

Sistemi Intelligenti Supervised learning

Alberto Borghese
Università degli Studi di Milano
Laboratorio di Sistemi Intelligenti Applicati (AIS-Lab)
Dipartimento di Scienze dell'Informazione
Alberto.borghese@unimi.it



A.A. 2014-2015

1/57

<http://borghese.di.unimi.it/>



Riassunto



- **Supervised learning**
- Regressione multi-scala
- Classificazione

A.A. 2014-2015

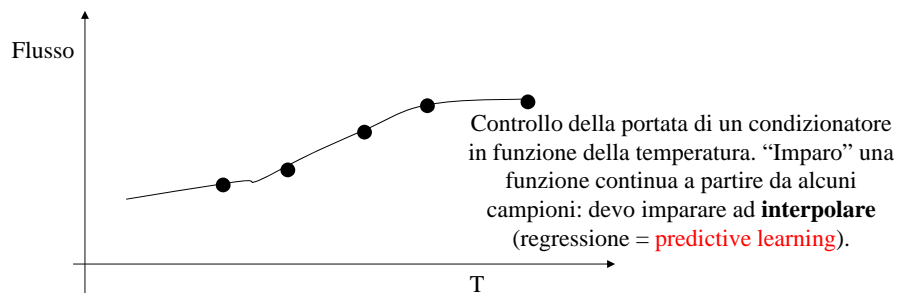
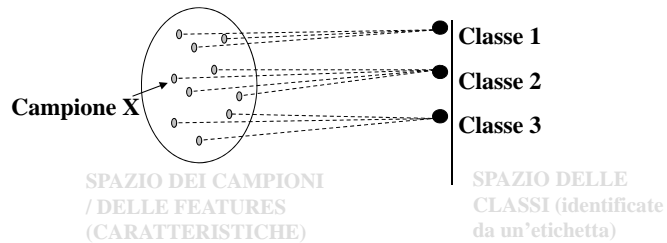
2/57

<http://borghese.di.unimi.it/>



Classificazione e regressione

Mappatura dello spazio dei campioni nello spazio delle classi.



A.A. 2014-2015

3/57

<http://borgnese.di.unimi.it/>





Ruolo dei modelli

- **Identificazione:** stimo i parametri di un modello a partire dai dati: identifico il modello.
- **Utilizzo:** utilizzo il modello per inferire informazioni su nuovi dati (controllo, regressione predittiva, classificazione).

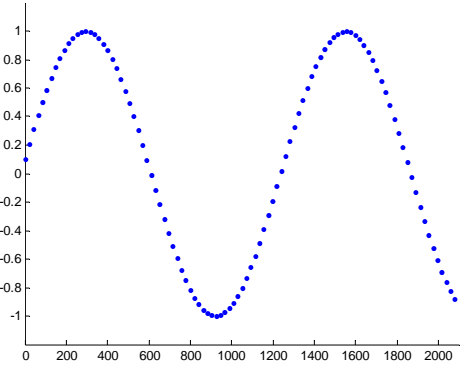
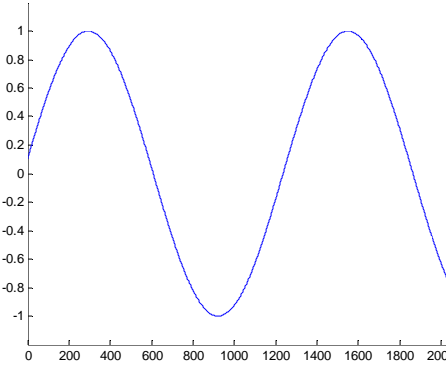
A.A. 2014-2015

4/57

<http://borgnese.di.unimi.it/>






Modello parametrico

I punti vengono fittati perfettamente da una sinusoida: $y = A \sin(\omega x + \phi)$. Devo determinare solo i 3 parametri della sinusoida (non lineare), i cui valori ottimali sono: $\omega = 1/200$, $\phi = 0.1$, $A = 1$. I parametri hanno un significato semantico.

A.A. 2014-2015
5/57
<http://borghese.di.unimi.it/>

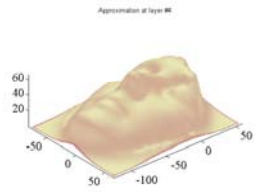
I modelli semi-parametrici

- L'approssimazione è ottenuta mediante funzioni “generiche”, dette di **base**, soluzione molto utilizzata nelle NN e in Machine learning. E' anche associato all' approccio «black-box» in cibernetica. Non si hanno informazioni sulla struttura dell'oggetto che vogliamo rappresentare.
- (Il concetto di Base in matematica è definito mediante certe proprietà di approssimazione che qui non consideriamo, consideriamo solo l'idea intuitiva). Il concetto di base è simile a quello dei “replicating kernels”.
- E' anche l'idea che sta alla base delle Reti Neurali Artificiali

$$z(p(x, y)) = \sum_i w_i G(p, p_i; \sigma)$$

Combinazione lineare di funzioni di base

Da calcolare

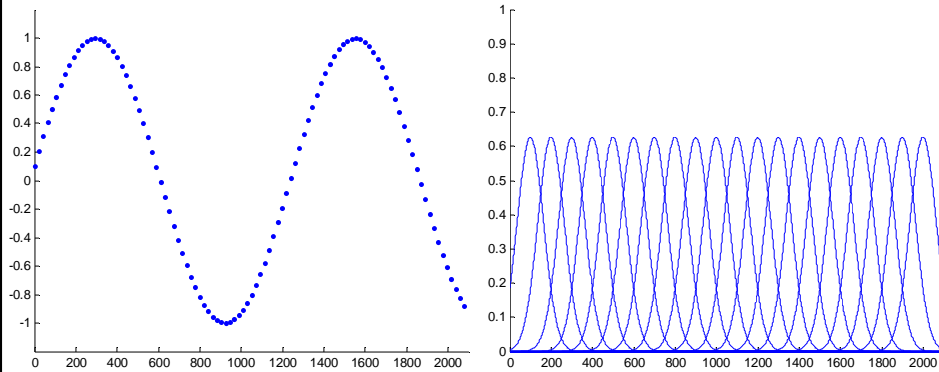


Funzione di base (fissate)

A.A. 2014-2015
6/57
<http://borghese.di.unimi.it/>



Approssimazione mediante un modello semi-parametrico (lineare)



Sinusoide $y = A \sin(\omega x + \phi)$ con $\omega = 1/200$, $\phi = 0.1$.

Vogliamo fittare i punti con l'insieme di Gaussiane riportate sulla dx. In questo caso hanno tutte $\sigma = 90$. Come le utilizzo?

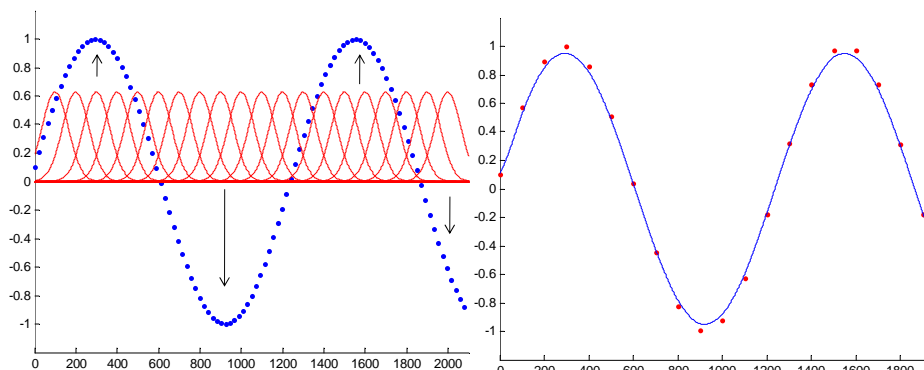
A.A. 2014-2015

7/57

<http://borghese.di.unimi.it/>



Funzionamento di un modello semi-parametrico (lineare)



$$y(x) = \sum_{i=1}^{20} w_i G(x - x_{o_i}; 90^\circ)$$

Devo definire, gli $M \{w_i\}$.
 $3 \ll M \ll N - \text{numero punti}$.

