UCB - CS189
Introduction to Machine Learning
Fall 2015
Lecture 7: Ridge regression

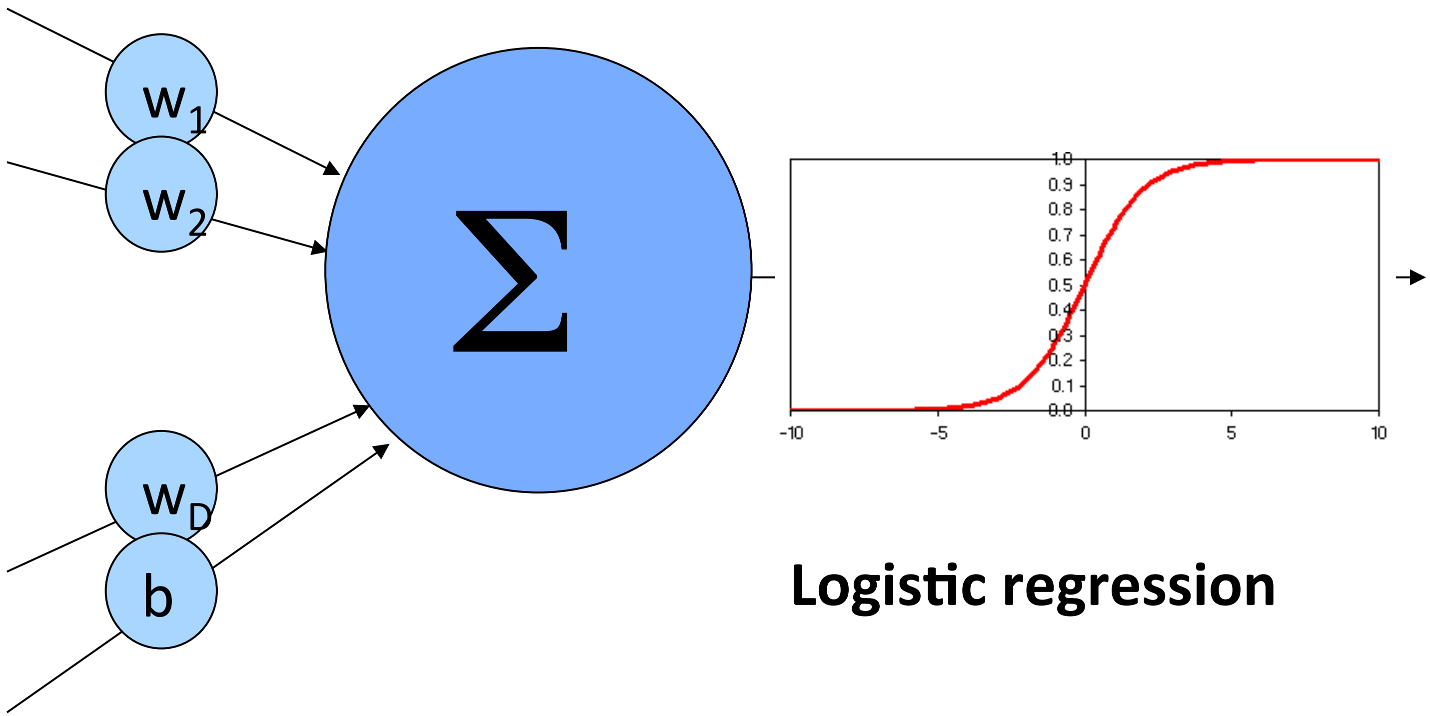
Isabelle Guyon

ChaLearn

Come to my office hours…

Wed 2:30-4:30 Soda 329

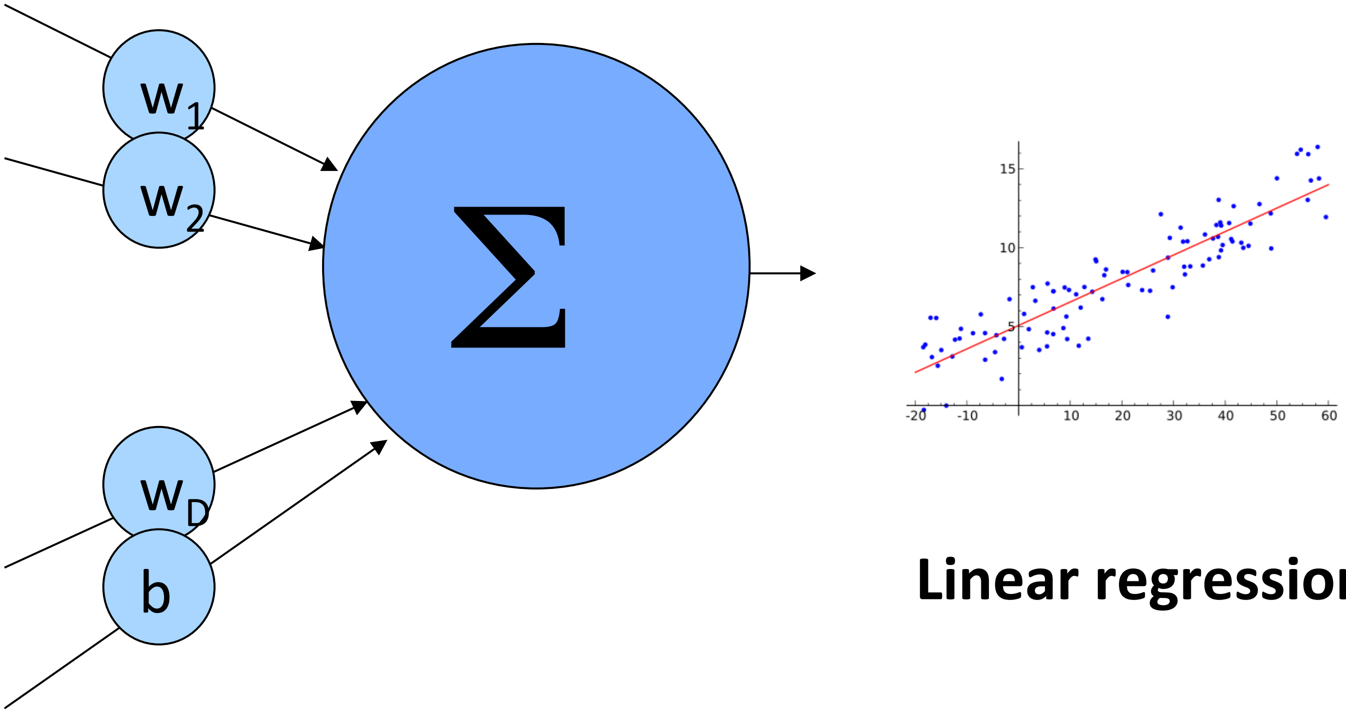
Last time



Logistic regression

Come to my office hours…
Wed 2:30-4:30 Soda 329

Today

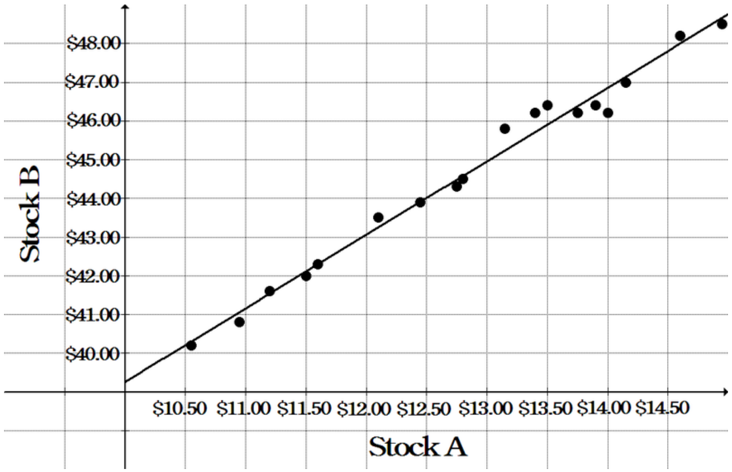


