECIDERE IN QUALE CLASSE DEGLI IPERPIANO PAGGIO SI VA A TROVARE:

$$\text{class}(\bar{x}) = \begin{cases} 1 & \text{se } f(\bar{x}, \bar{w}, b) \geq 0 \\ -1 & \text{se } f(\bar{x}, \bar{w}, b) < 0 \end{cases}$$

PERMETTE ALGORITMI DI LEARNER SU UN NUOVO DATASET

$$D = \{(\bar{x}_i, y_i) | i=1, \dots, N\}$$

NEL CASO IN CUI $\bar{w} \in b$ SONO TUTTI I CAMPIONI VERSO LA CLASSE VULNERABILE CORRETTO DELL'IPERPIANO:

$$\langle \bar{w}, \bar{x}_i \rangle + b \geq 0 \text{ quando } y_i = 1$$

$$\langle \bar{w}, \bar{x}_i \rangle + b < 0 \text{ quando } y_i = -1$$

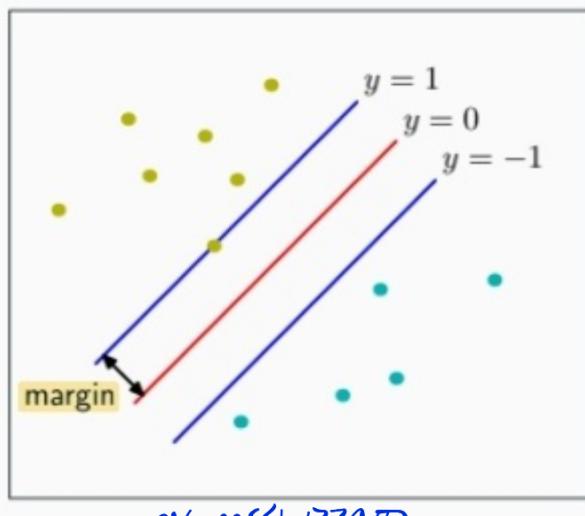
IN MANIERA COMPARSA:

$$y_i (\langle \bar{w}, \bar{x}_i \rangle + b) \geq 0$$

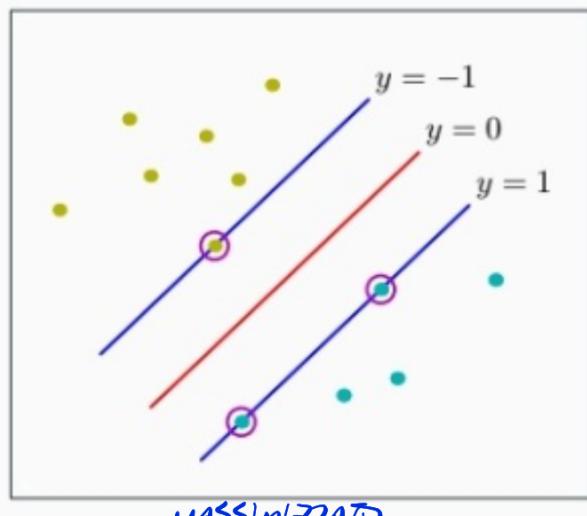
CONSIDERANDO PER CUI
TUTTI I VALORI DEL DATASET D SONO CONGRUITAMENTE

PUNTI DEFINISMO IL MARGINE COME LA DISTANZA TRA L'IPERPLANO ED IL PUNTO DEL DATASET PIÙ vicino all'IPERPLANO

L'OGGETTIVO È MASSIMIZZARE LA DISTANZA!



non massimizzato



massimizzato

MASSIMIZZARE IL MARGINE

SE DEFINISMO LA DISTANZA PERPENDIColare TRA UN PUNTO
PUNTO \bar{x}_n E L'IPERPLANO $(\bar{w}, \bar{x}) + b = 0$ COME:

$$\frac{y_n (\langle \bar{w}, \bar{x}_n \rangle + b)}{\|\bar{w}\|}$$

→ PROIEZIONE DI
 \bar{x}_n SULL'IPERPLANO

L'OGGETTIVO È QUELLO DI MASSIMIZZARE IL MINIMO DI QUESTA DISTANZA!

$$\arg \max_{\bar{w}, b} \left\{ \frac{1}{\|\bar{w}\|} \min_n \left[y_n (\bar{w}, \bar{x}_n) + b \right] \right\}$$

COSÌ I VECCHI CHE I PUNTI DEBBERO ESSERE CORRETTAMENTE CLASSIFICATI:

$$y_n (\bar{w}, \bar{x}_n) + b \geq 0$$

È UN PROBLEMA DI OTIMIZZAZIONE COMPLESSO PER VIA DELLA COESISTENZA DI MAX E MIN!