I σ sono tutti uguali ed uguali a 90° , le Gaussiane sono equispaziate.
 Le Gaussiane sono note tutte a priori, devono essere definiti i pesi.

A.A. 2014-2015

8/57

<http://borghese.di.unimi.it/>



Modelli lineari e non lineari



Classificazione alternativa dei modelli. Vengono utilizzate classi molto diversi di algoritmi per stimare i parametri di questi due tipi di modelli.

$$z(p(x, y)) = f(x) = \sum_i w_i x$$

$$z(p(x, y)) = \sum_i f_i(p; w)$$

$f(\cdot)$ è funzione lineare nei $\{w_i\}$

$f(\cdot)$ è funzione non lineare

e.g. $f(\cdot) = e^{w \cdot x}$
 $f(\cdot) = x \ln(w \cdot x)$
....



How to classify the error introduced by a model?



Is the model good enough?

Does it have enough parameters?

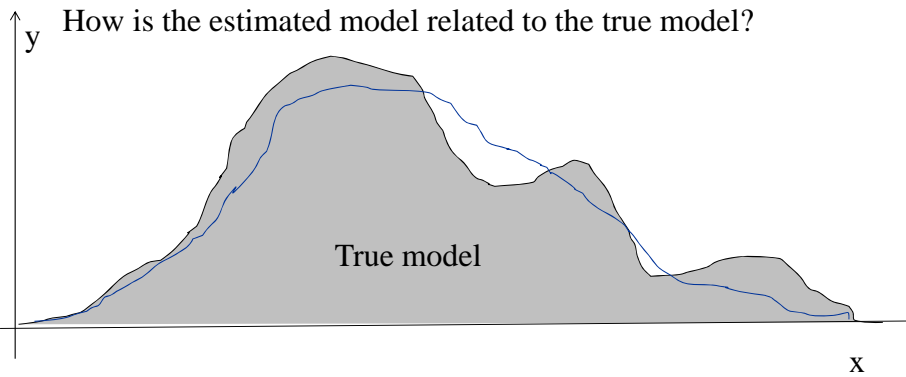
Does it cover the input domain (in all dimensions)?

This is not enough to obtain a good model!!

The model should be properly tuned to the data



How to classify the error introduced by a model?



Bias and variability trade-off

Bias is the distance of the model curve from the true unknown curve. It is associated to model error.

A.A. 2014-2015

11/57

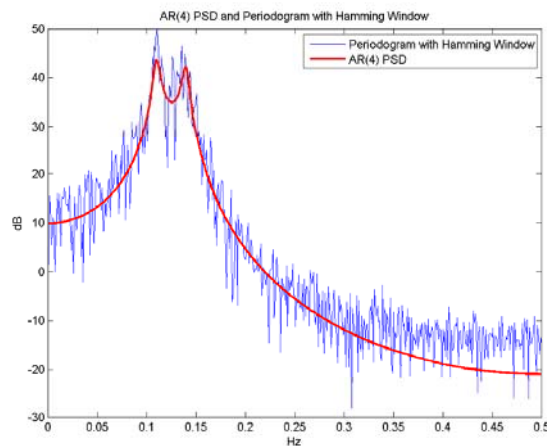
<http://borghese.di.unimi.it/>



Variability



How are the measured points related to the estimated model?



Given $P_{\text{mes}}(x_{\text{mes}}, y_{\text{mes}})$ and $y = f(x)$, the error is measured as: $\text{dist}(y_{\text{mes}}, f(x_{\text{mes}}))$, for instance Euclidean distance. It is associated to measurement error.

If variability goes to zero, bias increases and overfitting arises.

In a good model, variability tends to the statistics of the measurement noise.



Problemi nella procedura di apprendimento



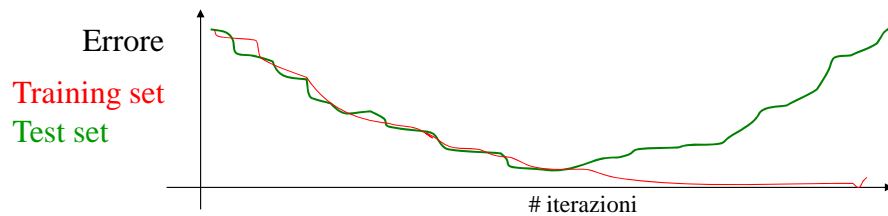
Quando si termina l'algoritmo di apprendimento?

Bootstrap – Vengono estratti pattern con ripetizioni.

Cross-Validation - Errore sull'insieme di training =
Errore sull'insieme di test.

Utilizzare lo “structural risk” invece dell’”empirical risk”.

Si vuole evitare che il modello si specializzi troppo sui pattern di training e non sia in grado di interpolare.



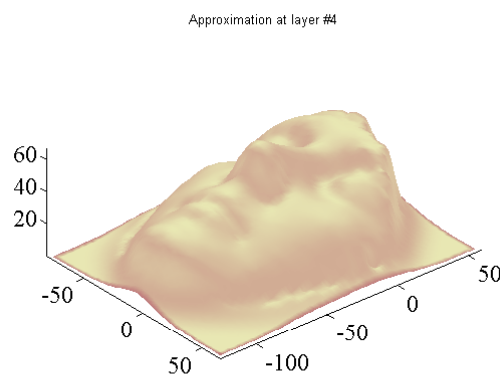
A.A. 2014-2015

13/57

<http://borgese.di.unimi.it/>



Problema dell'overfitting dovuto a sovrapparametrizzazione



Quante unità?

A.A. 2014-2015

14/57

<http://borgese.di.unimi.it/>



Riassunto



- Supervised learning
- **Regressione multi-scala**
- Classificazione

A.A. 2014-2015

15/57

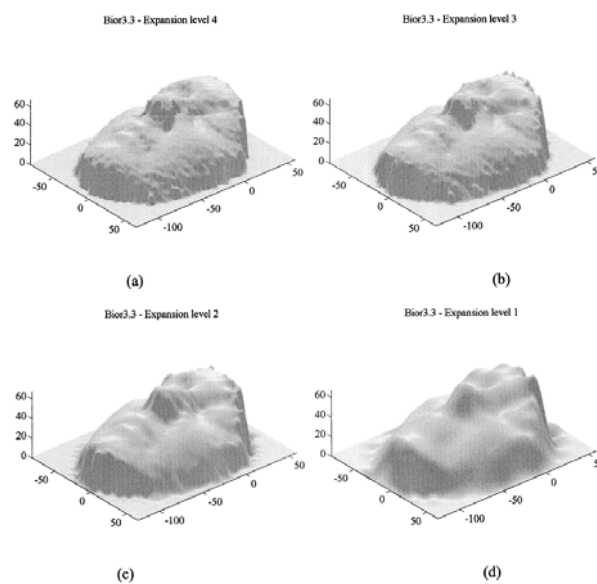
<http://borghese.di.unimi.it/>



Pyramidal reconstruction



- Which is the adequate scale?
- Which model is the closest to the true model?



A.A. 2014-2015



Surface reconstruction with filtering



- Convolution: $\hat{f}(x) = \int_{\mathbf{R}} f(c) G(x-c|\sigma) dc = f(x) * G(x; \sigma)$

we can reconstruct signals up to a certain scale, provided an adequate small value of σ .

