

VEDIAMO COME POTER ESTENDERE I MODELLI LINEARI AL FINITO DI COSTRUIRE **DECISION BOUNDARIES** CON ZONE.

POSSIAMO DERIVARE LE SUPPORT VECTORS MULITIPLICANDO UN **EMPICAL RISK** AL POSTO DI AUMENTARE IL MARGINE.

PER OTTENERE QUESTA VALUTAZIONE SI RICVRA LA **HINGE LOSS**:

$$\mathcal{L}(\tilde{\mathbf{w}}, b, D) = \sum_{n=1}^N \max \left\{ 0, 1 - y_n (\langle \tilde{\mathbf{w}}, \tilde{x}_n \rangle + b) \right\}$$

*! LAGRANGIANA*

$$= \sum_{n=1}^N l(\tilde{x}_n, y_n)$$

*L>LOSS*

*non c'è loss poiché il punto è classificato bene*

$$\text{dove } l(\tilde{x}, y) = \begin{cases} 0 & \text{SE } y(\langle \tilde{\mathbf{w}}, \tilde{x} \rangle + b) \geq 1 \\ 1 - y(\langle \tilde{\mathbf{w}}, \tilde{x} \rangle + b) & \text{SE } y(\langle \tilde{\mathbf{w}}, \tilde{x} \rangle + b) < 1 \end{cases}$$

QUINDI NELLA LOSS VAMO AD INCLUDERE SOLTANTO I VALORI CHE NON SONO CORRETTAMENTE CLASSIFICATI.

UNA VOLTA OTTENUTA LA FORMA ANALITICA DELLA LOSS, È POSSIBILE RIDURRE E RIDUCERE IL COMPLESSITÀ DEL MODELLO INCORPORANDO LA FORMA ANALITICA DELLA REGRESSIONE LINEARE:

$$(\bar{w}^*, b^*) = \underset{\bar{w}, b}{\operatorname{argmin}} \frac{1}{2} \|\bar{w}\|^2 + C \sum_{n=1}^N \max \left\{ 0, 1 - y_n (\langle \bar{w}, \bar{x}_n \rangle + b) \right\}$$

RECOLTORE

LOSS

ABBIAMO QUINDI OTTENUTO UNA FORMULAZIONE IN TERMINI DI LOSS E RECOLTORE, DOVE LA **MINIMIZZAZIONE DEL MARGINE** PIÙ PUÒ OSSERVI VISTA COME UNA FORMA DI RECOLTORE, DOVE L'**IPERPARAMETRO C** È SINTOSS UTILIZZATO PER **PESARE IL LOSS**

ABBIAMO SOLO CONSIDERATO LA PRIMA SVM. CONSIDERIAMO INVECE LA **FORMA DUALE DI SVM**:

$$\max_a \left\{ \sum_{n=1}^N a_n - \frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N a_n a_m y_n y_m \langle \mathbf{x}_n, \mathbf{x}_m \rangle \right\}$$

subject to  $0 \leq a_n \leq C$ , for  $n = 1, \dots, N$

$$\sum_{n=1}^N a_n y_n = 0$$

NOTIAMO CHE L'UNICO USO DEGLI INPUT SAMPLES VIENE FATTO ALL'INTERNO DEGLI **INNER PRODUCTS**. QUINDI, AL FINE DELL'APPRENIMENTO, CI BISOGNA SOLO CALCOLARE GLI INNER PRODUCTS TRA I CAMPIONI  $\bar{x}_m$ .

SÌ, VAI QUINDI AD UTILIZZARE UN **EMBEDDING ESPRESSO**:  $\phi: \mathbb{R}^d \rightarrow \mathcal{H}$

RAPPANDO I DATI IN UN NUOVO SPAZIO, IL NUOVO PROBLEMA DI OTTIMIZZAZIONE DIVENTA QUINDI:

**optimization problem**

→ **INNER PRODOTTO NEL NUOVO SPAZIO**

$$\max_a \left\{ \sum_{n=1}^N a_n - \frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N a_n a_m y_n y_m \langle \phi(\mathbf{x}_n), \phi(\mathbf{x}_m) \rangle_{\mathcal{H}} \right\}$$

subject to  $0 \leq a_n \leq C$ , for  $n = 1, \dots, N$

$$\sum_{n=1}^N a_n y_n = 0$$

GLI EMBEDDINGS DI QUESTO TIPO SONO CHIAMATI **FEATURE MAPS** POICHÉ SIAMO MAPPANDO LA RAPPRESENTAZIONE DI UNA FEATURE IN UNO SPAZIO IN CUI È MAIS RAPPRESENTABILE DELLE FEATURES NEL NUOVO SPAZIO. IL VANTAGGIO È QUINDI QUELLO DI UTILIZZARE COMBINAZIONI NON LINEARI DELLE FEATURES DI INPUT.

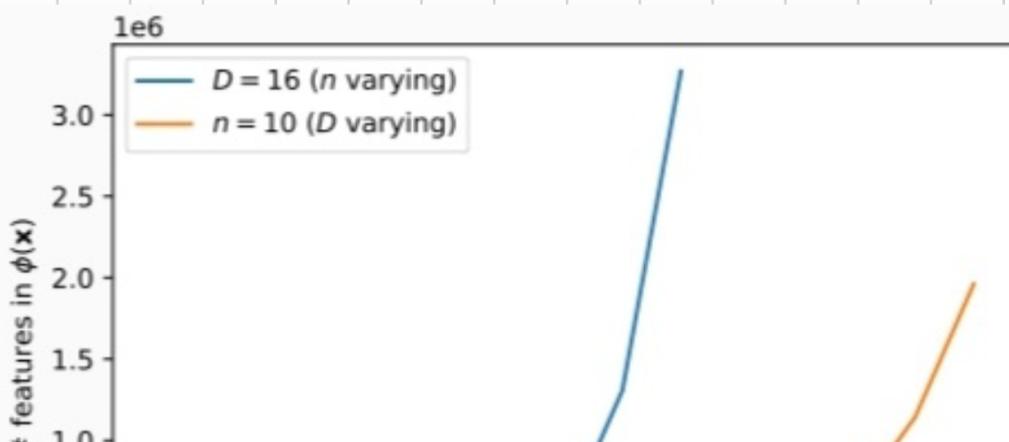
IL CLASSIFICATORE RISULTANTE È QUINDI LINEARE NEL NUOVO SPAZIO DELLE FEATURES, MA NON LINEARE NELLO SPAZIO ORIGINALE.