POSSIAMO TUTTAVIA SCRIVERE \bar{w} E b IN UN QUADRATO
SENZA AVERNE LA DISTANZA TRA I PUNTI E L'IPERPLANO!

POSSIAMO PUNTO SCRIVERE AL FINE DI OTTENERE:

$$\langle \bar{w}, \bar{x}_o \rangle + b = 1$$

PER IL PUNTO PREVIOUS \bar{x}_o

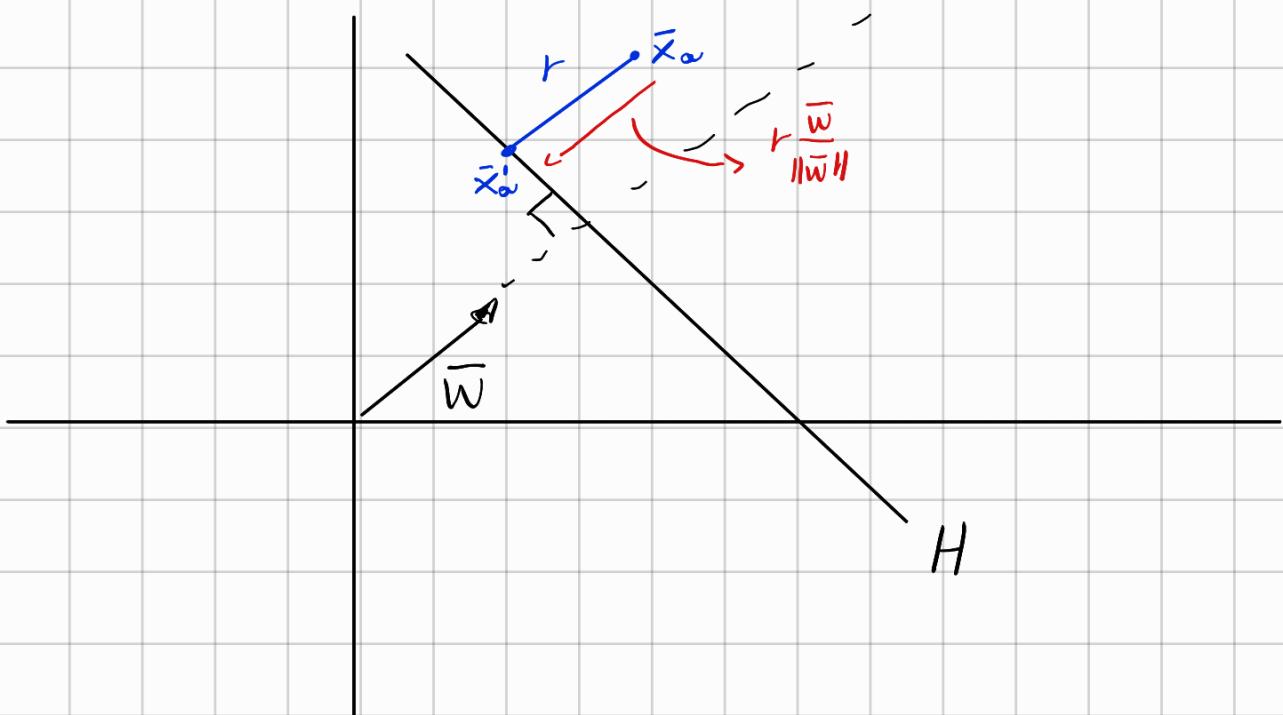
SE CONSIDERO r COME LA DISTANZA ORTOGONALE DA \bar{x}_o ALL'IPERPLANO, POSSO ALLORA CONSIDERARE LA PROIEZIONE ORTOGONALE DI \bar{x}_o SUL'IPERPLANO COME:

$$\bar{x}'_o = \bar{x}_o - r \frac{\bar{w}}{\|\bar{w}\|}$$

→ Proiezione di \bar{x}_o su
(perpend.)