- Discrete convolution: $\hat{f}(x) = f_i * G(x - x_{k_i}; \sigma) = \sum_{i=1}^N w_i G(x - x_{k_i}; \sigma)$

The reconstruction of the function, if $G(\cdot)$ is normalized, is obtained through digital filtering.

Extrapolation beyond the sample points. Reconstruction up to a given scale.



Filters and bases



$$\hat{f}(x) = \sum_k f_k * G(x - x_k; \sigma)$$

$$\hat{f}(x) = \sum_{k=1}^N f_k G(x, x_k, \sigma) \Delta x = \frac{\Delta x}{\sqrt{\pi} \sigma} \sum_{k=1}^N f_k e^{-\frac{(x-x_k)^2}{\sigma^2}} \quad \frac{\Delta x_k}{\sqrt{\pi} \sigma} \text{ Normalization factor}$$

Normalized Gaussians, filter = weighed sum of shifted (normalized) basis functions. Basis representation. Approximation space.

Riesz basis, the approximation space is characterized by the scale of the basis that determines the amplitude of the space.

A sequence of spaces can be defined according to σ :

$$\sigma_0 \rightarrow V_0; \sigma_1 \rightarrow V_1; \sigma_2 \rightarrow V_2, \dots$$

The number of representable functions increases.



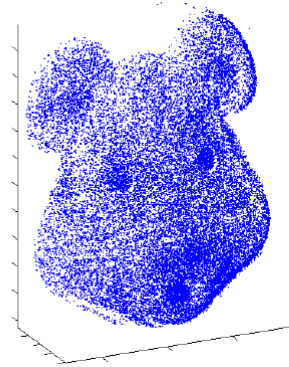
Advantages and problems



$$\hat{f}(x) = \sum_{k=1}^N f_k G(x, x_k, \sigma) \Delta x = \frac{\Delta x}{\sqrt{\pi} \sigma} \sum_{k=1}^N f_k e^{-\frac{(x-x_k)^2}{\sigma^2}}$$

Filters interpolates and reduces noise but...

Height in the function on a grid crossing should be known.



A.A. 2014-2015

19/57

<http://borgese.di.unimi.it/>



Gridding



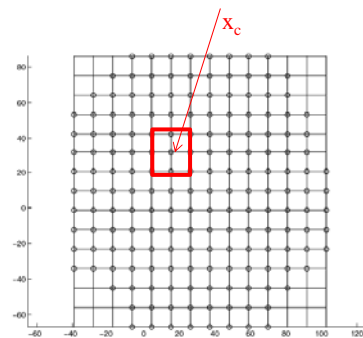
$$\hat{f}(x) = \sum_{k=1}^N f_k G(x, x_k, \sigma) \Delta x = \frac{\Delta x}{\sqrt{\pi} \sigma} \sum_{k=1}^N f_k e^{-\frac{(x-x_k)^2}{\sigma^2}}$$

How can we determine w_k from points clouds?

Local estimators. Nadaraya Watson estimator. *Lazy learning*.

$$\hat{f}(x_c) = \frac{\sum_i y_i K_\sigma(x_i, x_c)}{\sum_i K_\sigma(x_i, x_c)} = \frac{\sum_i y_i e^{-\frac{\|x_i - x_c\|^2}{\sigma^2}}}{\sum_i e^{-\frac{\|x_i - x_c\|^2}{\sigma^2}}}$$

$K_\sigma(\cdot)$ Gaussian



Parzen-window estimators.

A.A. 2014-2015

20/57

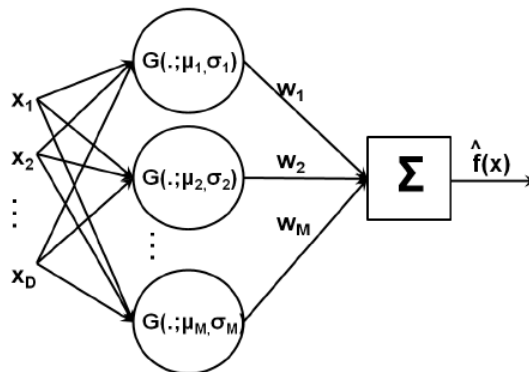
<http://borgese.di.unimi.it/>



RBF Network



Connessionism. Simple processing units combined with simple operations to create complex functions.



A.A. 2014-2015

21/57

Perceptron

<http://borghese.di.unimi.it/>

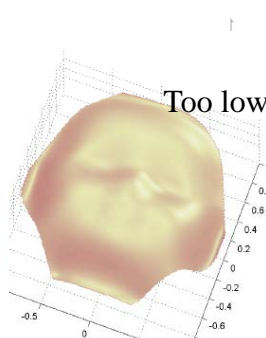
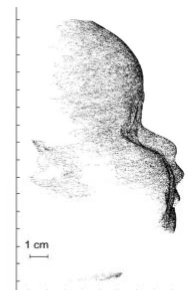
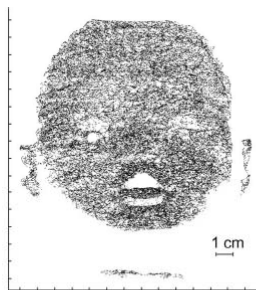


Surface Approximation



- Properties:
 - Redundancy.
 - Riesz basis (unique representation, given the height in the grid crossings).

Which scale?



<http://borghese.di.unimi.it/>



Incremental strategy



- Acquire more data in the more complex areas, less smooth, higher frequency.
- Acquire less data in the less complex areas, more smooth, lower frequency.

$$\hat{f}(x) = \sum_{k=1}^N f_k G(x, x_k, \sigma) \Delta x = \frac{\Delta x}{\sqrt{\pi} \sigma} \sum_{k=1}^N f_k e^{-\frac{(x-x_k)^2}{\sigma^2}}$$

- Can we use a single Δx ?

Incremental approximation with local adaptation.

A.A. 2014-2015

23/57

<http://borghese.di.unimi.it/>



Start from low resolution

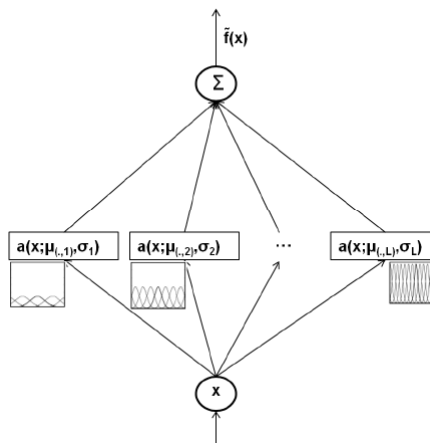


- Low resolution, small distance, $1/\Delta x > 2\nu_{\text{Max}}$

$$\hat{f}(x) = \sum_{k=1}^N f_k G(x, x_k, \sigma) \Delta x = \frac{\Delta x}{\sqrt{\pi} \sigma} \sum_{k=1}^N f_k e^{-\frac{(x-x_k)^2}{\sigma^2}}$$

σ determines the amount of overlap. It determines also the frequency content of the Gaussian $G(\cdot)$.

Once σ (or Δx is computed) the support is defined.





Determination of the surface height

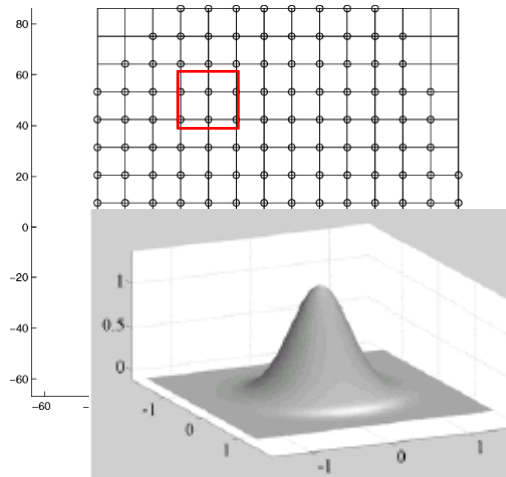


How many points to consider? The Gaussian has infinite support. Splines have a limited support.

$$\hat{f}(x) = \sum_{k=1}^N f_k G(x, x_k, \sigma) \Delta x$$

Apply local estimator to the data points in the neighbourhood of a grid crossing (Gaussian center) to compute f_k .

