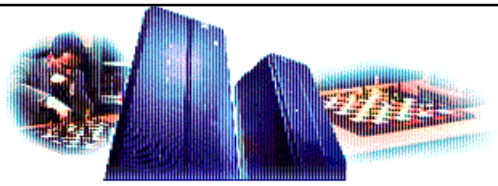
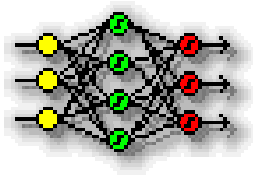




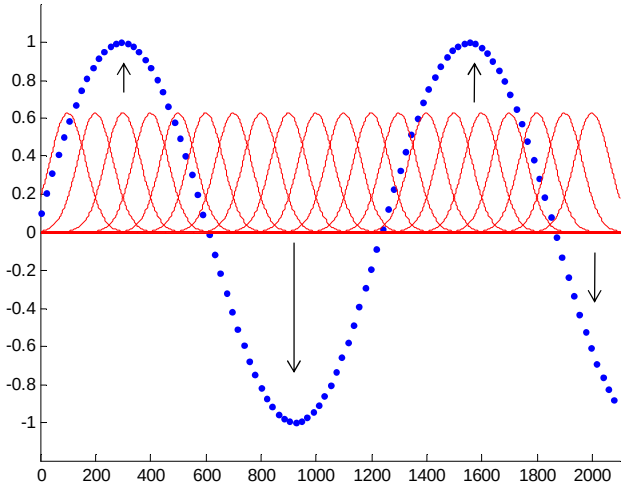
Sistemi Intelligenti
Supervised learning

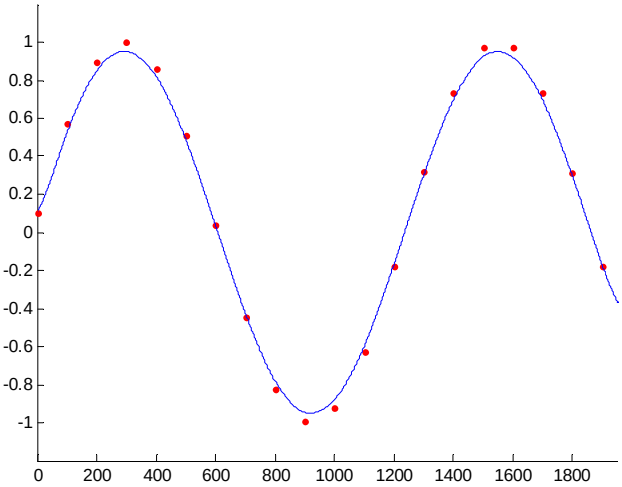
Alberto Borghese
Università degli Studi di Milano
Laboratorio di Sistemi Intelligenti Applicati (AIS-Lab)
Dipartimento di Informatica
Alberto.borghese@unimi.it

Riassunto

- Regressione multi-scala
- Scelta del modello
- Classificazione

Funzionamento di un modello semi-parametrico (lineare)





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

Devo definire, gli M $\{w_i\}$.
3 << M << N – 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.

Surface reconstruction with filtering

- Convolution: $\hat{f}(x) = \int_{\mathbb{R}} f(c)\, G(x-c|\sigma)\, \mathrm{d}c = f(x) * G(x;\sigma)$

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(.) is normalized, is obtained through digital filtering.

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

Filters and bases

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

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

$\frac{\Delta x_k}{\sqrt{\pi}\sigma}$ Normalization factor

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

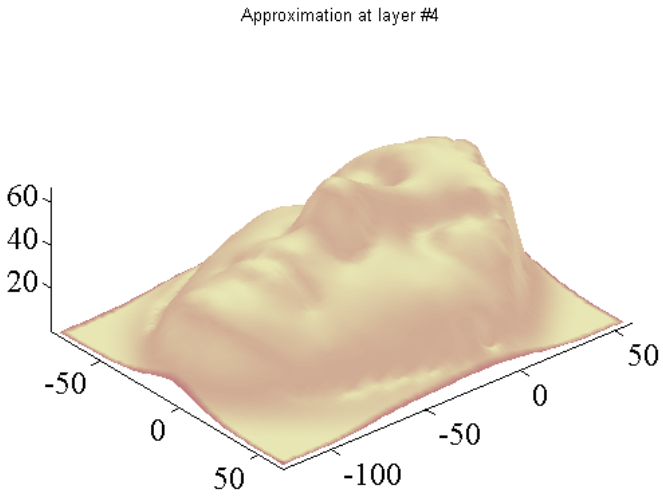
Riesz basis, the approximation space is characterized by the scale of the basis that determines the amplitude of the space.

A sequence of spaces can be defined according to σ:

σ_0 -> V_0; σ_1 -> V_1; σ_2 -> V_2….

The number of representable functions increases.

Problema dell'overfitting dovuto a sovraparametrizzazione



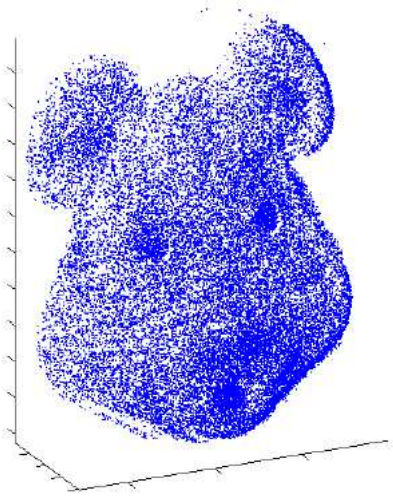
Quante unità?

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.



Gridding

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

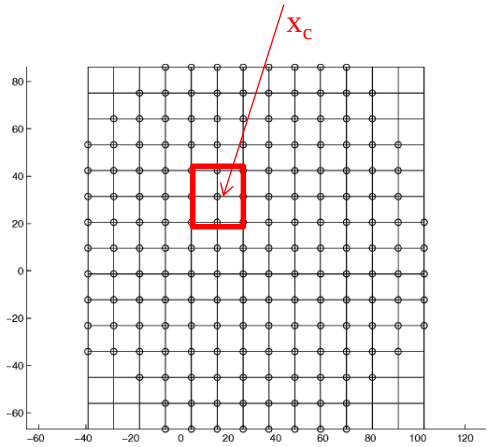
How can we determine w_k from points clouds?

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

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

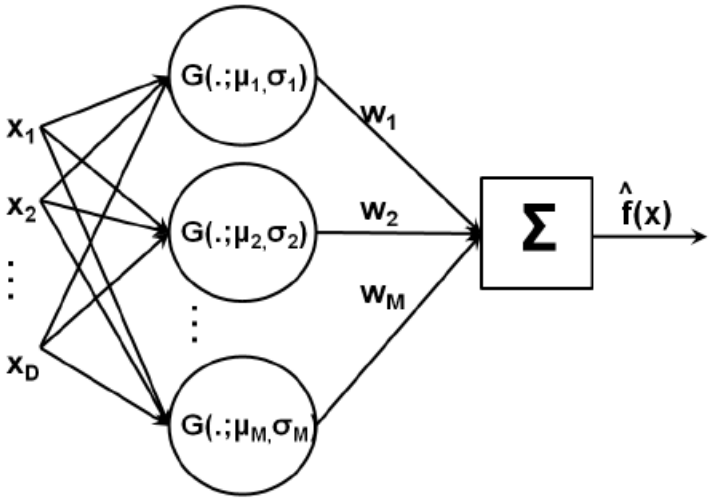
$K_\sigma(.)$ Gaussiana

Parzen-window estimators.



RBF Network

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



Perceptron

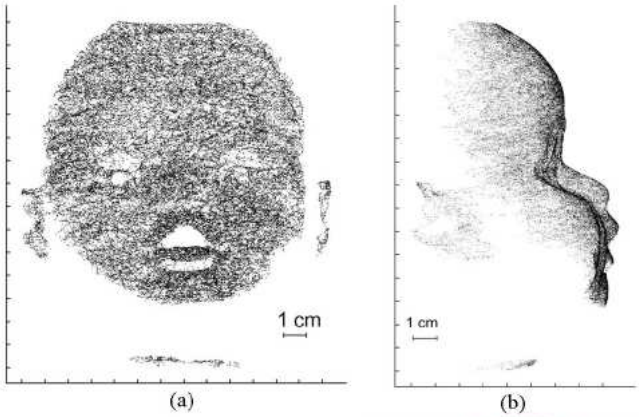
Surface Approximation

- Properties:

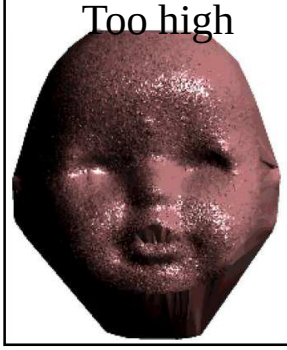
- Redundancy.