E' considerando che \bar{x}'_o è sul perpend.

$$\left\langle \bar{w}, \bar{x}_o - r \frac{\bar{w}}{\|\bar{w}\|} \right\rangle + b = 0$$



SFRUTTANDO IL PRODOTTO DI BILINEARITÀ:

$$\left\langle \bar{w}, \bar{x}_o \right\rangle + b - r \frac{\left\langle \bar{w}, \bar{w} \right\rangle}{\|\bar{w}\|} = 0$$

E' POSSIBILE CHE \bar{x}_0 SIA SUL MARGINE SE HA CHE $\langle \bar{w}, \bar{x}_0 \rangle + b = 1$

QUINDI SI OTTENE:

$$r = \frac{1}{\|\bar{w}\|}$$

\rightarrow MARGINE

QUINDI VEDIAMO COMBINARE LA MASSIMIZZAZIONE DEL MARGINE CON IL VINCITO DELL'CLASSI OTTENENDO:

$$\max_{\bar{w}, b} \frac{1}{\|\bar{w}\|} \text{ con } y_m(\langle \bar{w}, \bar{x}_m \rangle + b) \geq 1 \quad \forall m=1, \dots, N$$

G VEDRA FORMA CONVEXA DI UNO MARGINE SVM:

$$\min_{\bar{w}, b} \frac{1}{2} \|\bar{w}\|^2 \text{ con } y_m(\langle \bar{w}, \bar{x}_m \rangle + b) \geq 1 \quad \forall m=1, \dots, N$$

\curvearrowleft ESISTE UNICO UNICO
MINIMA

NB: QUINDI QUINDI OTTIENI UN PROBLEMA QUADRATICO CONVEXO
IN D VARIAZIONI CON VINCITO LINEARE

PER RISOLVERE QUESTO PROBLEMA SI USA LA FUNKTONE LEGGERA:

$$L(\bar{w}, b, \bar{x}) = \frac{1}{2} \|\bar{w}\|^2 - \sum_{m=1}^N \alpha_m \{ y_m(\langle \bar{w}, \bar{x}_m \rangle + b) - 1 \}$$

QUESTA FUNZIONE E' DECRESCENTE

now pass to train.

$$\bar{w} = \sum_{m=1}^N \alpha_m y_m \bar{x}_m$$

E

$$\sum_{m=1}^N \alpha_m y_m = 0$$



SOSTITUENDO QUESTE DUE ESPRESSIONI NELLA LEGGE DI KKT:

$$\max_{\alpha} \left\{ \sum_{m=1}^N \alpha_m - \frac{1}{2} \sum_{m=1}^N \sum_{m'=1}^N \alpha_m \alpha_{m'} y_m y_{m'} (\bar{x}_m, \bar{x}_{m'}) \right\}$$

$$\text{con } \alpha_m \geq 0 \quad \forall m = 1, \dots, N$$

QUESTA È UNA RAPPRESENTAZIONE SEMPLIFICATA DELLA MIGLIORE SVM
CHE È UNO DEI PROBLEMI QUADRATICI CONVESO SOLUBILE IN
N VARIABILI CON UNA SOLUZIONE CON COMPLESSITÀ $O(N^3)$

AL FINE DI UTILIZZARE IL CLASSIFICATORE SI SOSTITUISCE \bar{w} NELLA
FUNZIONE DI DECISIONE:

$$f(\bar{x}) = \sum_{m=1}^N \alpha_m y_m (\bar{x}, \bar{x}_m) + b$$

CON LE SEGUENTI CONDIZIONI SODDISFAITE:

$$\alpha_m > 0$$

$$y_m f(\bar{x}_m) \cdot 1 \geq 0$$

$$\alpha_m \{ y_m f(\bar{x}_m) - 1 \} = 0$$

POSSO PER TUTTI GLI m VALORI $\alpha_m > 0$ OPPURE $y_m f(\bar{x}_m) = 1$

E' IL VALORE DI \bar{x}_m PER IL quale $\alpha_m > 0$ E' $y_m f(\bar{x}_m) = 1$

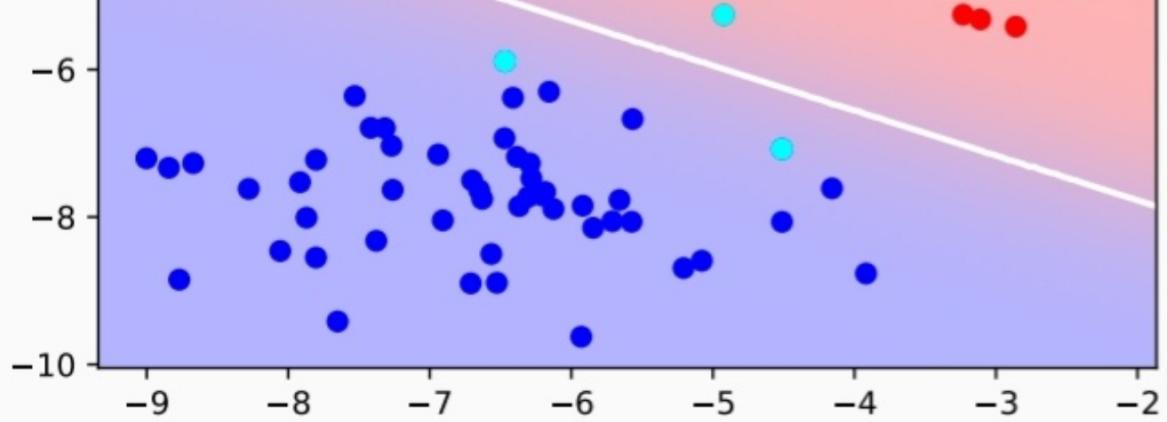
Sono i SUPPORT VECTORS

Sono i SUPPORT VECTORS CONTROBUISSANO ALLE CLASSIFICAZIONI, DEDICO?

$$f(\bar{x}) = \sum_{m=1}^N \alpha_m y_m \langle \bar{x}, \bar{x}_m \rangle + b = \sum_{m \in SV} \alpha_m y_m \langle \bar{x}, \bar{x}_m \rangle + b$$

ABBIAMO PUNTI OTENUTI IN CLASSIFICAZIONE LUNGHE CHE SONO ROBUSTI AI NUOVI DATI





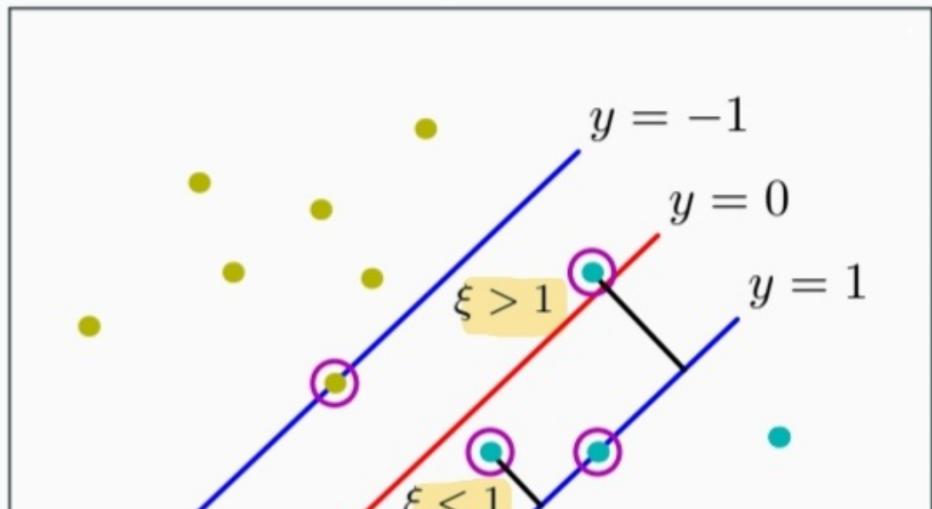
SOFT MARGIN CLASSIFIER

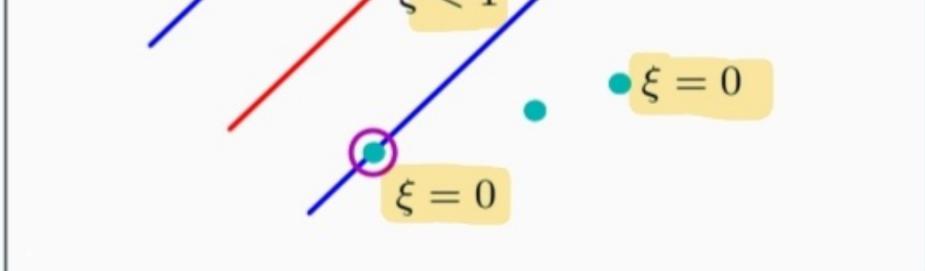
NEL CASO IN CUI IL PROBLEMA NON È LINEARMENTE SEPARABILE SI AVRA' SICURAMENTE UNO PERCENTAGGIO DI ERRORE CLASSIFICATORE.