Sorting of the data is made simple, they are subdivided into quads. Identified the points inside the neighbourhood is equivalent to extract all the points between two positions in the data vector.



A.A. 2014-2015

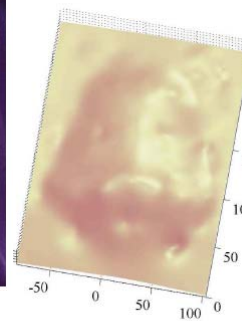
25/57



We can obtain a «poor» reconstruction



But it is a start. It can be seen as a modified support for successive approximations.



A.A. 2014-2015

26/57

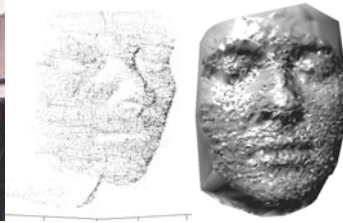
<http://borghese.di.unimi.it/>



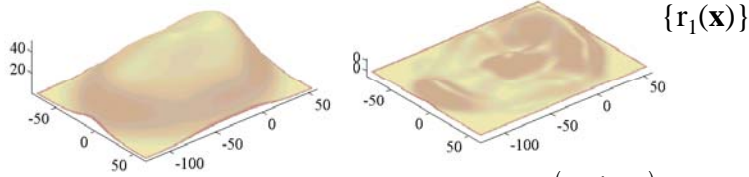
What can be done?



We can compute the residual for each data point.



Approximation at layer #1

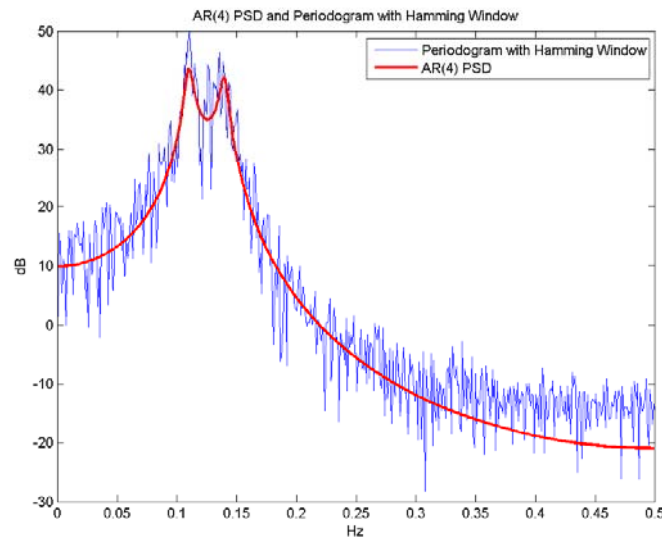


We evaluate the residual for each data point: $r_i = dist(y_m, \hat{f}(x_m))$

E.g.: $r_i = (y_m - \hat{f}(x_m))^2$ $r_i = |y_m - \hat{f}(x_m)|$



Where does this residual come from?



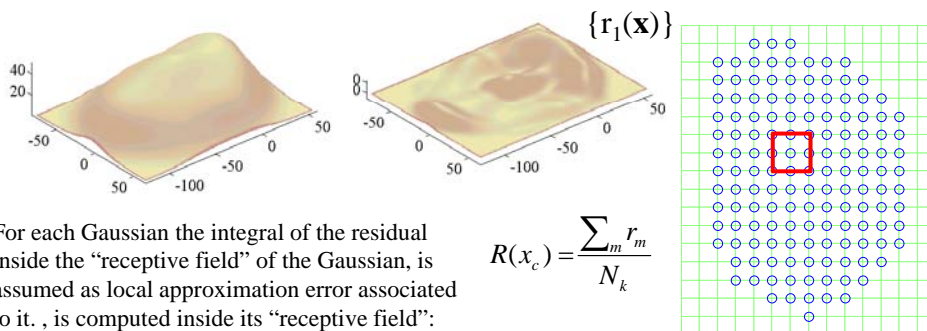
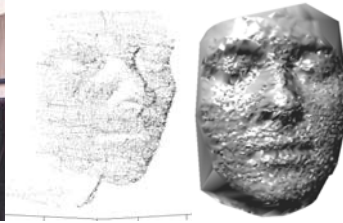
We want to eliminate variability due to noise.



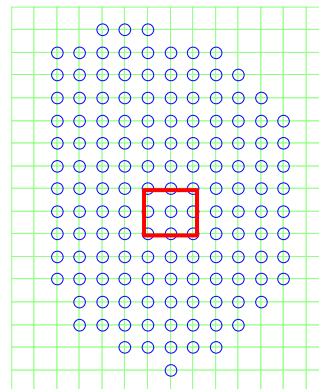
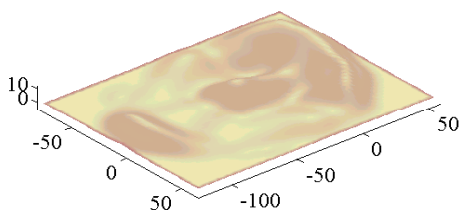
Is the residual adequate?



Approximation at layer #1




How can we evaluate the local adequacy of the reconstruction?




$$R(x_c) = \frac{\sum_m r_m}{N_k}$$

We compare the local residual it with a threshold:

- Degree of approximation
- Noise: RMS.

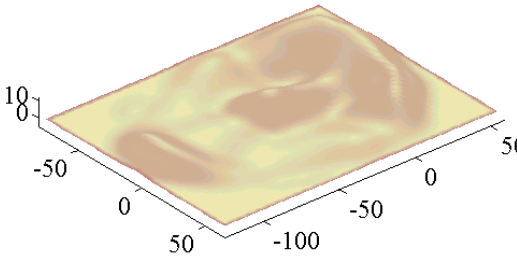


Layer 2

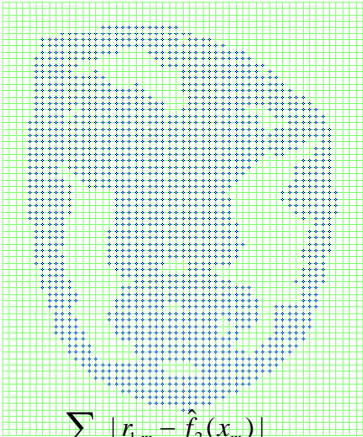


Input are the residuals, $r_{1,m} = |y_m - \hat{f}_1(x_m)|$
 Output is the model that approximates $r_{1,m}$: $f_2(x_m) \rightarrow r_{1,m}$

Output of layer #2



Layer #2




More packed Gaussians
 There should be enough points to have a reliable local estimate of not filled grid.