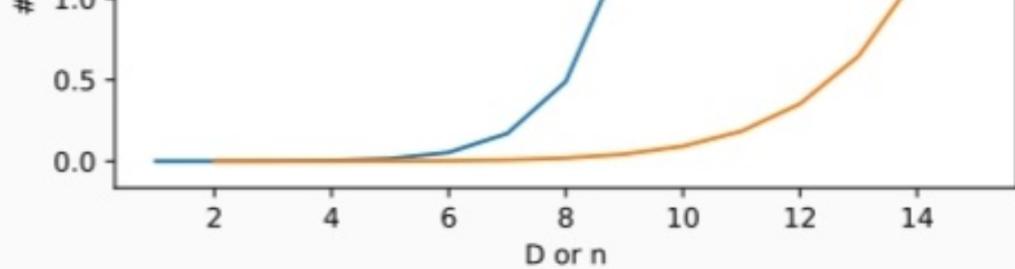
ESEMPPIO:

$$\bar{x} = [x_1, x_2] \in \mathbb{R}^2 \xrightarrow{d=2} \phi(\bar{x}) = \begin{bmatrix} 1 & x_1 & x_2 & x_1^2 & x_2^2 & x_1 x_2 \end{bmatrix}$$

PER I 12 VETTORI DI MONDO È OTTENUTO, PER  $\phi: \mathbb{R}^D \rightarrow \mathbb{R}^M$ :

$$\binom{D+n-1}{n} \frac{1}{(D-1)!} (n+1)^{D-1}$$





SI NOTA CHE AL CRESCERE DEL VALORE DI D aumentano molti LE FEATURES  $\Rightarrow$  rischioso!

AL FINE DI CONTENERE L'ESPLOSIONE DEI DATI SI USA IL **KERNEL TRICK**

### KERNEL TRICK

NELLA NUOVA FORMULAZIONE USATA DELLA **DEA SVM** SI OSSERVA PIÙ  
 CHE NON ABBIANO EFFETTUAMENTE BISOGNO DEI EMBEDDINGS  $\phi(\bar{x}_n)$  E  
 $\phi(\bar{x}_m)$  MA ABBIANO SOLO BISOGNO DI  $\langle \bar{x}_n, \bar{x}_m \rangle$ . SI HA PIÙ  
 BISOGNO DI:

- KERNEL FUNCTION H:**  $k(\bar{x}, \bar{z}) = \langle \phi(\bar{x}), \phi(\bar{z}) \rangle_H$

- GRAM MATRIX (Kernek) H:**  $K_{[n,m]} = \langle \phi(\bar{x}_n), \phi(\bar{x}_m) \rangle_H$

E LA **SIMMETRIA** DELL'INNER PRODUCT IMPLICA CHE  $k \in H$  SONO SIMMETRICHE, COSÌ  
 COME LA PROPIETÀ DI **DEFINITA POSITIVA** IMPLICA CHE  $H$  È UNA MATELLA  
SEMOGLIATIVA POSITIVA  $\Leftrightarrow \bar{x}^T K \bar{x} \geq 0 \quad \forall \bar{x}$

La forma più facile di funzione Kernel è il **KERNEL LINEARE**:

$$k(\bar{x}, \bar{z}) = \bar{x}^T \bar{z} \quad \text{o} \quad \phi(\bar{x}) = \bar{x}$$

quindi, in sostanza, il kernel trick consiste nell'usare le funzioni kernel per sostituire le inner products in moduli non lineari (?)

L'obiettivo sarà sempre di costruire direttamente i kernels senza ricorrere ai embeddings. Ad esempio: ( $V = \mathbb{R}^2$ )

$$\begin{aligned} u(\bar{x}, \bar{z}) &= (\bar{x}^\top \bar{z})^2 = (x_1 z_1 + x_2 z_2)^2 = \dots \\ &= (x_1^2, \sqrt{2}x_1 x_2, x_2^2) (z_1^2, \sqrt{2}z_1 z_2, z_2^2)^\top \end{aligned}$$

mentre  $u(\bar{x}, \bar{z}) = (\bar{x}^\top \bar{z})^2$  corrisponde all'embedding polinomiale  
 $\phi(\bar{x}) = (x_1^2, \sqrt{2}x_1 x_2, x_2^2)$

possiamo considerare vari tipi di kernels invertendo conseguentemente la semidefinizione positiva:

$$k(\mathbf{x}, \mathbf{x}') = ck_1(\mathbf{x}, \mathbf{x}') \quad (6.13)$$

$$k(\mathbf{x}, \mathbf{x}') = f(\mathbf{x})k_1(\mathbf{x}, \mathbf{x}')f(\mathbf{x}') \quad (6.14)$$

$$k(\mathbf{x}, \mathbf{x}') = q(k_1(\mathbf{x}, \mathbf{x}')) \quad (6.15)$$

$$k(\mathbf{x}, \mathbf{x}') = \exp(k_1(\mathbf{x}, \mathbf{x}')) \quad (6.16)$$

$$k(\mathbf{x}, \mathbf{x}') = k_1(\mathbf{x}, \mathbf{x}') + k_2(\mathbf{x}, \mathbf{x}') \quad (6.17)$$

$$k(\mathbf{x}, \mathbf{x}') = k_1(\mathbf{x}, \mathbf{x}')k_2(\mathbf{x}, \mathbf{x}') \quad (6.18)$$

$$k(\mathbf{x}, \mathbf{x}') = k_3(\phi(\mathbf{x}), \phi(\mathbf{x}')) \quad (6.19)$$

$$k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top \mathbf{A} \mathbf{x}' \quad (6.20)$$

$$k(\mathbf{x}, \mathbf{x}') = k_a(\mathbf{x}_a, \mathbf{x}'_a) + k_b(\mathbf{x}_b, \mathbf{x}'_b) \quad (6.21)$$

$$k(\mathbf{x}, \mathbf{x}') = k_a(\mathbf{x}_a, \mathbf{x}'_a)k_b(\mathbf{x}_b, \mathbf{x}'_b) \quad (6.22)$$

where  $c > 0$  is a constant,  $f(\cdot)$  is any function,  $q(\cdot)$  is a polynomial with nonnegative coefficients,  $\phi(\mathbf{x})$  is a function from  $\mathbf{x}$  to  $\mathbb{R}^M$ ,  $k_3(\cdot, \cdot)$  is a valid kernel in  $\mathbb{R}^M$ ,  $\mathbf{A}$  is a symmetric positive semidefinite matrix,  $\mathbf{x}_a$  and  $\mathbf{x}_b$  are variables (not necessarily in  $\mathbb{R}^M$ ),  $k_a(\cdot, \cdot)$  and  $k_b(\cdot, \cdot)$  are valid kernels in  $\mathbb{R}^{M_a}$  and  $\mathbb{R}^{M_b}$  respectively.