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

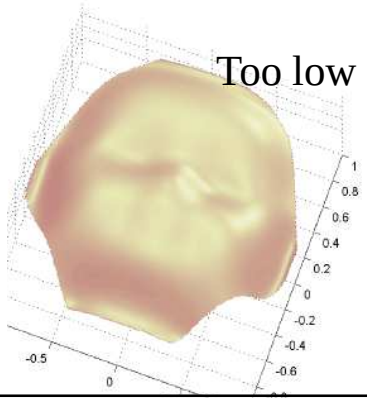
Which scale?

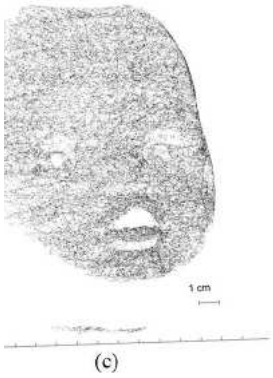


Too high



Too low

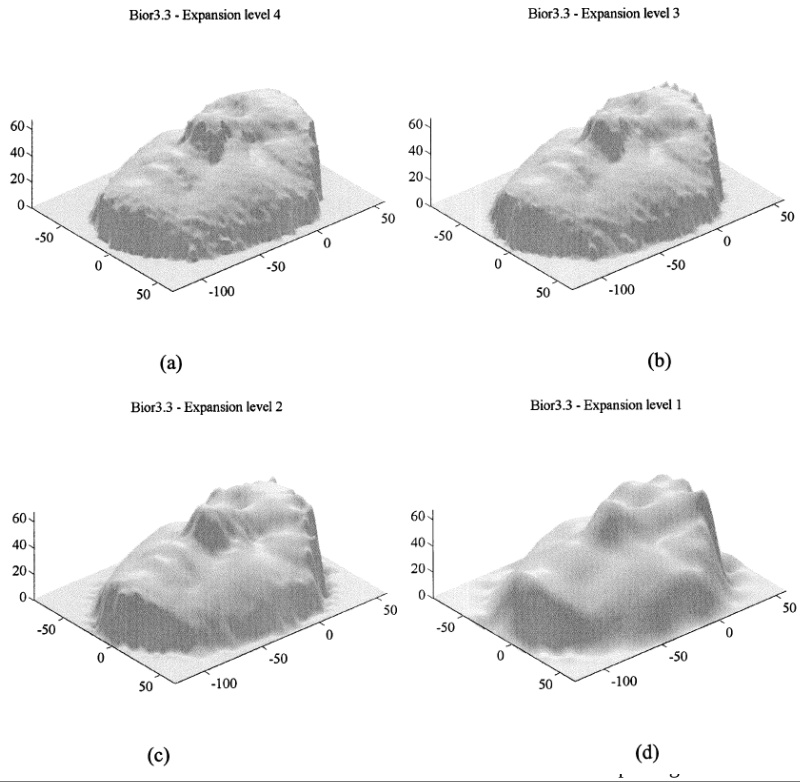




http:\borghese.di.unimi.it\

Pyramidal reconstruction

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



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.

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

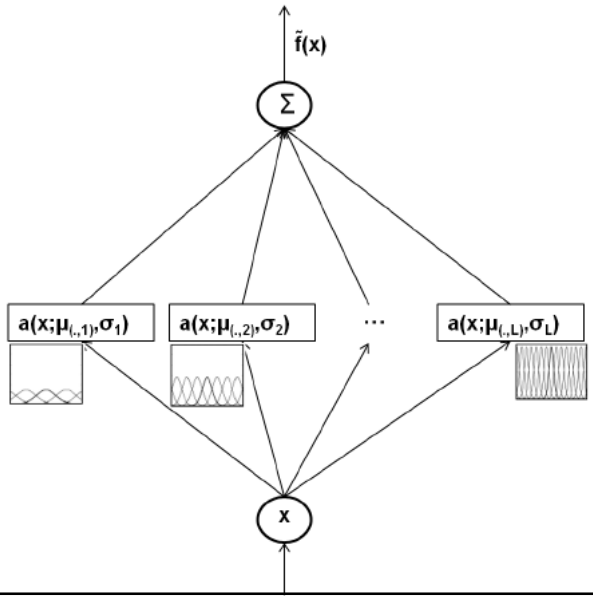
- Can we use a single Δx?

Incremental approximation with local adaptation.

Start from low resolution

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

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



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

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

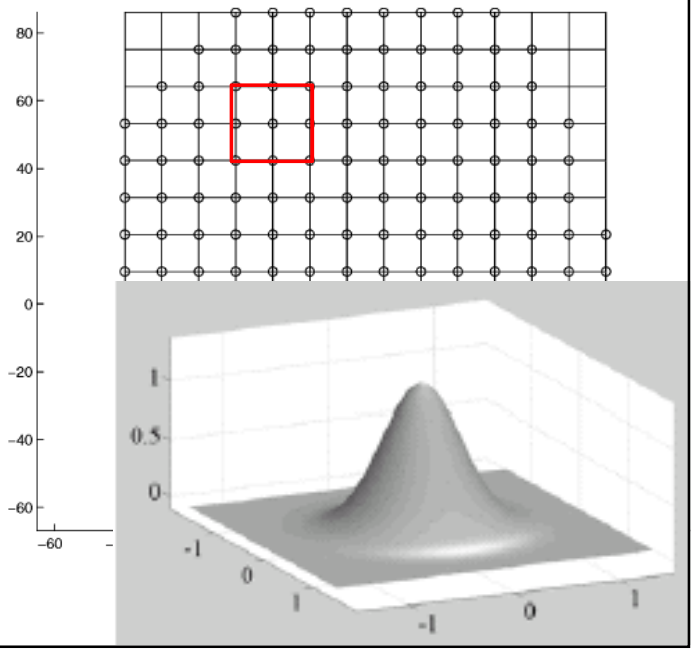
Determination of the surface height

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

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

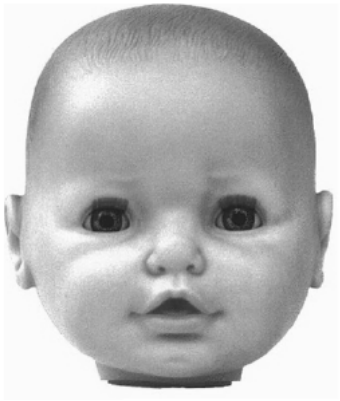
Apply local estimator to the data points in the neighbourhood of a grid crossing (Gaussian center) to compute f_k.

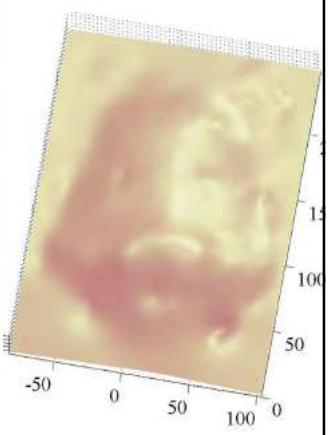
Sorting of the data is made simple, they are subdivided into quads. Identified the points inside the neighbourhood is equivalent to extract all the points between two positions in the data vector.



We can obtain a «poor» reconstruction

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



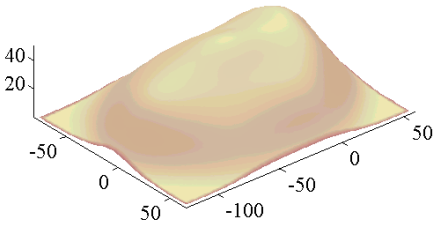


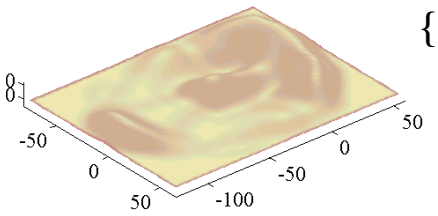
What can be done?

We can compute the residual for each data point.