$$R(x_c) = \frac{\sum_m |r_{1,m} - \hat{f}_2(x_m)|}{N_k}$$


A.A. 2014-2015

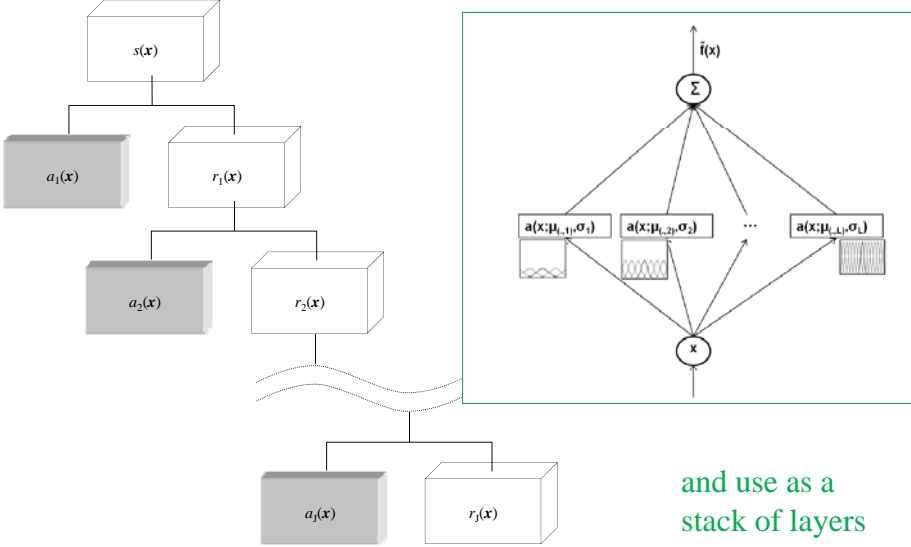
31/57

<http://borghese.di.unimi.it/>



Hierarchy construction





and use as a stack of layers

A.A. 2014-2015

32/57

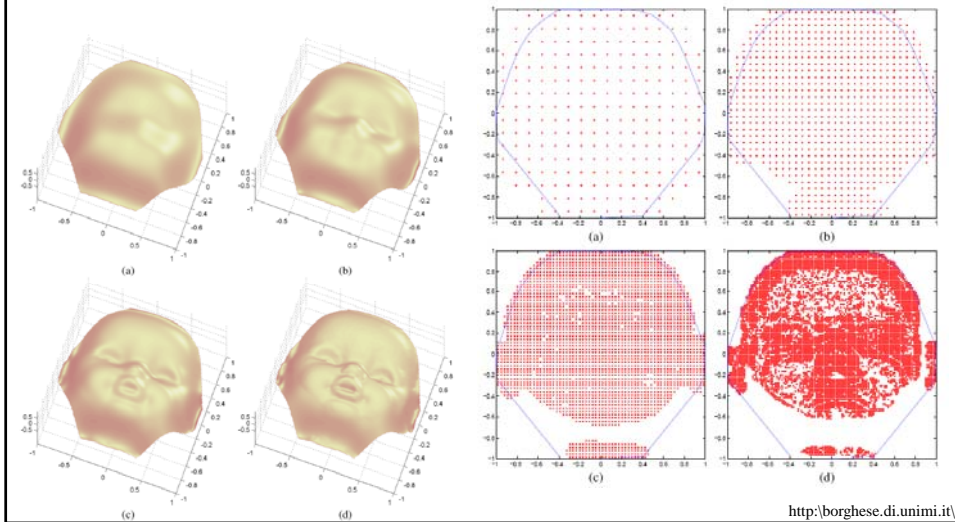
<http://borghese.di.unimi.it/>



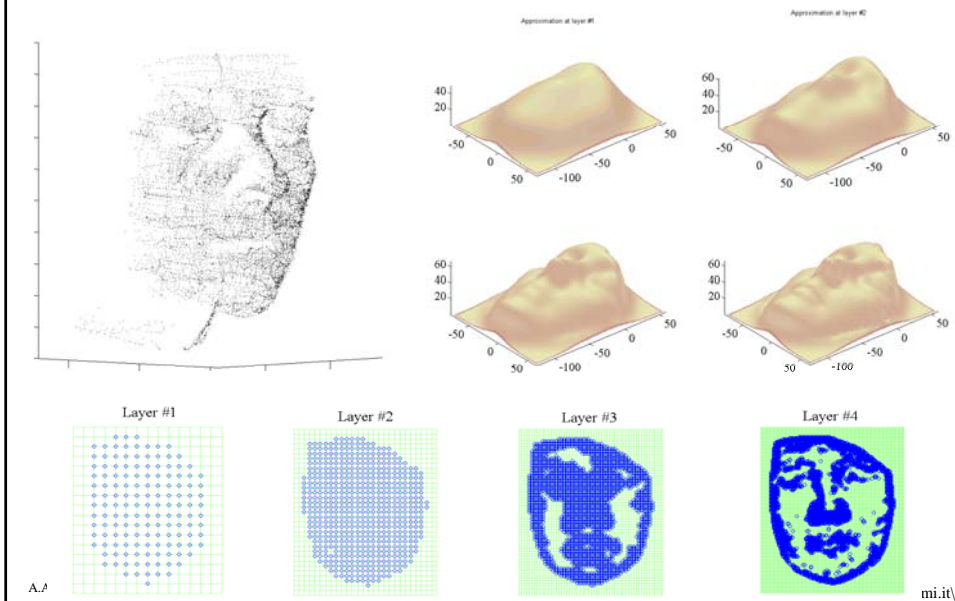
How to operate on large sets of data?



Recursive splitting of the quad domain -> local re-ordering of the data.



Applicazione della regressione





Characteristics of HRBF networks



- Not fully occupied layers
- Adaptive local scale
- Adaptive allocation of the resources
- Uniform convergence to a residual error
- Residual bias is recovered in the next layers.
- Relatively dense data sets are required to obtain a robust local estimate.
- Riesz basis, with a high degree of redundancy between the coefficients. The angle between two approximating spaces is not 90, but it is considerably smaller

$$\cos \alpha_j = \sup_{f(\cdot) \in V_j, h(\cdot) \in V_{j+1}} \frac{\langle f(\cdot), h(\cdot) \rangle}{\|f(\cdot)\|_2 \|h(\cdot)\|_2} = \cos \alpha_{j-1}$$

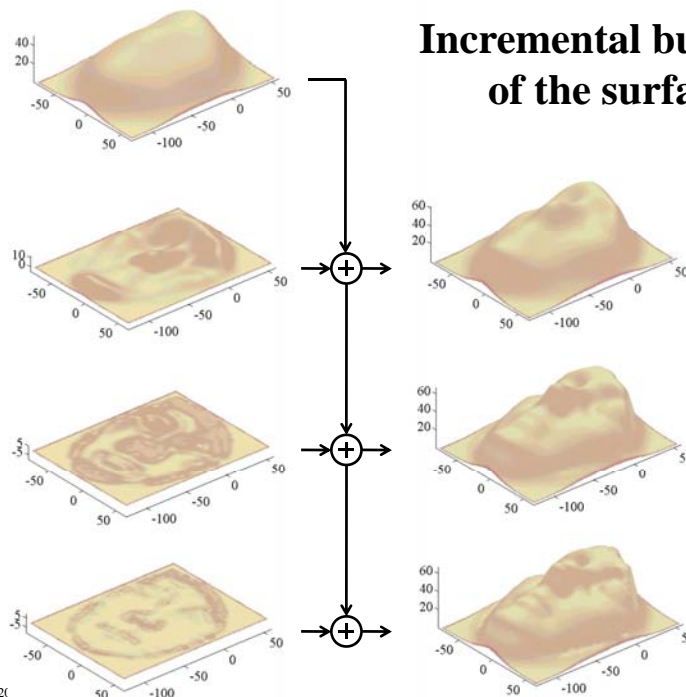
A.A. 2014-2015

35/57

<http://borghese.di.unimi.it/>



Incremental building of the surface



A.A. 2014-2015

<http://borghese.di.unimi.it/>



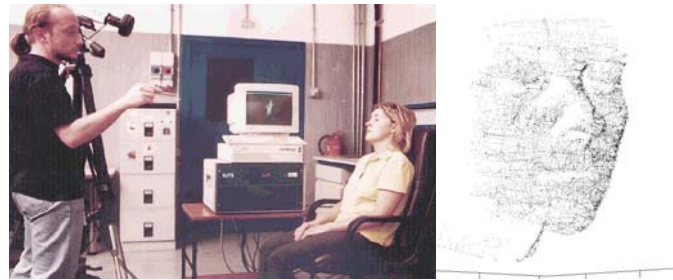
On-line version



- Data do not arrive all together (batch)
- One data at a time.
- Growing while scanning



hrbf_online.wmv



A.A. 2014-2015

37/57

<http://borghese.di.unimi.it/>



Observation



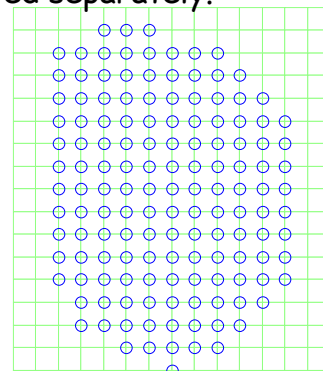
- Each new point, $y=f(x_k)$, modifies at least f_1 around x_k .
- This in turns can modify 4 values in the next layer and so forth.