necessarily disjoint) with  $\mathbf{x} = (\mathbf{x}_a, \mathbf{x}_b)$ , and  $k_a$  and  $k_b$  are valid kernel functions over their respective spaces.

uso OG) NERVOLE PIÙ USATE SONO QUADRI GAUSSIANI:

$$k(\bar{x}, \bar{z}) = \exp\left\{-\frac{\|\bar{x} - \bar{z}\|}{2\sigma^2}\right\}$$

USANDO LE FORMULE DELLA TABELLA OTTIENIAMO:

$$h(\bar{x}, \bar{z}) = \underbrace{\exp(-\bar{x}^\top \bar{x}/2\sigma^2)}_u \underbrace{\exp(-\bar{x}^\top \bar{z}/2\sigma^2)}_u \underbrace{\exp(-\bar{z}^\top \bar{z}/2\sigma^2)}_u$$

IL PUNTO CORRISPONDE ALLA EMBEDDING-CHE MAPPI UNO SPazio DELL'FEATURE A DIMENSIONE INFINTA.

PLANO IN DUAL FORMA DIVERSA:

$$\max_{\mathbf{a}} \left\{ \sum_{n=1}^N a_n - \frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N a_n a_m y_n y_m k(\mathbf{x}_n, \mathbf{x}_m) \right\}$$

subject to  $0 \leq a_n \leq C$ , for  $n = 1, \dots, N$

$$\sum_{n=1}^N a_n y_n = 0$$

E' ORA DI VEDERCI IL CLASSIFICATORE:

$$f(x) = \sum_{n=1}^N \alpha_n y_n K(\bar{x}, \bar{x}_n) + b$$

$$= \sum_{\text{ZS}SV} \alpha_n y_n K(\bar{x}, \bar{x}) + b$$

→ i nuovi campioni saranno avvicinamente confrontati con support vector e non più con tutto il dataset

quindi il kernel trick è un altro modo di APPROCCIO A PROBLEMI NON LINEARMENTE SEPARABILI ed è il modo principale per INCREMENRTE LA COMPLICATITÀ del modello senza modificare la formulazione

METODI NON PARAMETRICI

COSÌ SI PUÒ MISURARE LE PRESTAZIONI DEI CLASSIFICATORI?

NEL CASO DI PROBLEMA DI CLASSIFICAZIONE BILOVATO SI UTILIZZA IL PARAMETRO DI ACCURATEZZA DEL CLASSIFICATORE:

$$\text{ACCURACY} = \frac{\# \text{ campioni di test correttamente classificati}}{\# \text{ campioni di test}}$$

NEL CASO INVECE DI PROBLEMI BILOVATI CONVIENE UTILIZZARE LA MATRICE DI CONFUSIONE:

- **TRUE POSITIVES**: PRESENTA CORRETTSIA SULLA CLASSE TRUE
- **TRUE NEGATIVES**: PRESENTA CORRETTSIA SULLA CLASSE FALSE
- **FALSI POSITIVI**: PREDICTO TRUE MA IL CAMPIONE NON APPARTIENE ALLA CLASSE
- **FALSI NEGATIVI**: PREDICTO FALSE MA IL CAMPIONE APPARTIENE ALLA CLASSE

Si definiscono quindi le seguenti metriche

$$\text{PRECISION} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

↗ INDICA LA PROPORTIONE DI ISANZE PREDICTE POSITIVE CHE EFFETTIVAMENTE LO SONO. MASSIMIZZARLE CA PRECISIONE SIGNIFICA RIDURRE AL MINIMO I FALSI POSITIVI  
(CLASSIFICARE NEGLI COME POSITIVI)

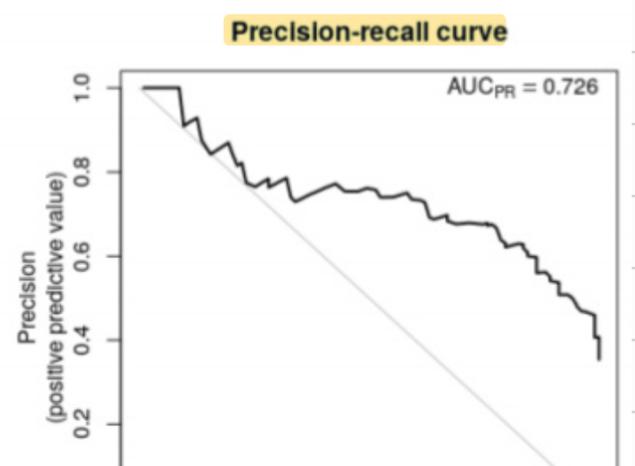
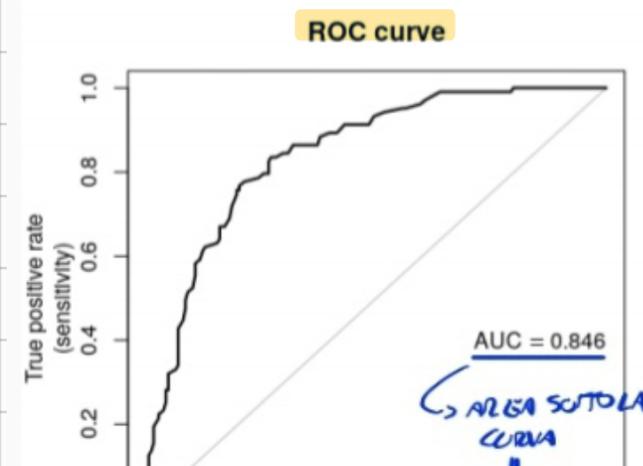
$$\text{RECALL} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

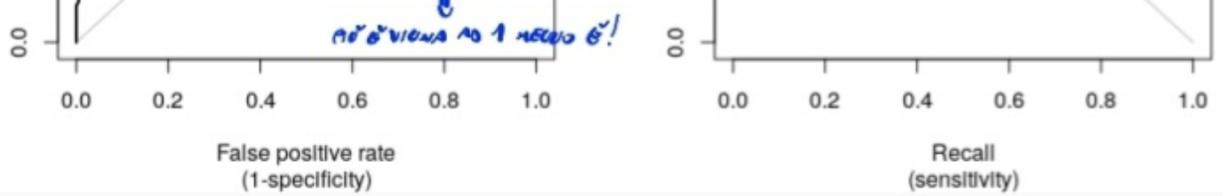
↗ INDICA PROPORTIONE DI ISANZE POSITIVE INQUA CHE SONO STATE PREDICTE CORRETTAMENTE. MASSIMIZZARLE CA RECALL SIGNIFICA RIDURRE AL MINIMO I FALSI NEGATIVI  
(CLASSIFICARE POSITIVI COME NEGLI)

$$\text{F1}(c) = \frac{\text{Precision}(c) * \text{Recall}(c)}{\text{Precision}(c) + \text{Recall}(c)}$$