Approximation at layer #l

$\{r_1(\mathbf{x})\}$

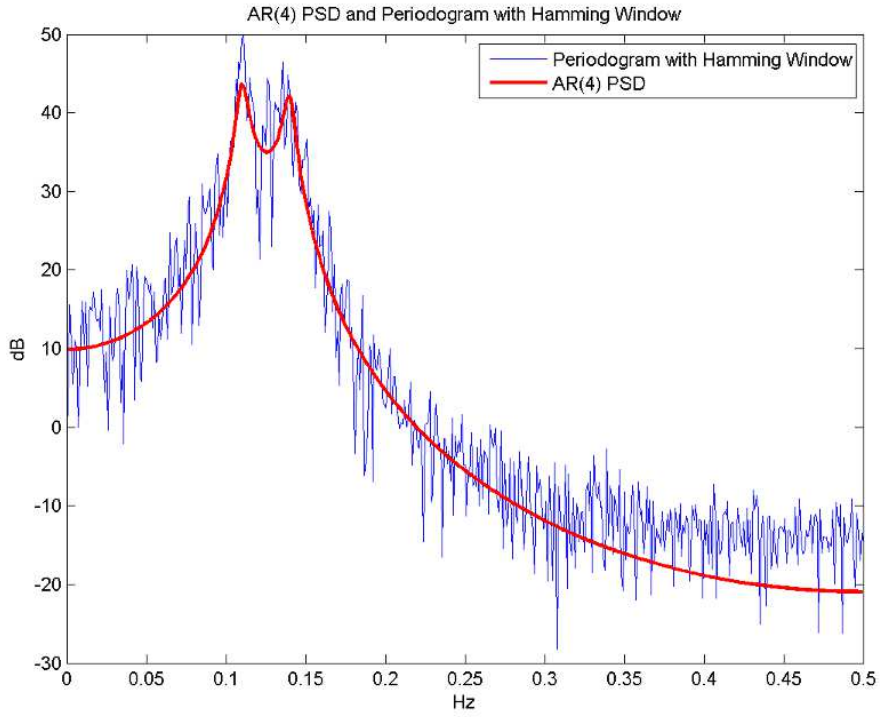




We evaluate the residual for each data point: $r_1 = dist\left(y_m, \hat{f}(x_m)\right)$

E.g.: $r_1 = \left(y_m - \hat{f}(x_m)\right)^2$ $r_1 = \left|y_m - \hat{f}(x_m)\right|$

Where does this residual come from?

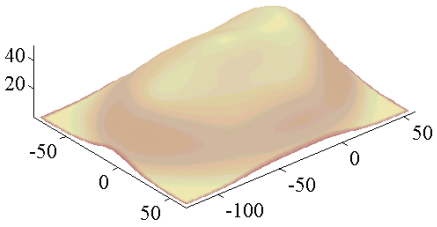


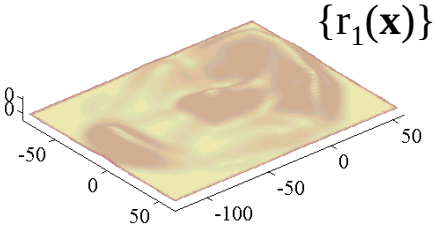
We want to eliminate variability due to noise.

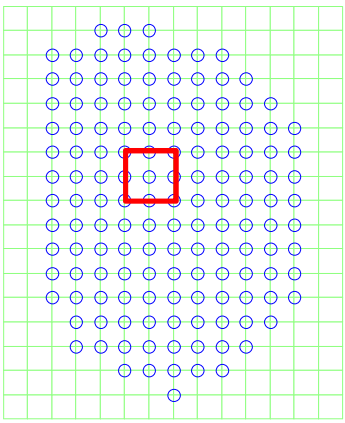
Is the residual adequate?

Approximation at layer #1

$\{r_1(\mathbf{x})\}$



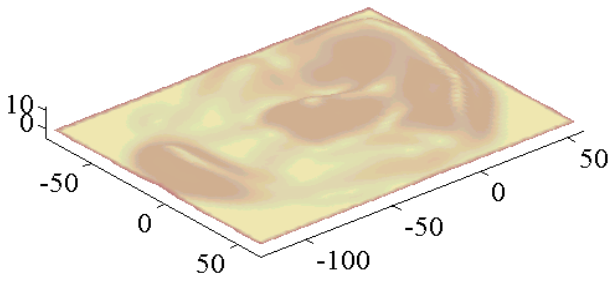




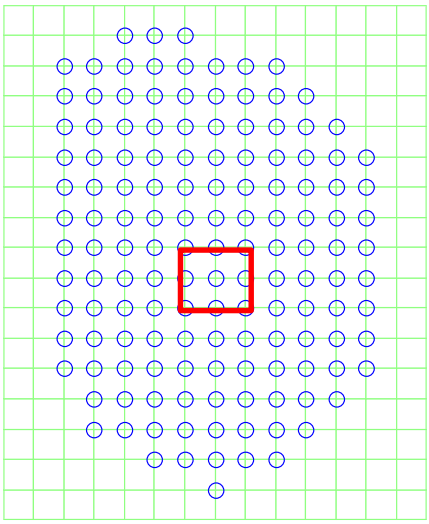
For each Gaussian the integral of the residual inside the “receptive field” of the Gaussian, is assumed as local approximation error associated to it. , is computed inside its “receptive field”:

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

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:

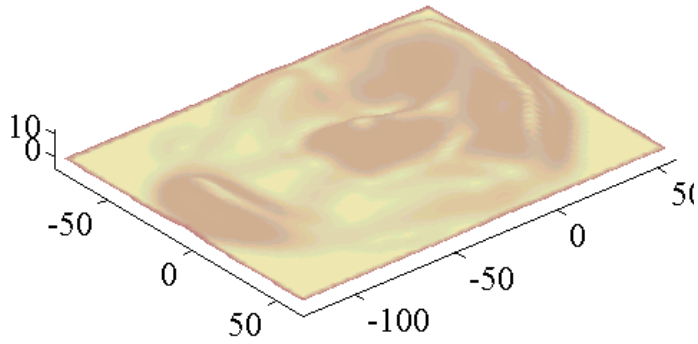
- Degree of approximation
- Noise: RMS.

Layer 2

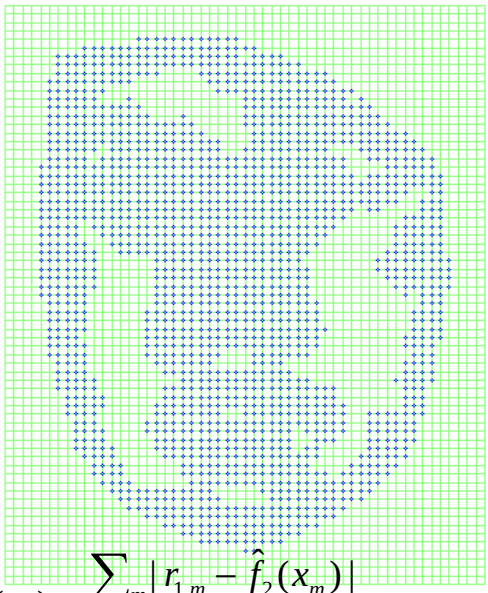
Input are the residuals, $r_{1,m} = \left| y_m - \hat{f}_1(x_m) \right|$