Recomputation can be simplified:

Numerator and denominator are stored separately.

$$\hat{f}(x) = \frac{\sum_i y_i K_\sigma(x_i, x)}{\sum_i K_\sigma(x_i, x)} = \frac{\sum_i y_i e^{-\frac{\|x_i - x\|^2}{\sigma^2}}}{\sum_i e^{-\frac{\|x_i - x\|^2}{\sigma^2}}}$$

For each new point a new term is added and the ratio is recomputed.



A.A. 2014-2015

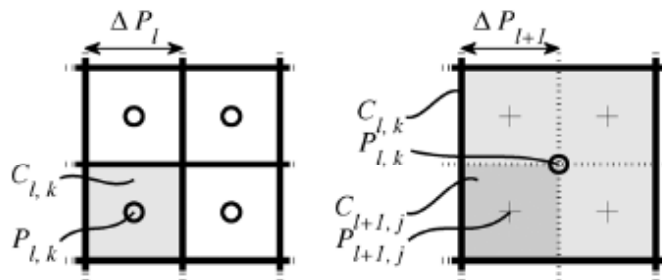
38/57



Local operations



- Local splitting of each quad is achieved when:
 - Residual is higher than threshold
 - Enough points have been sampled



A.A. 2014-2015

39/57

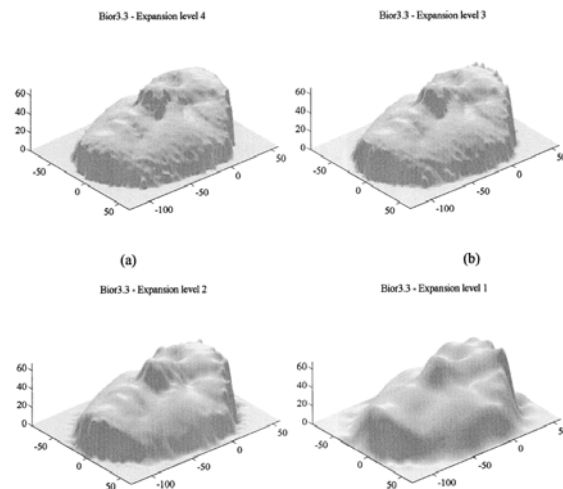
<http://borgese.di.unimi.it/>



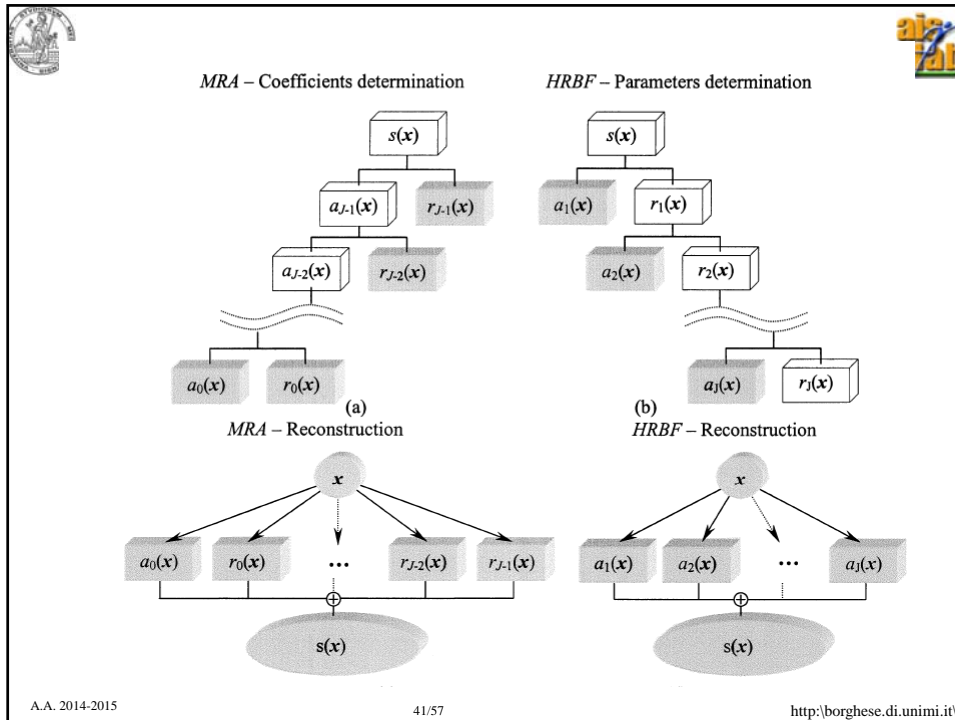
Comparison with Wavelets



- Fast incorporation of the content (high angles between approximating spaces \rightarrow 90 degrees)
- No control on the residual.



A.A. 2014-2015



Beyond Wavelet

Portilla et al., Image Denoising Using Scale Mixtures of Gaussians in the Wavelet Domain, 2003.

Coefficients reduction through a model of the noise.

RBF and Wavelet have excellent for CUDA implementation as all bases with limited support.

A.A. 2014-2015 42/57 http://borghese.di.unimi.it/



Riassunto



- Supervised learning
- Regressione multi-scala
- **Classificazione**

A.A. 2014-2015

43/57

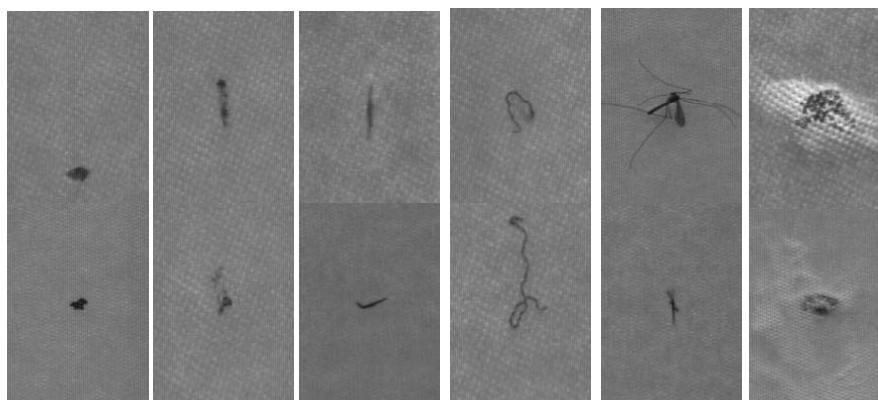
<http://borghese.di.unimi.it/>



CLASSIFICAZIONE: Riconoscimento difetti in linee di produzione



(progetto finanziato da Electronic Systems: 2006-2007)



regolari

irregolari

allungati

fili

insetti



macchie su
denso

Difetti – Classificazione real-time e apprendimento mediante **boosting**.
Committee (linear combination) of weak (binary) classifiers.