Si visualizzano quindi due tipi di curve:





STIMAZIONE DI DENSITÀ: 15000MMI

FINO AD ORA ABBIAMO VISTO SOLTANTO MODELLI PROBABILISTICI CHE SONO  
BASATI SU MODELLI PARAMETRICO, DOVÈ SI USI A STIMARE UN PICCOLO  
NUMERO DI PARAMETRI (AD ESEMPIO MEDIA/COVARIANZA DI UNA GAUSSIANA).

PURETTA QUESTE PRIOR ASSUMPTIONS SONO TUTTAVIA CUE UN'ESTRAZIONE DI QUESTI APPROCCIO, SOLO QUANDO QUESTE ASSUNZIONI POSSANO ESSERE OBTENUTE

DEFINISCIAMO QUINDI UN 15000MMI COME UNA PARZIONE DEL DOMINIO IN BINI  
DI LARGHEZZA FISSATA. POI ANDIAMO A CONTARE IL NUMERO DI OSSERVAZIONI CHE  
RICADONO IN OGNI BIN.

QUESTO È QUINDI UNA RAPPRESENTAZIONE DI UNA PROBABILITÀ DI DENSITÀ MIGLIORATA DIVIDENDO  
IL NUMERO TOTALE DI OSSERVAZIONI  $N$  IN LARGHEZZA DEI BINI:

$$P_i = \frac{m_i}{N\Delta_i}$$

(IMPORTE TOTALI)

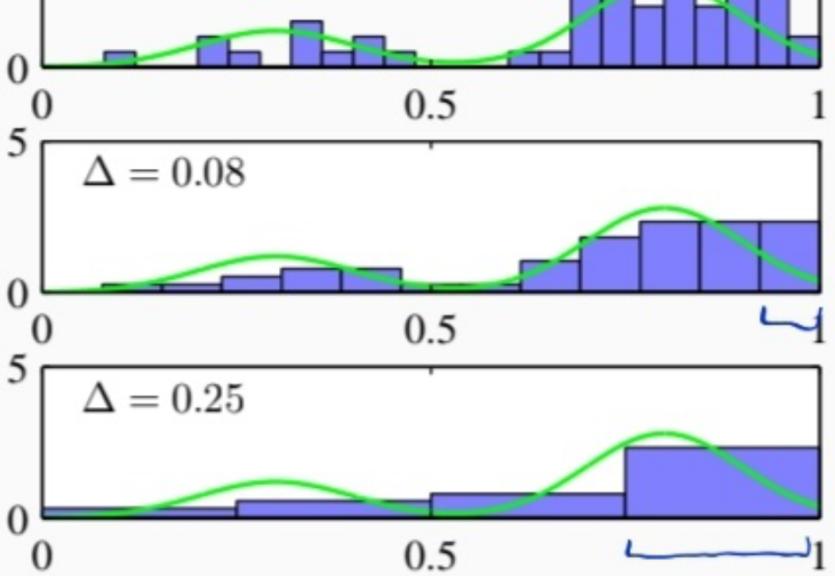
→ Campioni in quel bin

Larghezza bin

PURETTA CI POSSI AD UNA DENSITÀ  $p(x)$  CHE È COSTANTE UNTO TUTTA LA LARGHEZZA DI OGNI BIN.

ABBIAMO QUINDI OTTENUTO UNO STIMATORE DI DENSITÀ CON UN SIMILE IPERPARABOLICO





NOTIAMO CHE LA SELEZIONE DELLA LARGHEZZA  $\Delta$  È IMPORTANTE PERCHÉ:

- SE È TROPPO PICCOLA ALLORA LO STIMA CON UNA VARIANZA <sup>VARIANZA</sup>
- SE È TROPPO GRANDE ALLORA LI CORRIS APPROSSIMA MOLTO <sup>BIAS</sup>

INSERIRE QUESTO MODELLO È DISCONTINO NEI CONFINI DEI BINS. IL VANTAGGIO È CHE IL DATASET NON È RICHIESTO DURANTE IL TEST MA LO SI PREGA DI CUI CON ISOGRAFICO SCHEMA MOLTI CON LE DIMENSIONI

## HERTZEL DENSITY ESTIMATORS

L'OGGETTO È ESTENDERE L'IDEA FORNITA NELL'ISOGRAFO AD UNO STIMA CON PURAMENTE LOCALE; PER FARLO PUÒSSO SI CONSIDERA LA DISTRIBUZIONE BINOMIALE, LA QUALE RAPPRESENTA LA PROBABILITÀ DI OTENERE  $m$  SUCCESSI IN UNA SEQUENZA DI  $N$  PROVE DI BERNOULLI.