Output is the model that approximates $r_{1,m}$: $f_2(x_m) \rightarrow r_{1,m}$

Output of layer #2



Layer #2

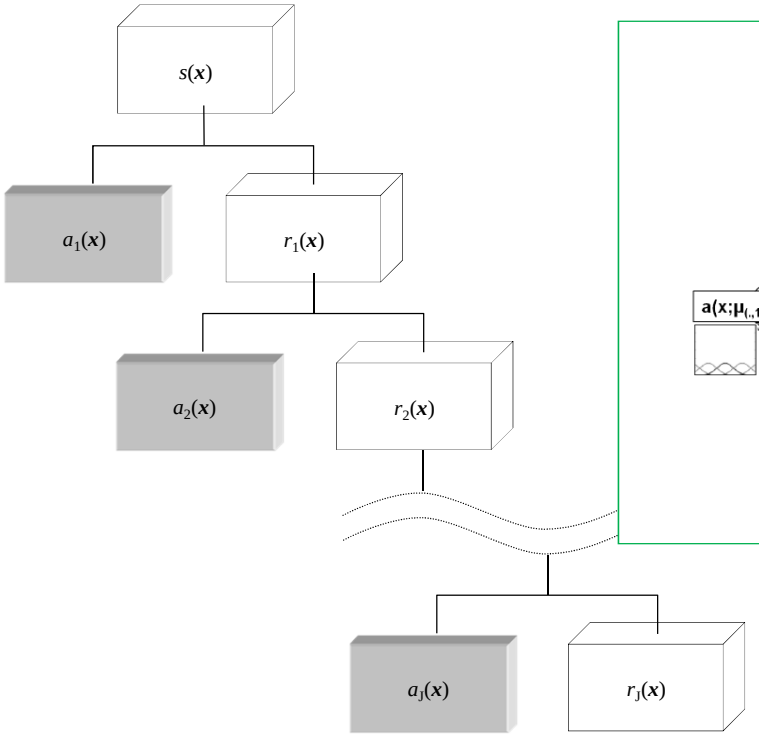


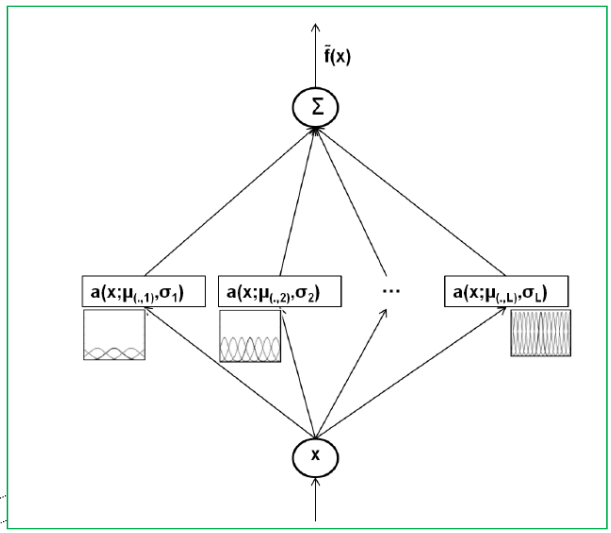
More packed Gaussians

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

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

Hierarchy construction

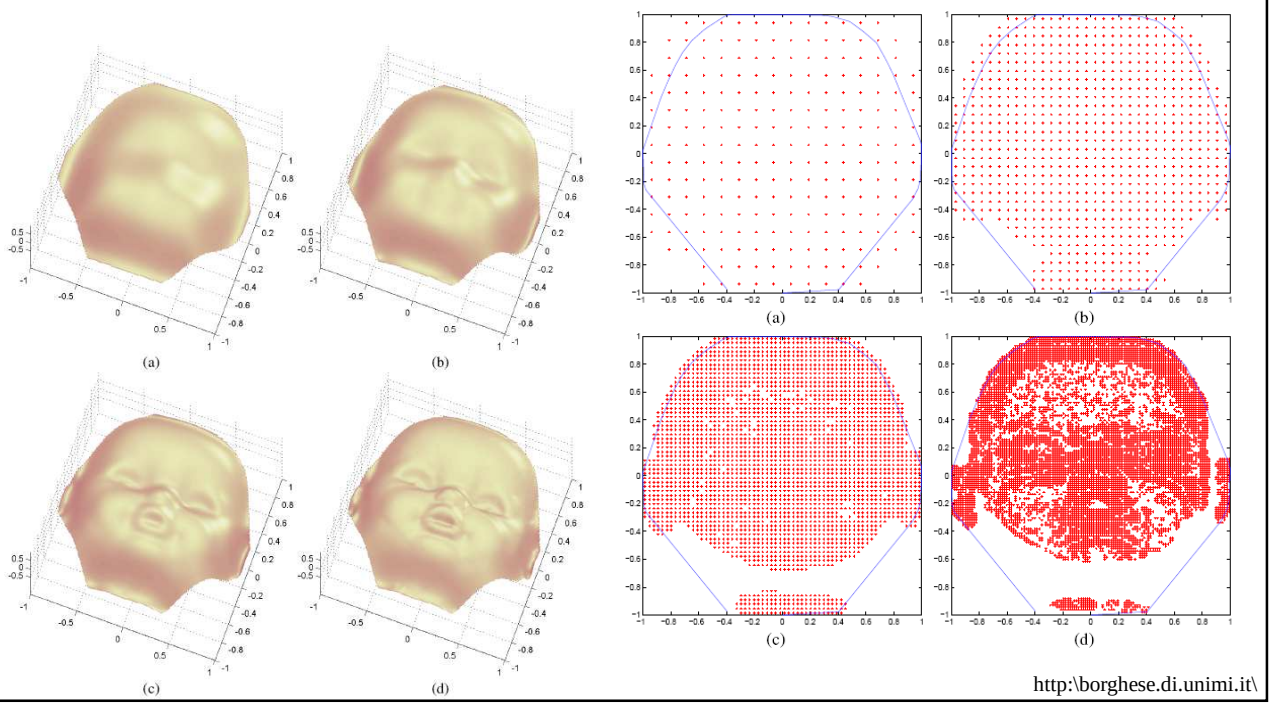




and use as a stack of layers

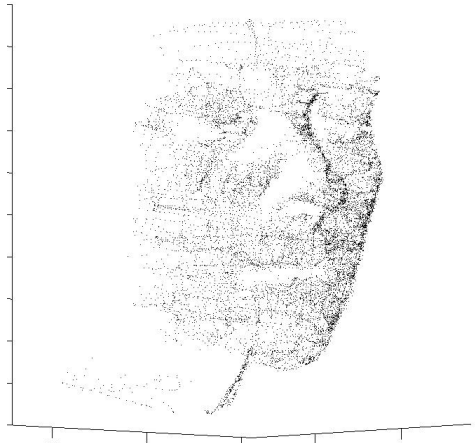
How to operate on large sets of data?

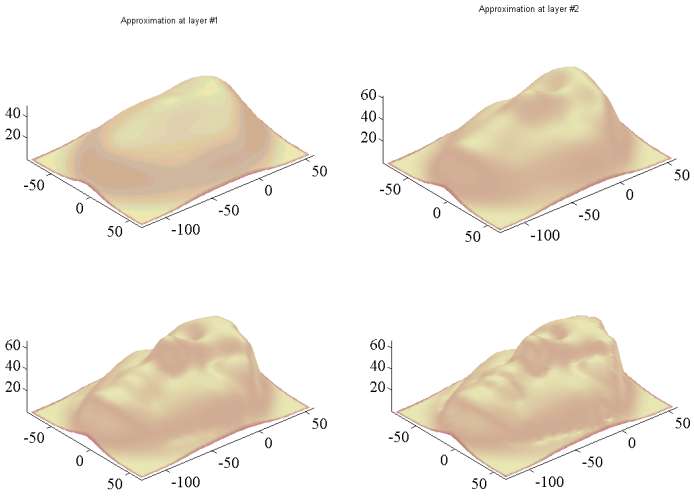
Recursive splitting of the quad domain -> local re-ordering of the data.