A.A. 2014-2015

44/57



<http://borghese.di.unimi.it/>

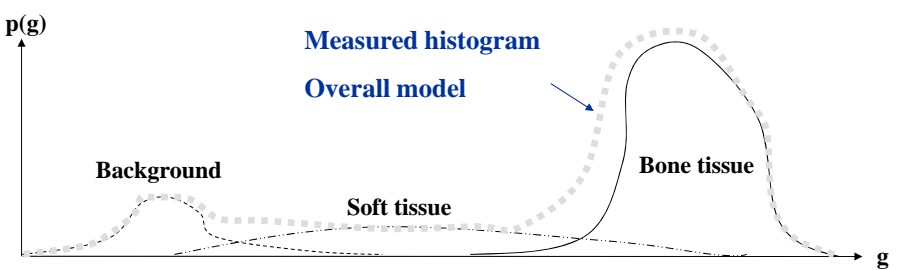
**Apprendimento Supervisionato:
Classificazione**

Task di classificazione
Uscita intera (etichetta o label della classe)

A.A. 2014-2015 http://borghese.di.unimi.it/

I modelli parametrici



$$p(g) = \sum_{j=1}^M P(j) \cdot p(g | j) = \sum_{j=1}^M w_j \cdot p_j(g)$$

La probabilità di avere un livello di grigio g è la somma pesata delle tre probabilità di avere background, $p_1(g)$, tessuto molle, $p_2(g)$ o tessuto osseo, $p_3(g)$.

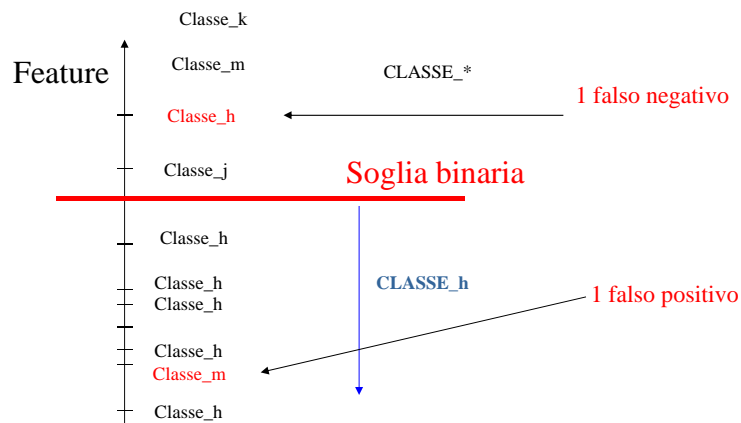
A.A. 2014-2015 http://borghese.di.unimi.it/



Classificatore binario

Classificatore binario. Si seleziona una feature e si sceglie la soglia ottimale.

Un classificatore binario è costituito da: feature, soglia, verso.



A.A. 2014-2015

47/57

<http://borghese.di.unimi.it/>



ADA(ptive) Boosting

Spesso non ci sono singole feature che consentono la classificazione corretta. Abbiamo feature “deboli”.

Il boosting consiste in un metodo incrementale che produce un classificatore composto da più classificatori elementari, binari, $h(I, threshold, sign, feature)$, in grado di minimizzare l’errore sul training set.

Combina più classificatori “deboli” in un classificatore performante.

Il risultato del booster è dato dal voto di maggioranza dei risultati pesati ottenuti da classificatori binari elementari che sono stati selezionati durante l’esecuzione del boosting.

$$H = \text{sign} \left(\sum_t \alpha_t h_t(\text{Image}; \text{threshold}, \text{sign}, \text{feature}) \right)$$

A.A. 2014-2015

48/57

<http://borghese.di.unimi.it/>



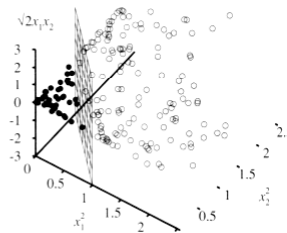
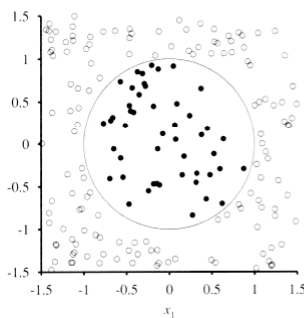
Support Vector Machines



Partizionamento dello spazio mediante una retta (V. Vapnik, 1998).

Minimizzano l'errore di (mis)classificazione e, tra tutti i classificatori che minimizzano questo errore, scelgono quello che **massimizza il margine**.

Applicazione di una trasformazione non lineare che mappa lo spazio di input in uno spazio a più dimensioni in cui i dati risultano linearmente separabili.



Mapping is defined by:

$$f_1 = x_1^2 \quad f_2 = x_2^2$$

$$f_3 = \sqrt{2}x_1x_2$$

Russel Norvig, 2nd edition

A.A. 2014-2015

49/57

<http://borghese.di.unimi.it/>



Support vector machines (geometrical view)



$w \cdot x = b$ is the plane

$\frac{b}{\|w\|}$ distance from O

For all points holds:

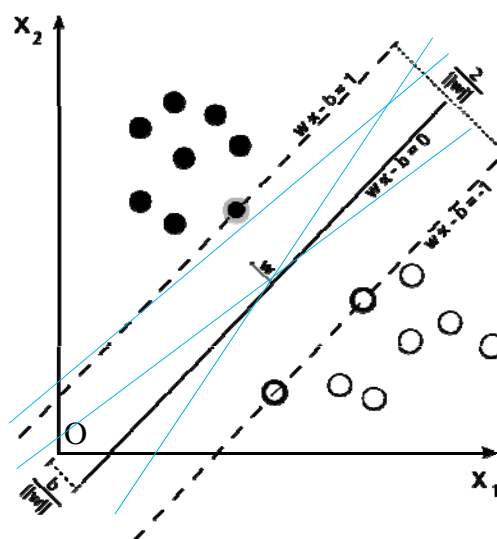
a) $w \cdot x_i - b \geq +1$ for $y_i = 1$

b) $w \cdot x_i - b \leq -1$ for $y_i = -1$

In compact form:

$$y_i(w \cdot x_i - b) - 1 \geq 0 \quad \forall i$$

Many w and b can be chosen, only one pair minimizes the space between the line and the closest point (the margin).



A.A. 2014-2015

50/57

<http://borghese.di.unimi.it/>



Support vector machines: the margin amplitude



For all points holds:

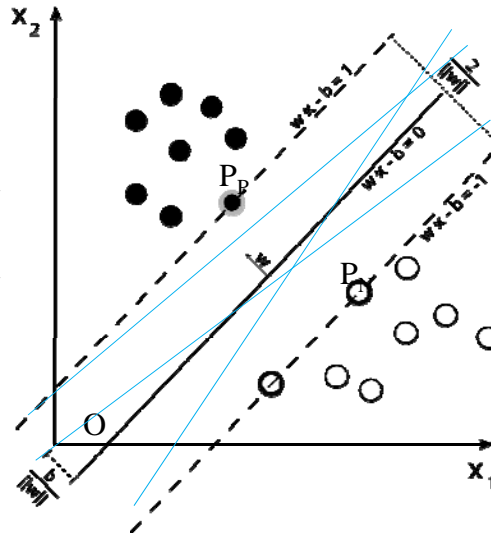
- a) $w x_i - b = +1$ for closest positive point, P_P .
- b) $w x_i - b = -1$ for closest negative point, P_N .

Distance between the line and P_P is: $\frac{1+b}{\|w\|}$

Distance between the line and P_N is: $\frac{1-b}{\|w\|}$

Total margin is: $\frac{2}{\|w\|}$

Come determinare w e b ?