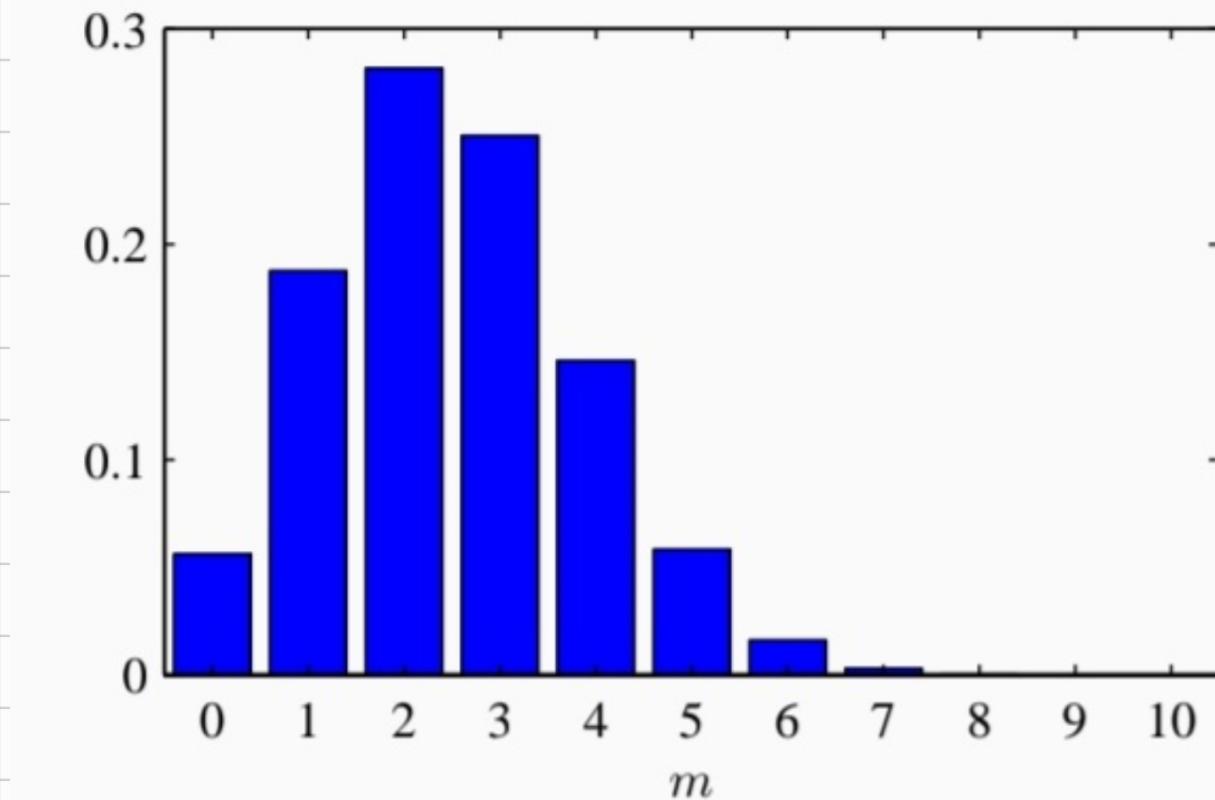
→ PROBABILITÀ DI SUCCESSO

$$B_{sm}(m|N, \mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m}$$

$$E(m) = \sum_{m=0}^N B_{m,n}(m|N, \mu)$$

=  $N\mu$

Le parole presenti in un documento cotengono approssimativamente una gaussiana.



ASSUMIAMO DI AVERE OSSERVAZIONI MISURATE DI UNA CERTA DENSITÀ DI PROBABILITÀ SCONosciuta  $p(\bar{x})$  IN UN SPazio A DIMENSIONE  $D$ .  
IN MODO SIMILE A UN ISTATISTICO, SI CONSIDERA UNA PICCOLA REGIONE  $\mathbb{R}$  ATORE ADO  $\bar{x}$ ; LA MASSA DI PROBABILITÀ ASSOCIAA CON  $\mathbb{R}$  VALE QUINDI:

$$P = \int_{\mathbb{R}} p(\bar{x}) d\bar{x}$$

$\rightarrow$  DENSITÀ DI PROBABILITÀ

SE ABBIANO  $N$  OSSERVAZIONI OTTIMESE DI  $p(x)$ , CHIUSO CON UNA PROBABILITÀ  $P$  DECIDERE NELLO SPAZIO  $M$ .

POSSIAMO POSSERCI DETERMINARE IL NUMERO TOTALE DI PUNTI  $K$  CHE APPARISCONO IN  $M$  TRAMITE LA BINOMIALE:

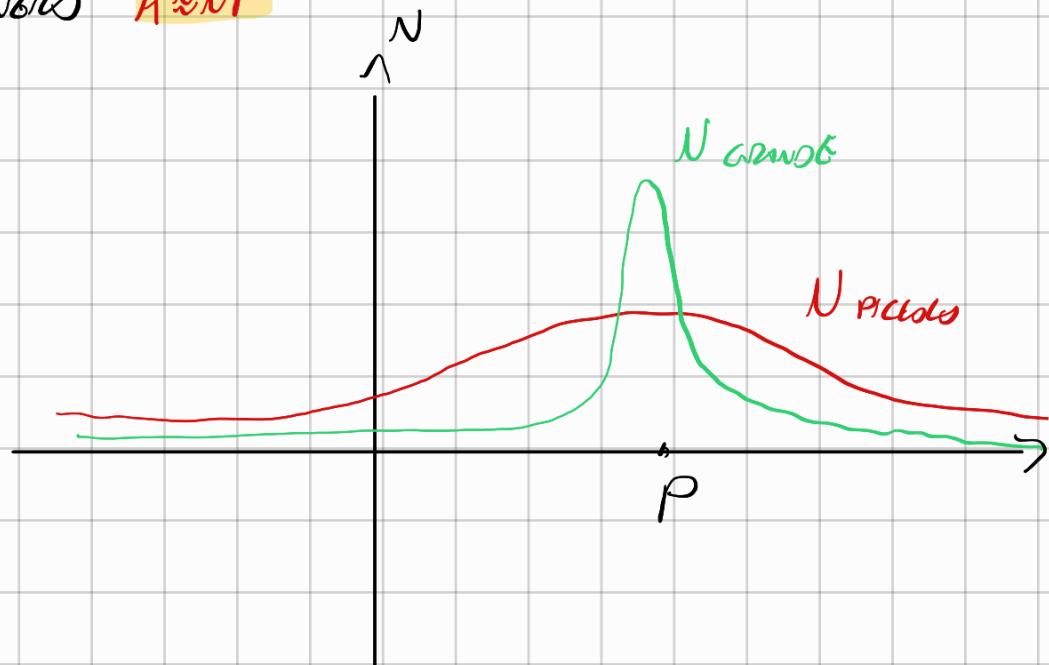
$$B_{1,n}(k|N, p) = \binom{N}{k} p^k (1-p)^{N-k}$$

CHE AVRA' LE SEGUENTI STATISTICHE:

$$\mathbb{E}(K/N) = p$$

$$\text{Var}(K/N) = \frac{p(1-p)}{N}$$

QUANDO PER  $N$  DIVENTA SOLO UNO SI HA CHE QUESTA DISTRIBUZIONE SARÀ CONCENTRATA AROUND  $p$  CHIUSO  $H \times NP$



MENTO INVECE SE ASSUMIAMO CHE  $M$ , CON UN CERTO VOLUTO VABBASTANZA PICCOLO, POSSIAMO DIRE CHE  $p(x)$  È COSTANTE, QUINDI  $P \propto p(x)V$ :

$$p(\bar{x}) \approx \frac{K}{NV}$$

DENSITÀ DI PROBABILITÀ È COSTANTE SE SI ASSUNGONO  
N punti in un piccolo volume V