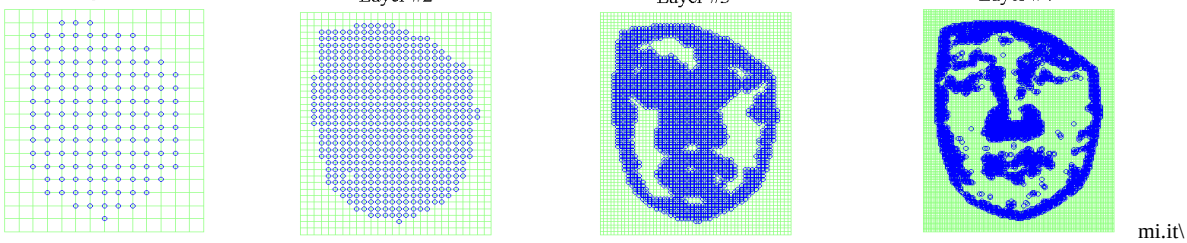


http:\borghese.di.unimi.it\

Applicazione della regressione





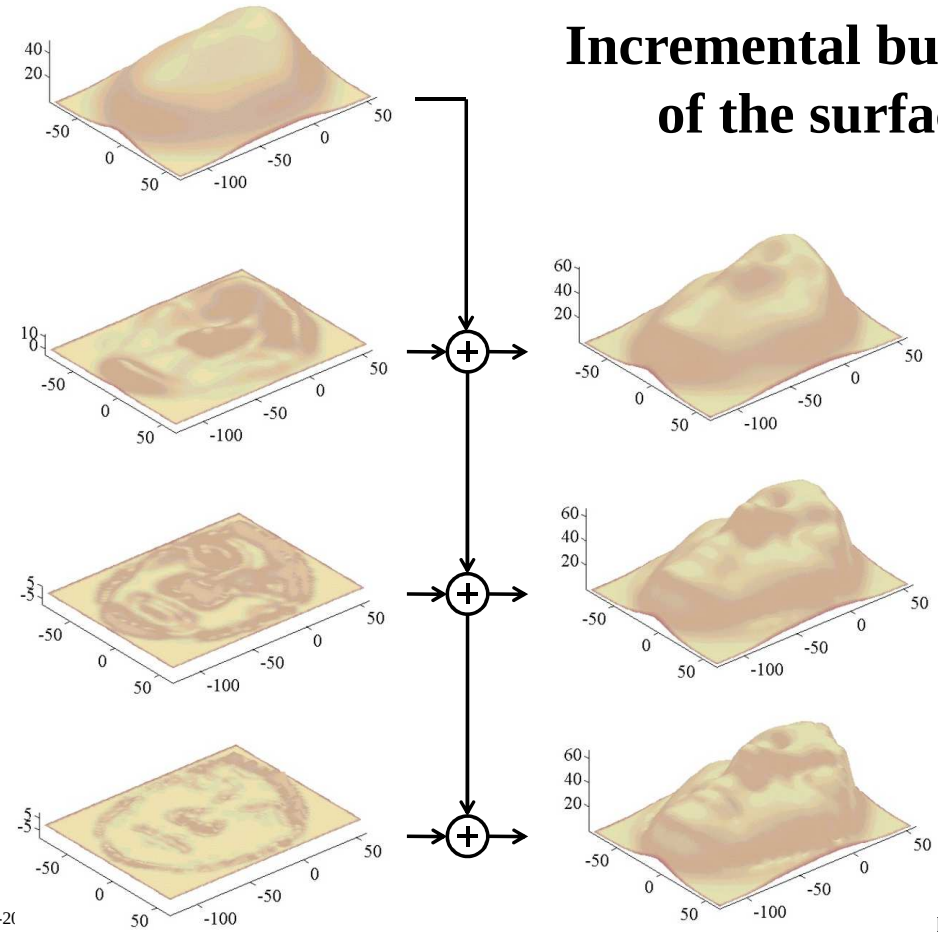


Characteristics of HRBF networks

- 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

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

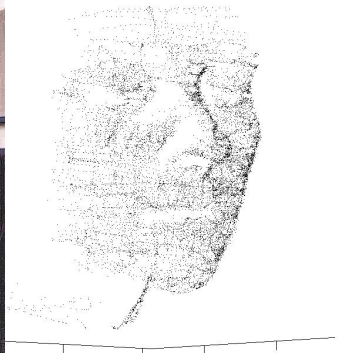
Incremental building of the surface



On-line version

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

hrbf_online.wmv



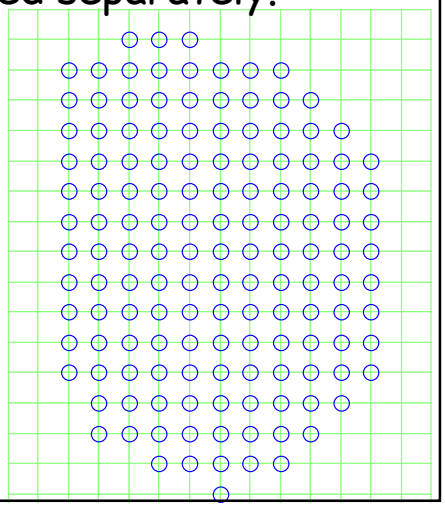
Observation

- 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.

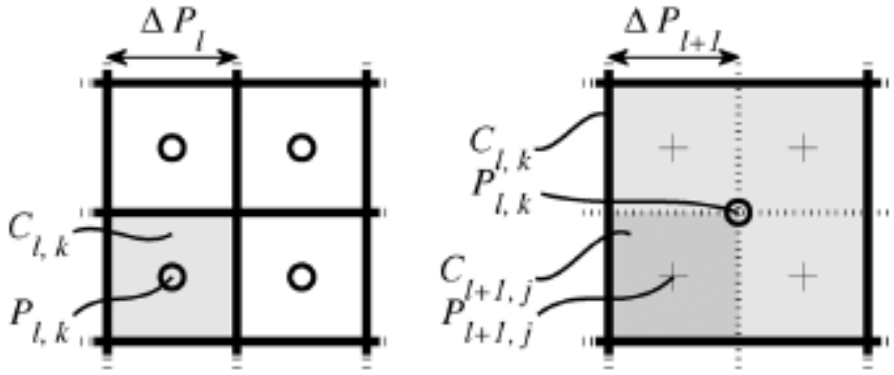
$$\hat{f}(x) = \frac{\sum_i y_i K_\sigma(x_i,x)}{\sum_i K_\sigma(x_i,x)} = \frac{\sum_i y_i \, e^{\frac{-||x_i-x||^2}{\sigma^2}}}{\sum_i e^{\frac{-||x_i-x||^2}{\sigma^2}}}$$

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



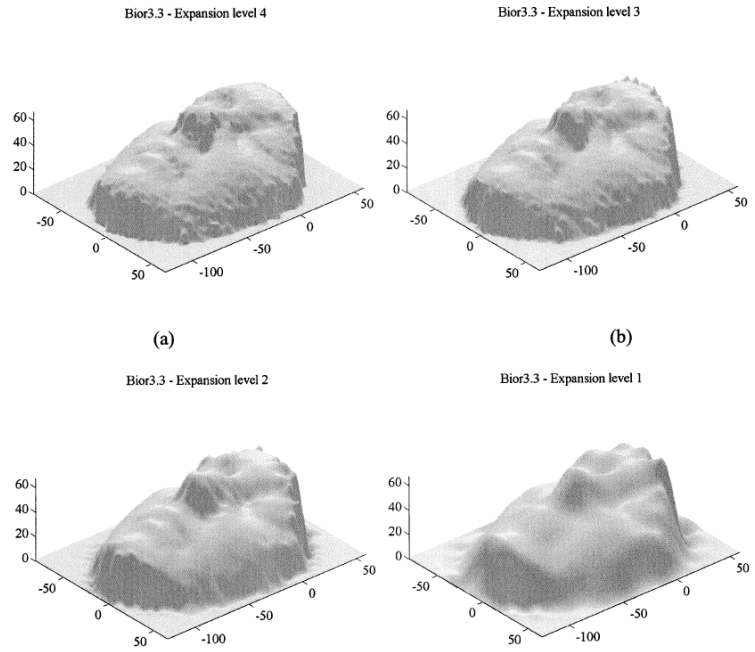
Local operations

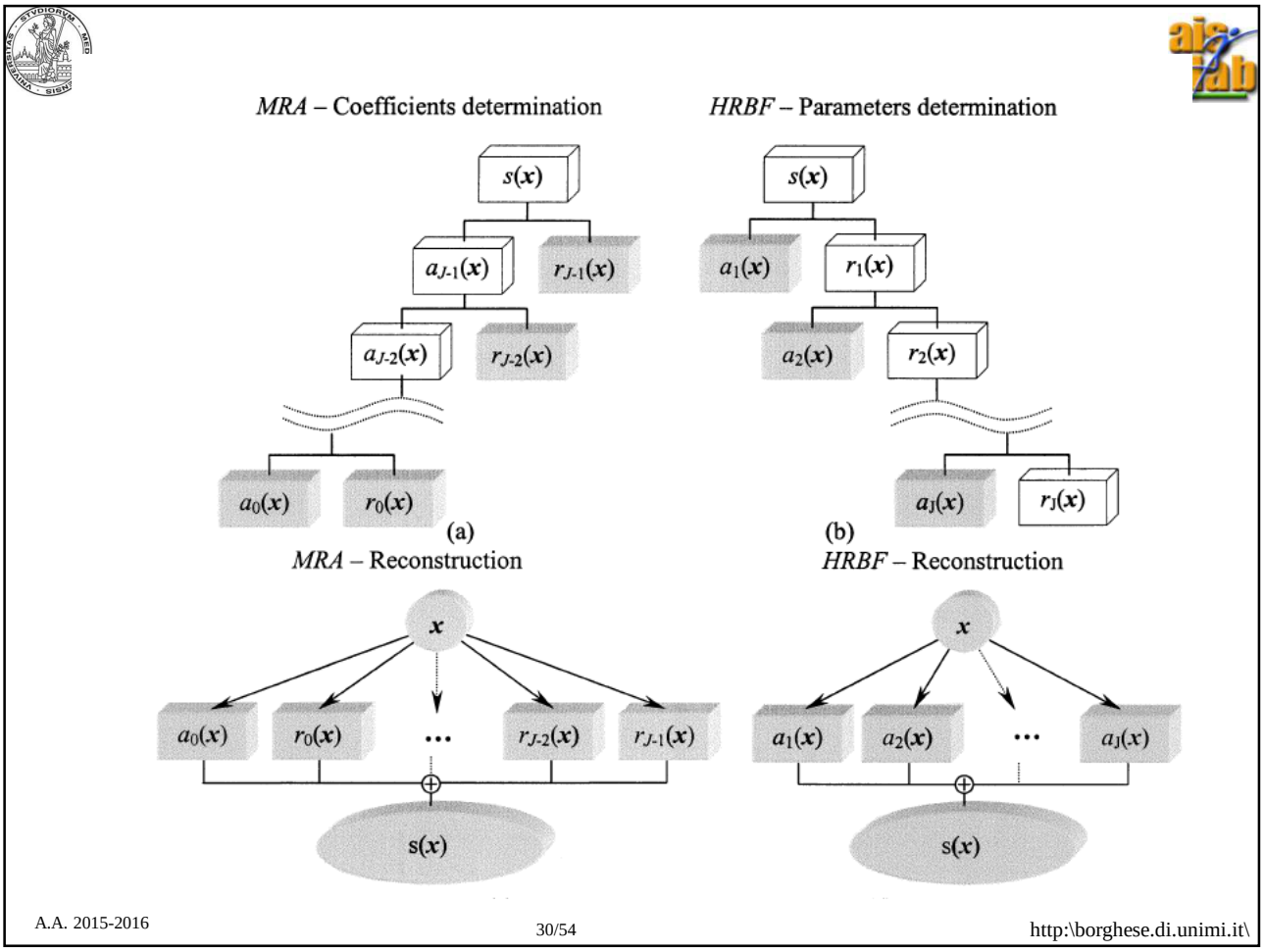
- Local splitting of each quad is achieved when:
  - Residual is higher than threshold
  - Enough points have been sampled



Comparison with Wavelets

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




MRA – Coefficients determination
HRBF – Parameters determination
s(x)
a_{J-1}(x)
r_{J-1}(x)
a_{J-2}(x)
r_{J-2}(x)
a_0(x)
r_0(x)
(a)
a_1(x)
r_1(x)
a_2(x)
r_2(x)
a_J(x)
r_J(x)
(b)
MRA – Reconstruction
HRBF – Reconstruction
x
s(x)

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.