SI INTRODUCE IL VARIABILE SLACK ξ :

$$\xi = \begin{cases} 0 & \\ |y_m - \langle \bar{w}, \bar{x}_m \rangle - b| & \end{cases}$$

SE \bar{x}_m È NELLA ZONE DEL MARGINE
NON VIENE PUNTO





QUASI IL MENO PROBLEMI DI OTTIMIZZAZIONE DIVERTA:

$$\min_{\bar{w}, b} \frac{1}{2} \|\bar{w}\|^2 + C \sum_{n=1}^N \xi_n$$

regularizzazione

IC OBIOSCE
quanto sia grande
il margine per
avere più
slack

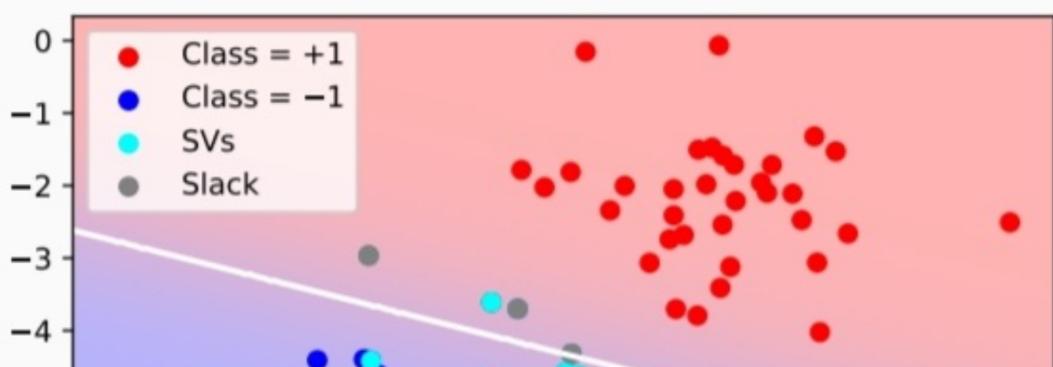
con $y_n (\langle \bar{w}, \bar{x}_n \rangle + b) \geq 1 - \xi$ $\forall n = 1, \dots, N$

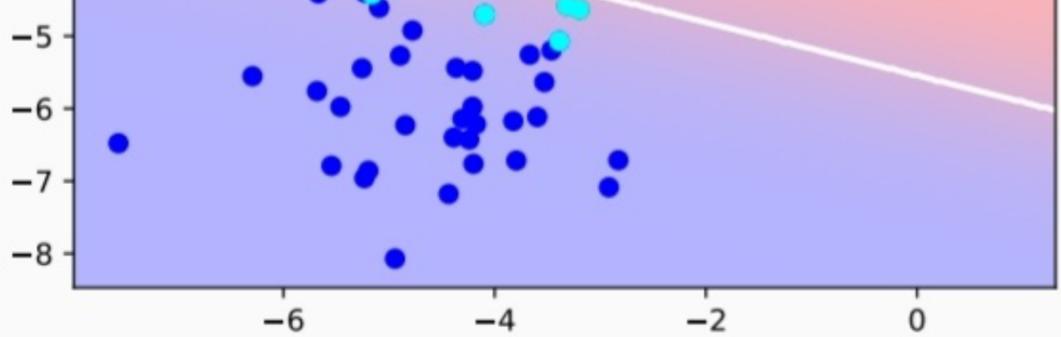
6 SI RISOLVE NELLO STESSO MODO CON LE FORMULE LOGARITMICHE

IN QUESTI FORMI SI INCLUDONO QUASI 1 MILIONE MISCLASSIFIED

DENTRO LA PENALITÀ DI MISCLASSIFICATION SONO CONTATI CON

ξ . IN QUESTO CASO SI PENSÀ A ROBUSTEZZA SULLE OUTLIERS!





PUNTO SVM È ROBUSTA ANCHE DELL'ESO
DI Hard Margin, mentre nel caso di problemi non separabili
si può arrivare al costo per di introdurre un **penalty** (C)
che esige un trade-off tra il costo di missclassificazione
con la massimizzazione del margine

SVM HA IL VANTAGGIO DI ESSERE UN PROBLEMA QUADRATICO
CONVESSO E PUNTO HA SOLUZIONE UNICA ED EFFICIENTE!