TUTT'UNA QUESTA STIMA DI  $p(\bar{x})$  È BASATA SU DUE ASSUNZIONI CONTRADDITTORIE:

- N DEVE ESSERE ABBASTanza PICCOLO IN MODO CHE  $p(\bar{x})$  VI SIA COSTANTE  
PODOLANTE
- N DEVE ESSERE ABBASTanza GRANDE IN MODO CHE IL NUMERO DI PUNTI NELLO SPazio SIA RAPPRESENTABILE CON UNA BINOMIALE  
BINOMIALE

A QUESTO PUNTO SI HANNO DUE SCHEMI POSSIBILI:

- ① FISSARE K E USARNE CHE LE INFORMAZIONI IN INGRESSO DETERMININO COME DEBBA ESSERE V → MENOS - NEIGHBOURS
- ② FISSARE V E USARNE CHE LE INFORMAZIONI IN INGRESSO DETERMININO COME DEBBA ESSERE K → KERNEL-DENSITY ESTIMATION

STIMA KERNEL DENSITY

CONSIDERIAMO N come un PERCEPBO UNITARIO CENTRATO NELL'ORIGIN. POSSIAMO DESCRIVERLO TRAMITE UNA FUNZIONE KERNEL:

$$y_i(\bar{x}) = \begin{cases} 1 & \text{se } |y_i| \leq \frac{1}{2} \\ 0 & \text{per } i=1, \dots, D \end{cases}$$

$\left. \begin{array}{c} n(u) \\ \vdots \end{array} \right\} \circ \text{NURSMENT}$

QUINDI, PER UN QUALSiasi PUNTO  $\bar{x}$  SI HA CHE:

$$K\left(\frac{(\bar{x} - \bar{x}_m)}{h}\right) = 1$$

SE IL PUNTO  $\bar{x}_m$  SI TROVA ALL'INTERNO DELL'IPERCUOLO DI DIMENSIONE  $h$  E CENTRATO SU  $\bar{x}$

QUINDI IL MUNERO TOTALE DI PUNTI K CHE SI TROVANO ALL'INTERNO DELL'IPERCUOLO ATTORNO AD  $\bar{x}$  È PAR. A:

$$K_h = \sum_{m=1}^N K\left(\frac{\bar{x} - \bar{x}_m}{h}\right)$$

TORNANDO ALL'EQUAZIONE RELATIVA AL PROBABILITÀ:

$$p(\bar{x}) = \frac{K}{NV} = \frac{1}{N} \sum_{m=1}^N \frac{1}{h^d} K\left(\frac{\bar{x} - \bar{x}_m}{h}\right)$$

IL QUALE SOFFRE DALLE DISCONTINUITÀ NELLE I CONFINI DELL'IPERCUOLO, SI POSSA QUINDI ALL'ESTERNO GUARIRE DEL KERNEL:

$$p(\bar{x}) = \frac{1}{N} \sum_{m=1}^N \frac{1}{(2\pi h^2)^{d/2}} \exp\left\{-\frac{\|\bar{x} - \bar{x}_m\|^2}{2h^2}\right\}$$

CHE VENGONO CHIAMATI **KERNEL (O PARZIEN) DENSITY ESTIMATOR** DOVE  
 L'IPERPARMETRO  $h$  È CHIAMATO **KERNEL BANDWIDTH** CHE CONTROLLA IL "SMOOTHEING"  
 CHE VENGONO APPLICATO.

MENTRE LE FUNZIONI KERNEL DEBONO SODDISFARE:

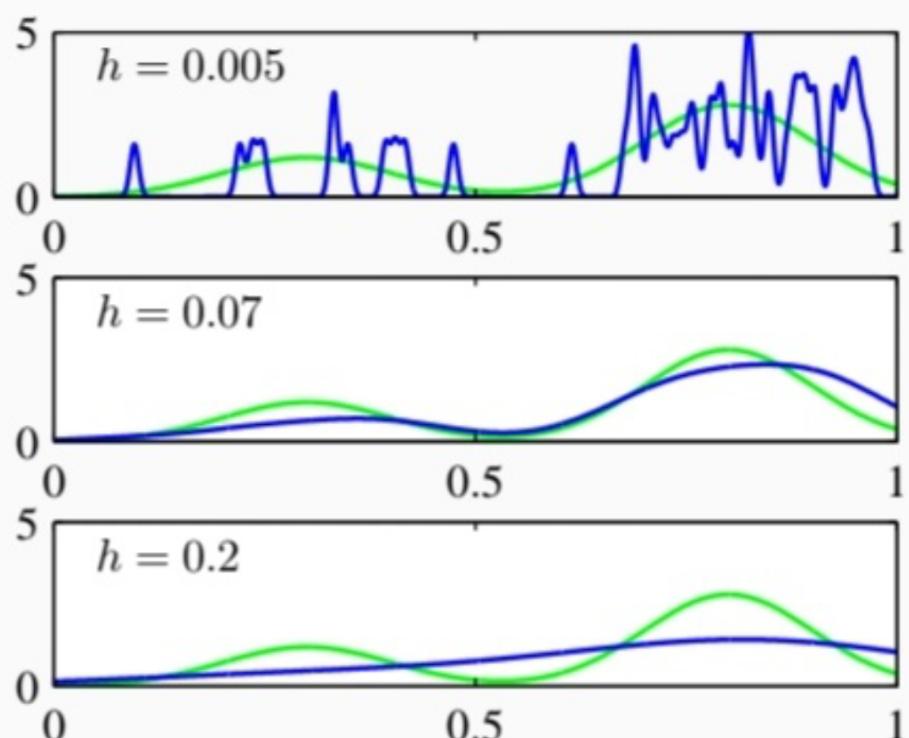
$$\cdot H(\bar{u}) \geq 0$$

$$\cdot \int \bar{u} H(\bar{u}) d\bar{u} = 0$$

$$\cdot \int H(\bar{u}) d\bar{u} = 1$$

QUESTI METODI SONO ANCHE CHIAMATI **METHOD LOCAL** POICHÉ VENGONO A STUARLE PIZZI

GUARDANDO AI DATI DI INPUT CHE SONO VISIONI AD  $\bar{x}$



**NADARAYA-WATSON KERNEL REGRESSION**

E POSSIBILE APPLICARE QUESTO TIPO DI METODO LOCALI A PROBLEMI DI

## REGRESSIONE.

SUPPONIAMO CHE UN **PROFILONE OTTIMALE** PER LA REGRESSIONE È DATO DA:

$$E(t|\bar{x}) = \int t p(t|\bar{x}) dt = \int t \frac{p(\bar{x}, t)}{p(\bar{x})} dt$$