Riassunto

- Regressione multi-scala
- Scelta del modello
- Classificazione

How to classify the error introduced by a model?

Is the model good enough?

Does it have enough parameters?

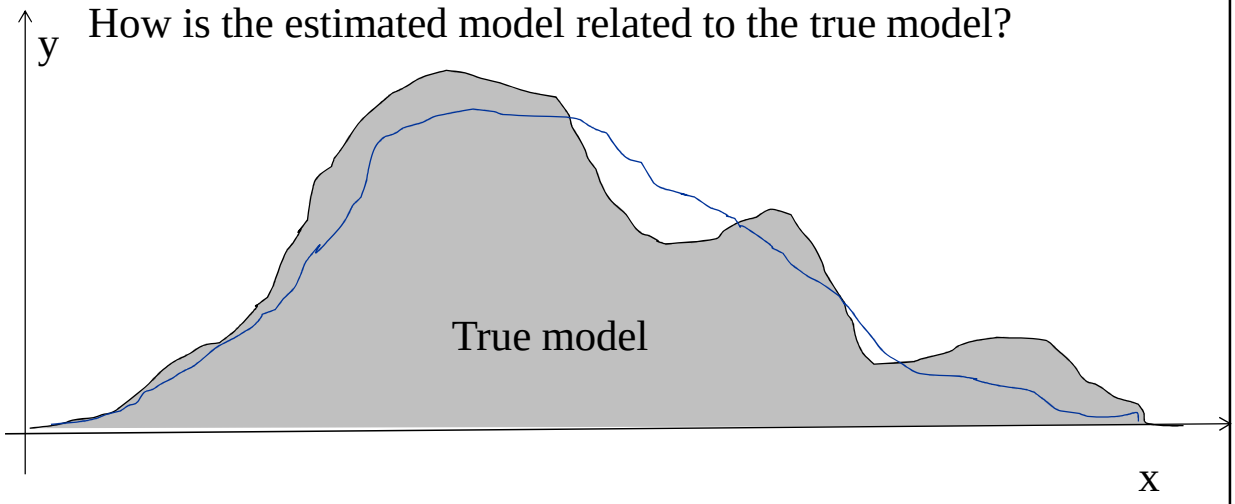
Does it cover the input domain (in all dimensions)?

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?

How is the estimated model related to the true model?

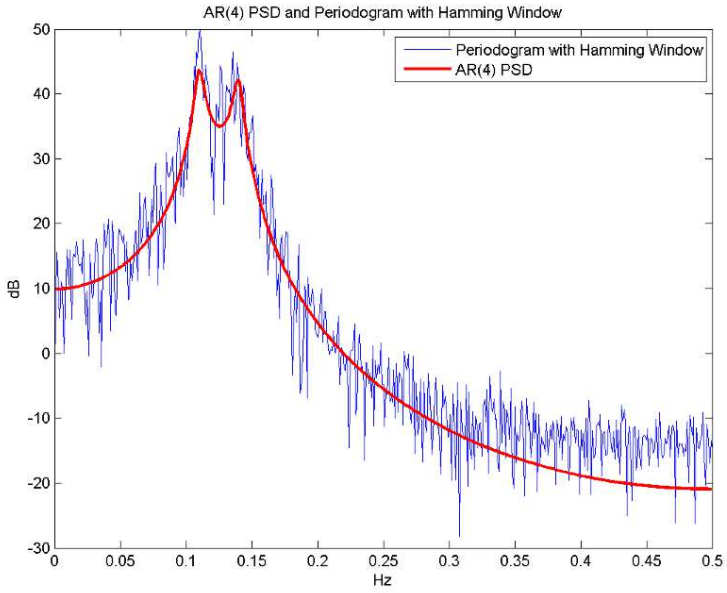


Bias and variability trade-off

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

Variability

How are the measured points related to the estimated model?



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

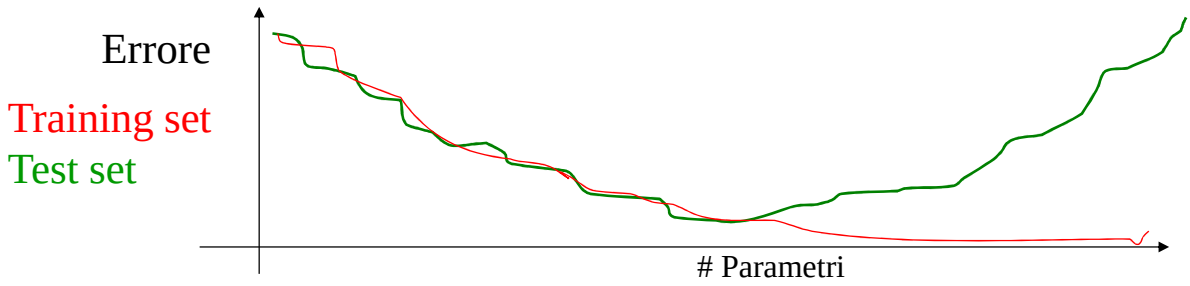
If variability goes to zero, bias increases and overfitting arises.
In a good model, variability tends to the statistics of the measurement noise.

Scelta empirica – cross-validation

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

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

*Il numero di parametri viene aumentato fino a quando **entrambi** gli errori diminuiscono.*



Scelta teorica

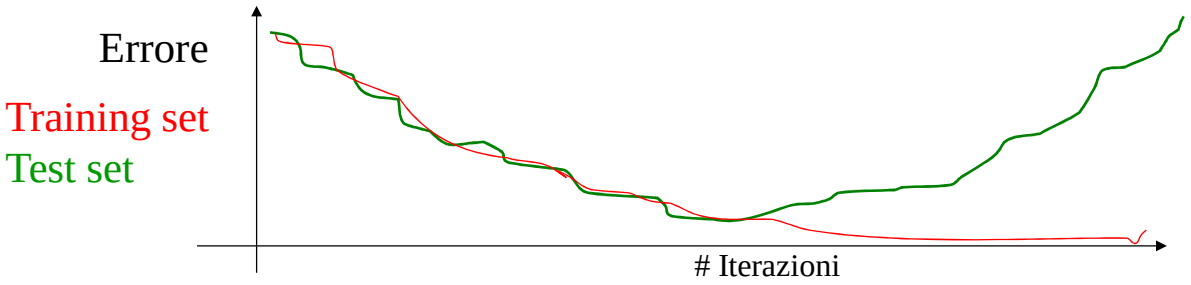
Quale funzione costo minimizzo? Come posso inserire l'informazione di complessità nella funzione costo?

Penalizzo i modelli con tanti parametri. Regularization with Reproducible Hilbert Kernels as regularizers

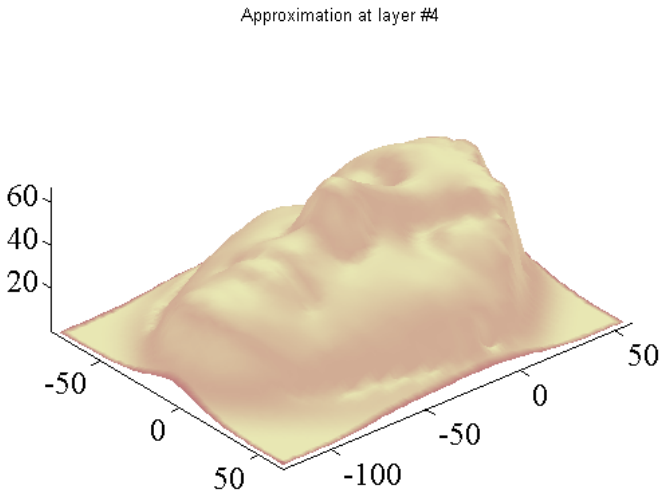
Altri approcci

Semi-convergenza: non porto l'algoritmo fino alla convergenza nel punto di ottimo ma arresto le iterazioni prima.

Il modello non sarà perfettamente aderente ai dati, ma il residuo sarà tendenzialmente l'errore di misura.