A.A. 2014-2015

51/57

<http://borghese.di.unimi.it/>



Support vector machines: computation of w and b



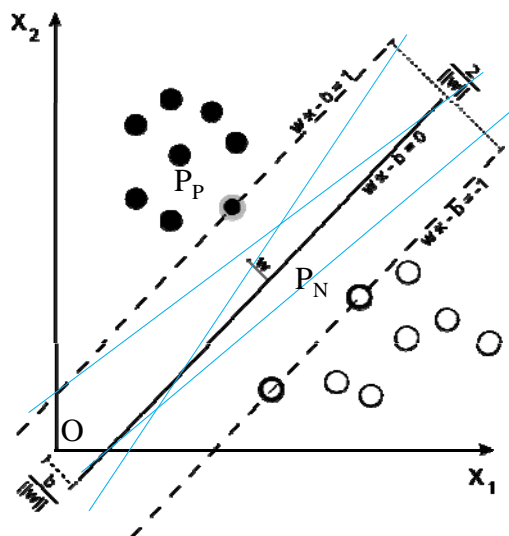
A) $y_i(w x_i - b) - 1 \geq 0 \forall i$

B) Total margin is: $\frac{2}{\|w\|}$

Massimizzo $\|w\|$ (B, angular coefficient of the line) with the constraint that classification, A, is correct.

The solution will be a line equally distant from two parallel lines through P_P and P_N .

Problema di massimo vincolato che viene trasformato in un problema di minimo vincolato di $\|w\|^2$



A.A. 2014-2015

52/57

<http://borghese.di.unimi.it/>



Support vector machines: Lagrange multipliers



Scrivo la funzione di Lagrange: $L(w, b) = \frac{1}{2} \|w\|^2 - \sum_i \alpha_i y_i (x_i \cdot w - b) + \sum_i \alpha_i$

Problema quadratico convesso con 1 minimo

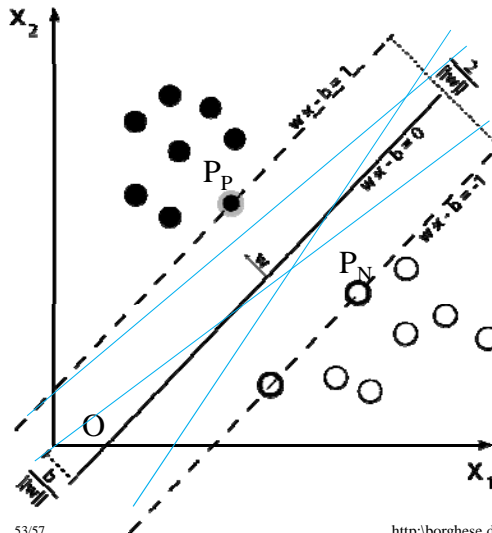
We have to compute:

$$\min_{w, b, \{\alpha_i\}} L(w, b)$$

with:

$$y_i (w x_i - b) - 1 \geq 0$$

$$\alpha_i \geq 0$$



A.A. 2014-2015

53/57

<http://borgnese.di.unimi.it/>



Condizioni KKT



Scrivo la funzione di Lagrange: $L(w, b) = \frac{1}{2} \|w\|^2 - \sum_i \alpha_i y_i (x_i \cdot w - b) + \sum_i \alpha_i$

We have to compute:

$$\min_{w, b, \{\alpha_i\}} L(w, b)$$

with:

$$y_i (w x_i - b) - 1 \geq 0$$

$$\alpha_i \geq 0$$

Karush-Kuhn-Tucker conditions applied:

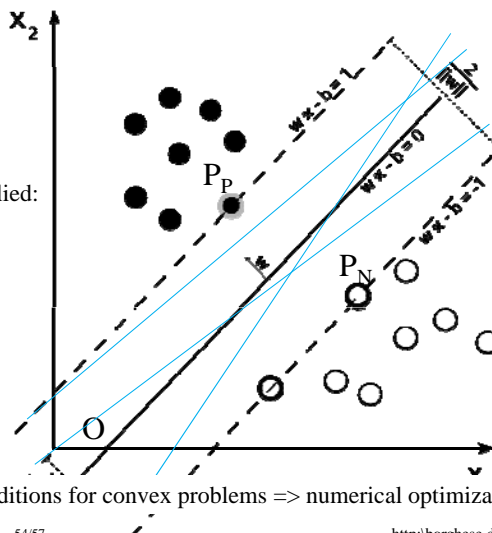
$$\frac{\partial L(\cdot)}{\partial w} = w - \sum_i \alpha_i y_i x_i = 0$$

$$\frac{\partial L(\cdot)}{\partial b} = -\sum_i \alpha_i y_i = 0$$

$$\alpha_i (y_i (x_i \cdot w + b) - 1) = 0$$

$$y_i (x_i \cdot w + b) - 1 \geq 0 \quad \forall i$$

$$\alpha_i \geq 0$$



KKT are necessary and sufficient conditions for convex problems \Rightarrow numerical optimization.

A.A. 2014-2015

54/57

<http://borgnese.di.unimi.it/>



Support vector machines: dual formulation



$$\min_{w,b,\{\alpha_i\}} L(w,b) \text{ with } \alpha_i \geq 0$$

$$L(w,b) = \frac{1}{2} \|w\|^2 - \sum_i \alpha_i y_i (x_i \cdot w - b) + \sum_i \alpha_i$$

Problema quadratico convesso con 1 minimo:
 - Funzione obbiettiva convessa
 - I punti che soddisfano i vincoli costituiscono dei semipiani -> spazi convessi.

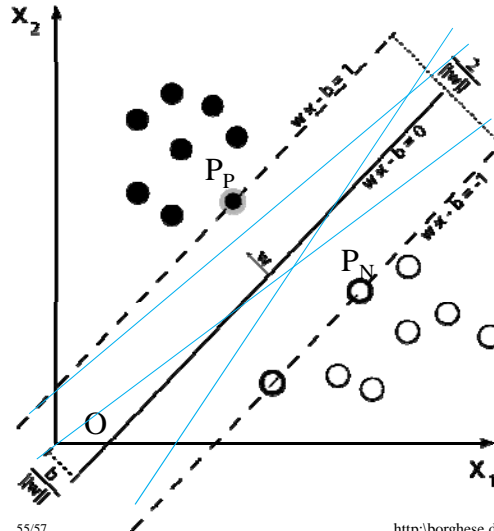
We solve the dual problem:

$$\max_{\{\alpha_i\}} L(w,b)$$

with constraints:

$$\frac{\partial L}{\partial w} = 0 \quad \frac{\partial L}{\partial b} = 0$$

$$\alpha_i \geq 0$$



A.A. 2014-2015

55/57

<http://borghese.di.unimi.it/>



Dual formulation contains an inner product



$$\max_{\{\alpha_i\}} L(w,b)$$

$$L(w,b) = \frac{1}{2} \|w\|^2 - \sum_i \alpha_i y_i (x_i \cdot w - b) + \sum_i \alpha_i$$

$$\frac{\partial L}{\partial w} = 0 \quad \sum_i \alpha_i y_i x_i = w$$

$$\frac{\partial L}{\partial b} = 0 \quad \sum_i \alpha_i y_i = 0$$

Si possono sostituire nell'equazione del massimo, eliminando così i vincoli e rimanendo con la sola funzione da massimizzare:

$$\max_{\{\alpha_i\}} L(w,b)$$

$$L(w,b) = -\frac{1}{2} \sum_{i,j} \alpha_i y_i y_j x_i \cdot x_j + \sum_i \alpha_i$$

Only the inner product of the data appears here.

A.A. 2014-2015

56/57

<http://borghese.di.unimi.it/>



Riassunto



- Supervised learning: predictive regression.
- Regressione multi-scala
- Classificazione