→ SI RISOLVE LA CONDIZIONE CON LA CONDIZIONE NORMALIZZATA

POSSIAMO QUINDI USARLO COME STIMA DI DENSITÀ COME APPROSSIMAZIONE.

$$\hat{p}(\bar{x}, t) = \frac{1}{N} \sum_{n=1}^N K_h(\bar{x} - \bar{x}_n) K_h(t - t_n)$$

$$\hat{p}(\bar{x}) = \frac{1}{N} \sum_{n=1}^N K_h(\bar{x} - \bar{x}_n)$$

LE QUINDI L'APPROSSIMAZIONE DELLO STIMATORE OTTIMALE DIVENTA:

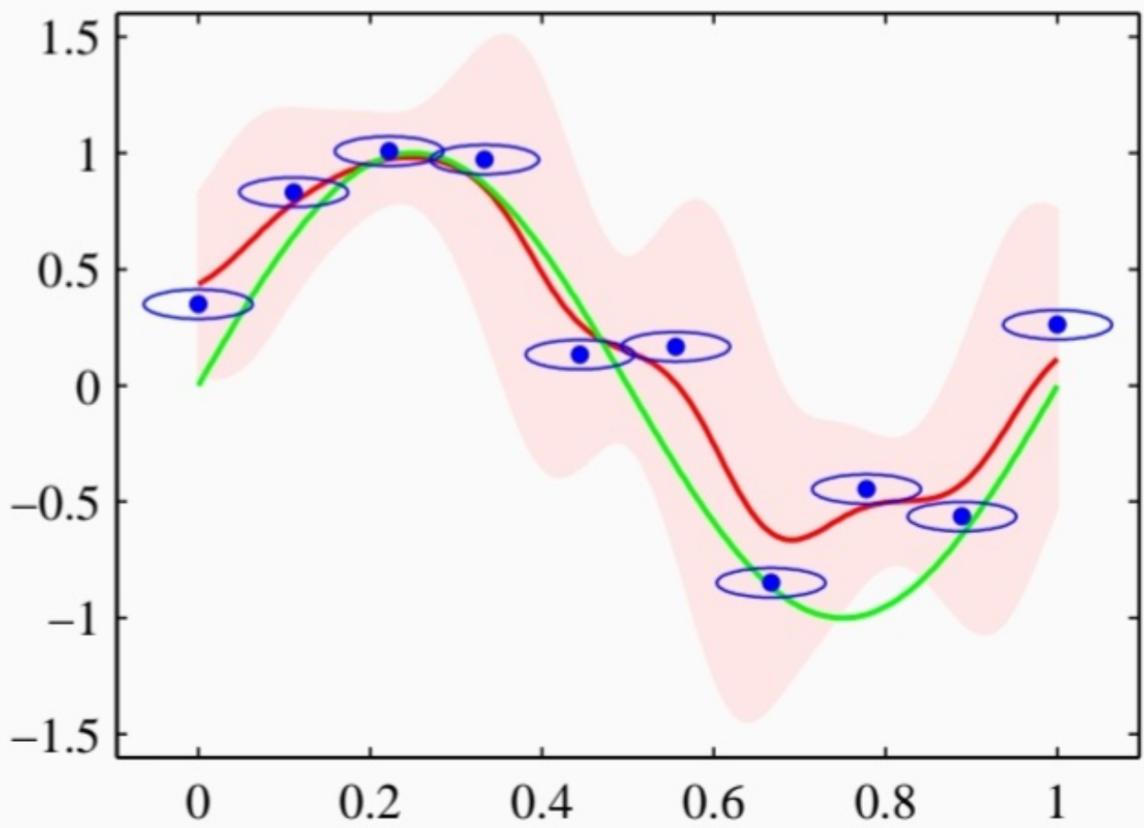
$$E(t|\bar{x}) = \int t \frac{p(\bar{x}, t)}{p(\bar{x})} dt$$

$$= \int t \frac{\frac{1}{N} \sum_{n=1}^N K_h(\bar{x} - \bar{x}_n) K_h(t - t_n)}{\frac{1}{N} \sum_{n=1}^N K_h(\bar{x} - \bar{x}_n)} dt$$

POICHE SI INTEGRA  
SULLO SPAZIO DI OUTPUT  
E NON QUELLO DI INPUT

$$= \frac{\sum_{n=1}^N K_h(\bar{x} - \bar{x}_n) \int t K_h(t - t_n) dt}{\sum_{n=1}^N K_h(\bar{x} - \bar{x}_n)}$$

$$= \frac{\sum_{n=1}^N K_h(\bar{x} - \bar{x}_n) t_n}{\sum_{n=1}^N K_h(\bar{x} - \bar{x}_n)}$$



## H-NEAREST NEIGHBORS

UN PROBLEMA CON LA KERNEL DENSITY ESTIMATION È IL FATTO CHE LA BANDWIDTH  $h$  È FISSATA PER TUTTI I KERNELS. QUINDI SE ABBIANO PARTI DELLO SPAZIO CON MOLTI CAMPIONI QUESTO C'È PIÙ PROBABILE DI OVERSMOOTHING E PERDITA DI DETTAGLIO SE  $h$  È TROPPO GRANDE.  
SE ANDIAMO A RIDURRE  $h$  SI HA CHE IN REGIONI A BASSA DENSITÀ SI AVRAANO SISTMI È RUMOROSI.

QUINDI IL NUOVO APPROCCIO È QUELLO DI FISSARE  $h$ !