Problema dell'overfitting dovuto a sovraparametrizzazione

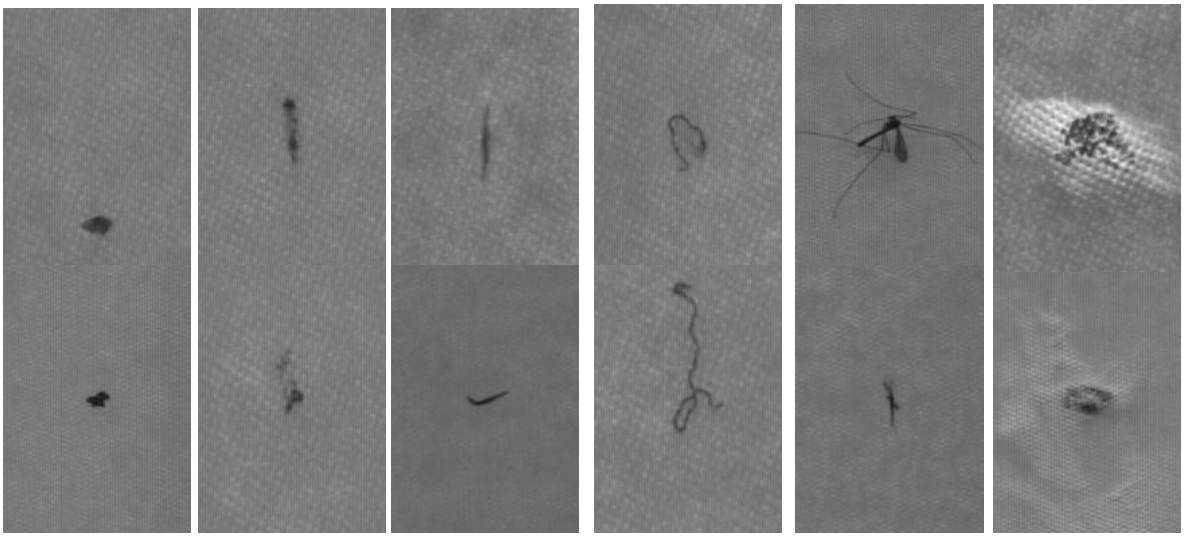


Quante unità?

Riassunto

- Supervised learning
- Regressione multi-scala
- Classificazione

CLASSIFICAZIONE: Riconoscimento difetti in linee di produzione
(progetto finanziato da Electronic Systems: 2006-2007)



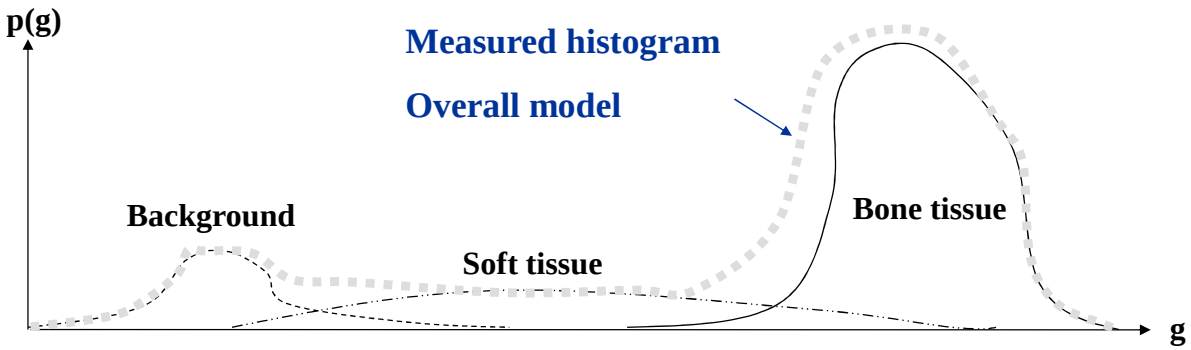
regolari | irregolari | allungati | fili | insetti | macchie su denso

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

Apprendimento Supervisionato: Classificazione

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

I modelli parametrici



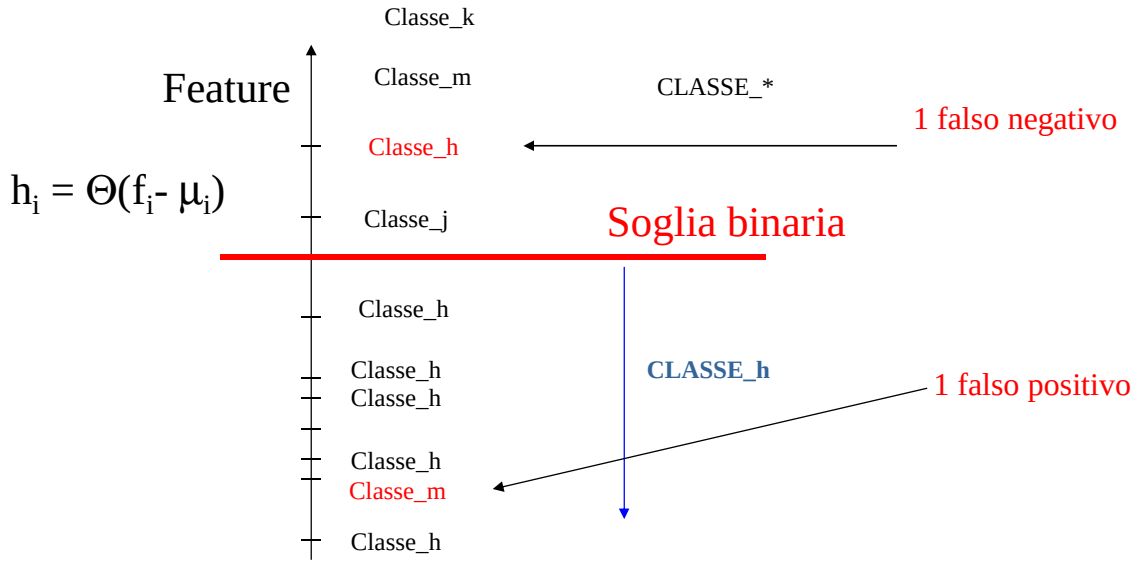
$$p(g) = \sum_{j=1}^{M} P(j) \cdot p(g \mid j) = \sum_{j=1}^{M} w_j \cdot p_j(g)$$

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, $p_3(g)$.

Classificatore binario

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

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



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.

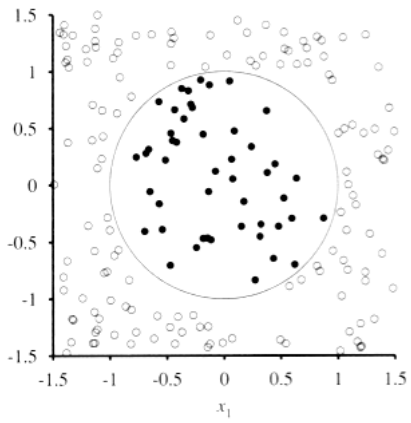
$$H = sign\left(\sum_t \alpha_t h_t(Image; threshold, sign, feature)\right)$$

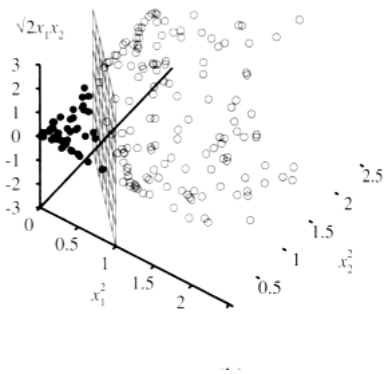
Support Vector Machines

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 (mediante funzioni Kernel) che mappa lo spazio di input in uno spazio a più dimensioni in cui i dati risultano linearmente separabili





Mapping is defined by:
$f_1 = x_1^2$ $f_2 = x_2^2$
$f_3 = \text{sqrt}(2)x_1x_2$
Russel Norvig, 2nd edition

Support vector machines (geometrical view)

w x = b is the plane

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

For all points holds:

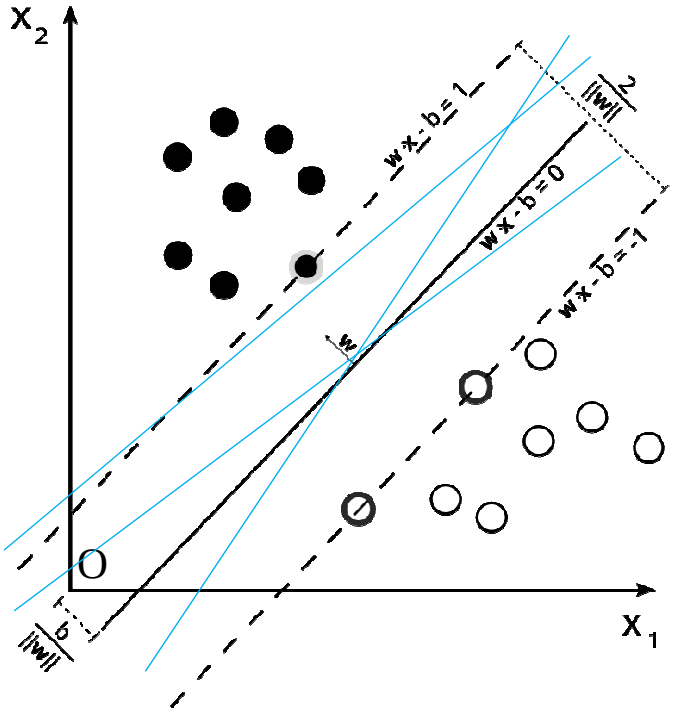
a) $w x_i - b \geq +1$ for $y_i = 1$

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

In compact form:

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

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



Support vector machines: the margin amplitude

For all points holds:

a) $w\,x_i - b = +1$ for closest poisitive point, P_P.