$h$  GRANDE  $\rightarrow$  OVERSMOOTHING  
 $h$  PICCOLO  $\Rightarrow$  SISTMI RUMOROSI

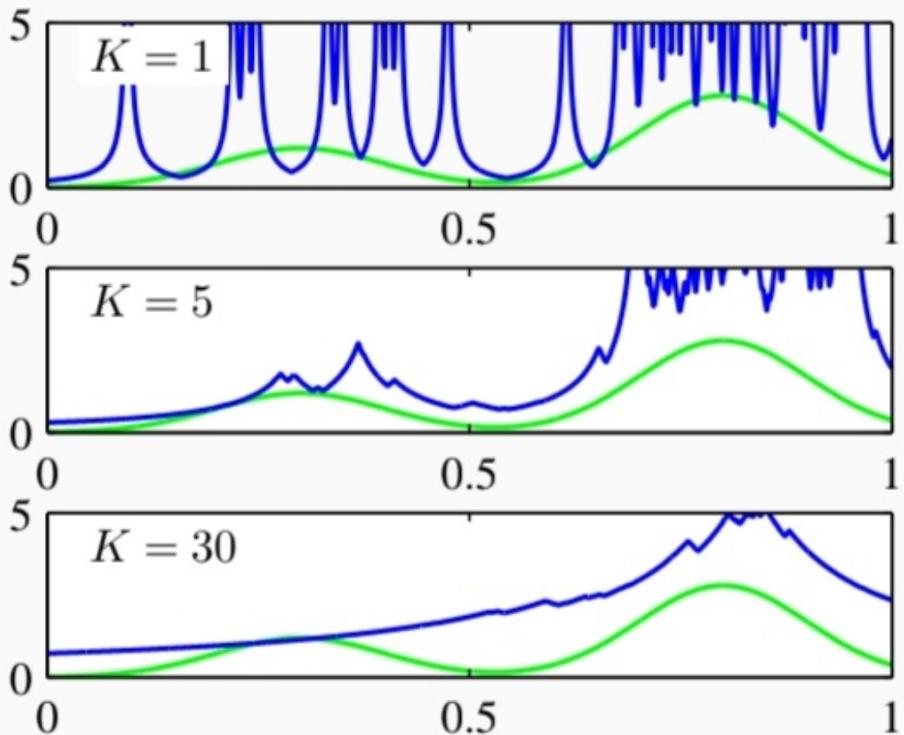
H-NEAREST NEIGHBORS È ALLORINCO:

① SI FISSA  $M$  E SI CONSIDERA UN'IPERSFERA A TUTTO AL PUNTO  $\bar{x}$  DOVE SI VUOLE STIMARE  $p(x)$

② SI INCREMENTA IL RAGGIO DELL'IPERSFERA FINO A QUANDO **ESPIAMENTO  $M$**   
**PUNTI DEL TRAINING SET SONO AL SUO INTERNO**

③ SI CALCOLA INFINE  $p(\bar{x}) = \frac{M}{MV}$  PER  $V$  ORA UALE AL VOLUME DELL'IPERSFERA  
RISULTANTE.

QUESTO APPROCCIO RISULTA **DISCONTINO E RUMOROSO PER IL SISTEMA DI DENSITÀ**



TUTTAVIA QUESTO METODO È MOLTO CONVENIENTE PER LA CLASSIFICAZIONE.  
POSSIAMO INFATTI CREARE UNA **HANNA SIMA DI DENSITÀ PER OGNI CLASSE**.  
**DENSITÀ CONDIZIONATE** E APPENA SUCCESSIVAMENTE USARE LE REGOLE DI BOYD'S

SUPPONIAMO DI avere  $N_k$  ESEMPI DI OGNI CLASSE  $k$  ( $\sum_k N_k = N$ ). PER  
CLASSIFICARE UN NUOVO PUNTO  $\bar{x}$  SI UTILIZZA IL RAGGIO DELL'IPERSFERA ASSUNENDO CHE  
IL VOLUME INTORNO AD  $\bar{x}$  NECESSARIO AD INCLUDERE  $N_k$  PUNTI SIA  $V_k$   
QUINDI SE NELL'IPERSFERA SI HANNO  $N_k$  PUNTI DELLA CLASSE  $k$ :

$$P(\bar{x}|C_H) = \frac{\mu_H}{N_H V_H}$$

E quindi siamo vicini a una classificazione con Bayes:

$$P(\bar{x}) = \frac{N}{N_H V_H}$$

$\hookrightarrow$  marginale

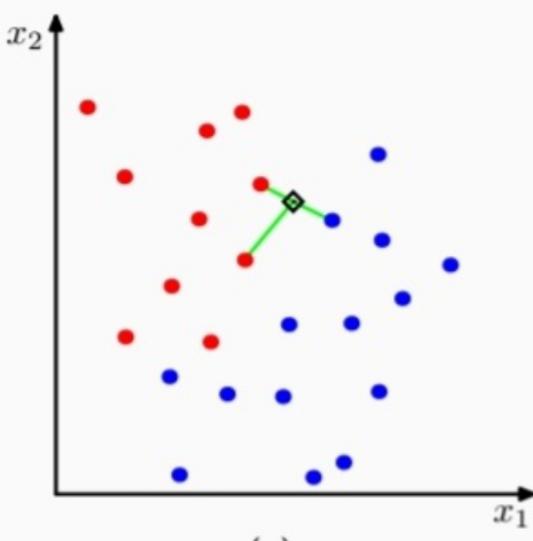
$$P(C_H) = \frac{N_H}{N}$$

$\hookrightarrow$  prior

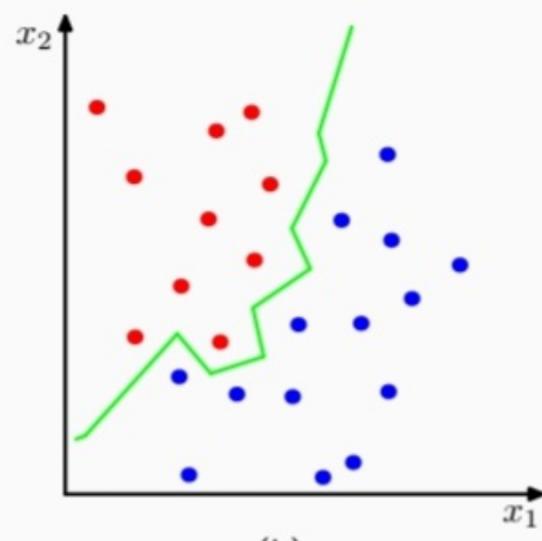
APPPLICAZIONE BAYES:

$$P(C_H | \bar{x}) = \frac{P(\bar{x}|C_H) P(C_H)}{P(\bar{x})} = \frac{\frac{1}{V_H}}{\frac{N}{N_H V_H}} = \frac{1}{N}$$

$\hookrightarrow$  per  $N$  piccolo  $\Rightarrow$  overfitting



(a)



(b)

PER CONCLUDERE:

- QUESTE TECNICHE HANNO UN PICCOLO NUMERO DI IPERPARAMETRI: LA **BROWNSIANA DEL NEURONE h** APRE IL NUMERO DI NEURONI **H**

- QUESTI METODI HANNO BLOCCO PROBLEMI DI CONVERGENZA
- PER FARLE PREDIRE QUESTI METODI HANNO BISOGNO CHE TUTTO IL TRAINING SIA DISPONIBILE AD INFERNAL TIME