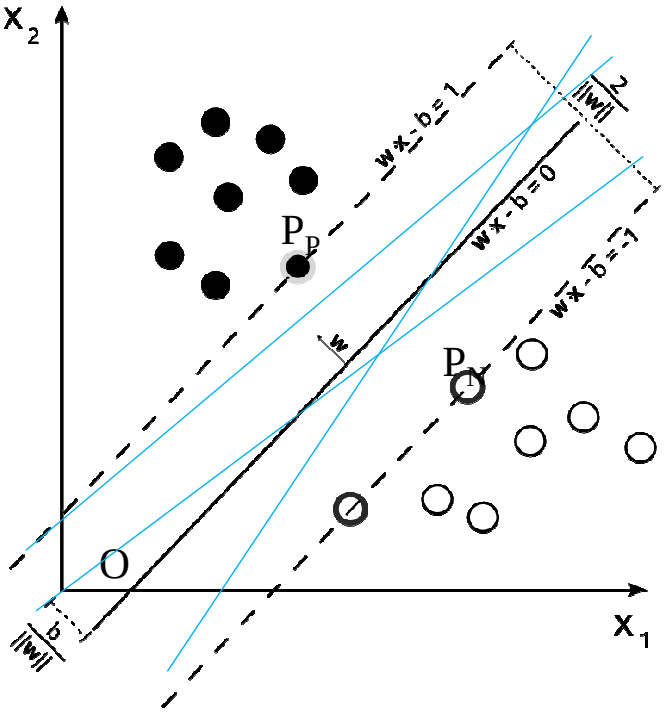
b) $wx_i - b = -1$ for closest negative point, P_N.

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?



Support vector machines: computation of w and b

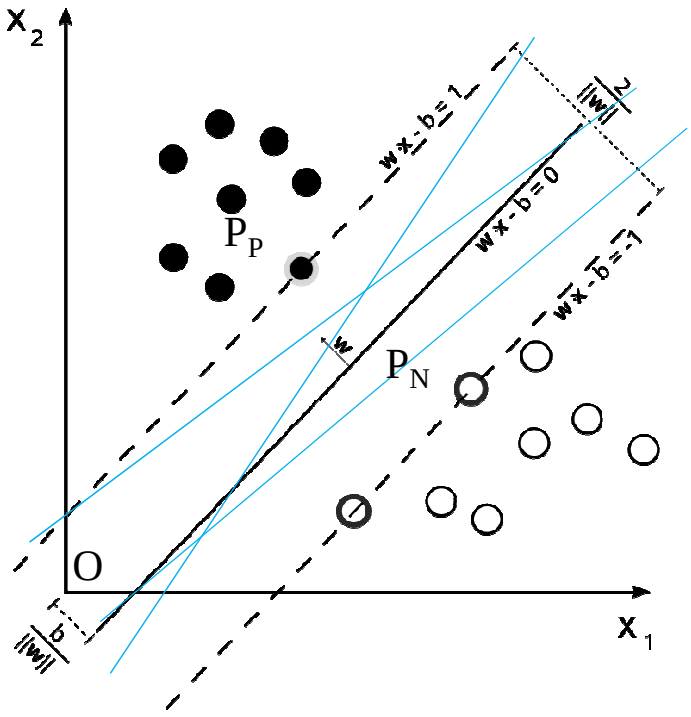
A) $y_i(wx_i-b) - 1 \geq 0 \ \forall i$

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 $\|w\|^2$



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

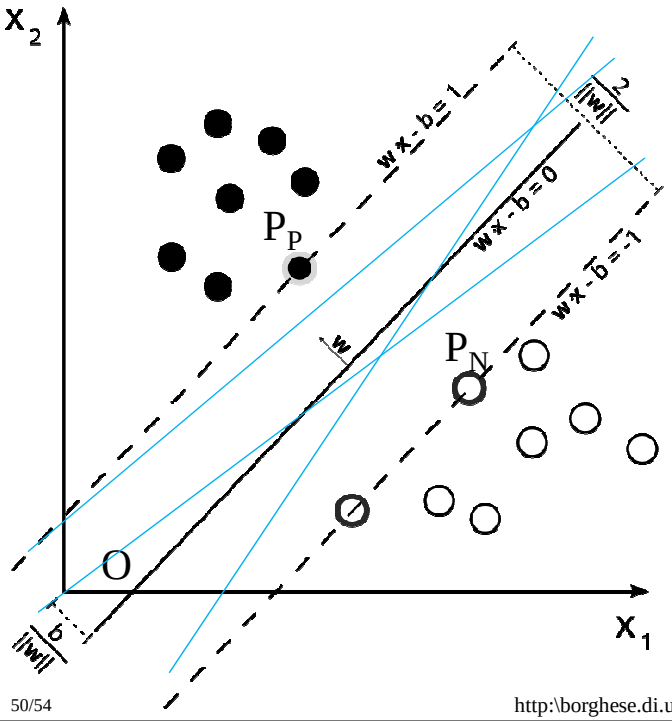
We have to compute:

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

with:

$$y_i(wx_i - b) - 1 \geq 0$$

$$\alpha_i \geq 0$$



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:

$y_i(wx_i-b) - 1 \geq 0$

$\alpha_i \geq 0$

Karush-Kuhn-Tucker conditions applied:

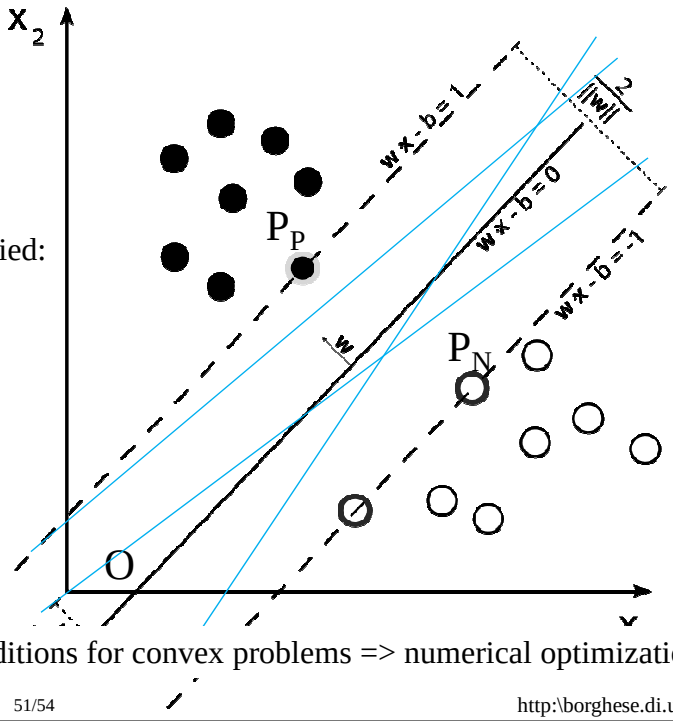
$$\frac{\partial L(.)}{dw} = w - \sum_i \alpha_i y_i x_i = 0$$

$$\frac{\partial L(.)}{db} = -\sum_i \alpha_i y_i = 0$$

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

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

$$\alpha_i \geq 0$$



KKT are necessary and sufficient conditions for convex problems => numerical optimization.

Support vector machines: dual formulation

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

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

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

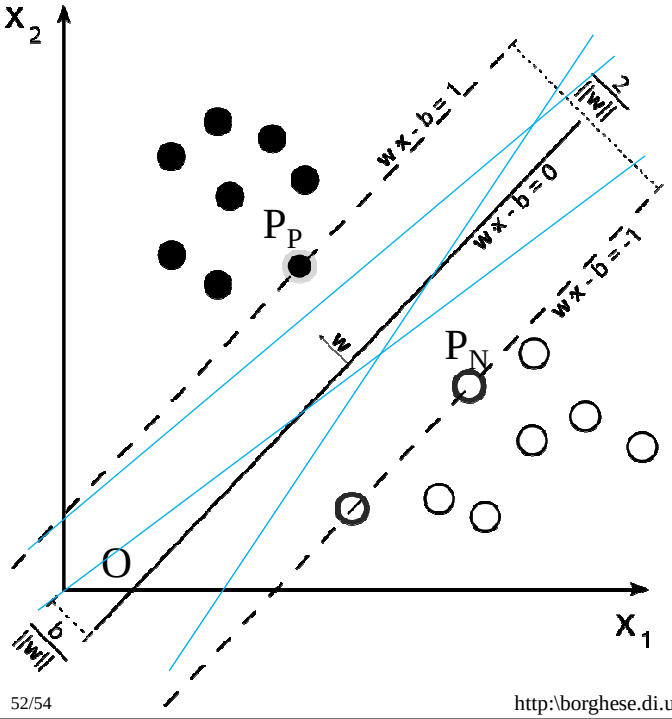
We solve the dual problem:

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

with constraints:

$$\frac{\partial L}{dw} = 0 \quad \frac{\partial L}{db} = 0$$

$$\alpha_i \geq 0$$



Dual formulation contains an inner product

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

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

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

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

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

Only the inner product of the data appears here.

Riassunto

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