Linear regression

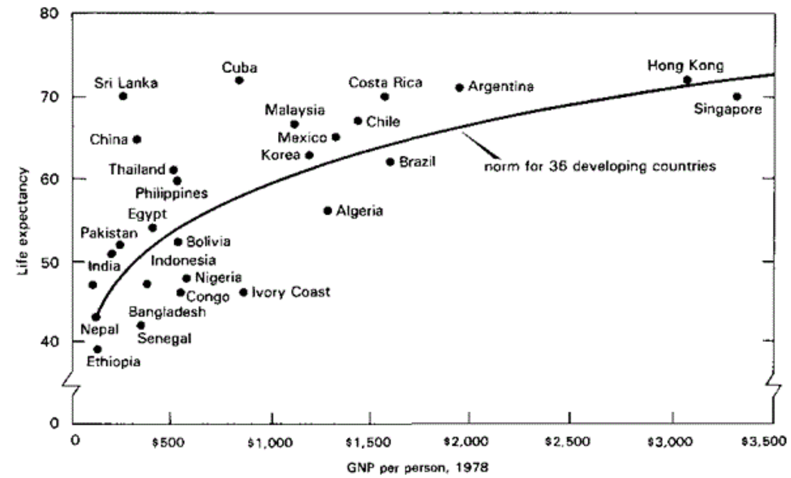
Math prerequisites

- Vectors and matrices
- Matrix multiplication
- Matrix inverse, determinant
- Matrix diagonalization, eigen vectors, eigen values, rank of a matrix
- Pseudo-inverse

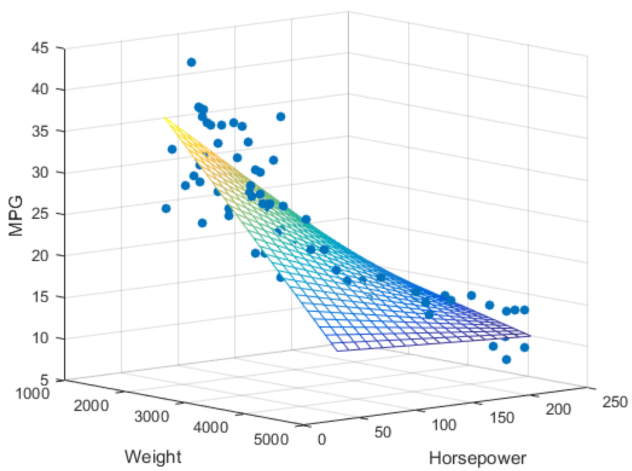
Uses of regression



Economy and Finance



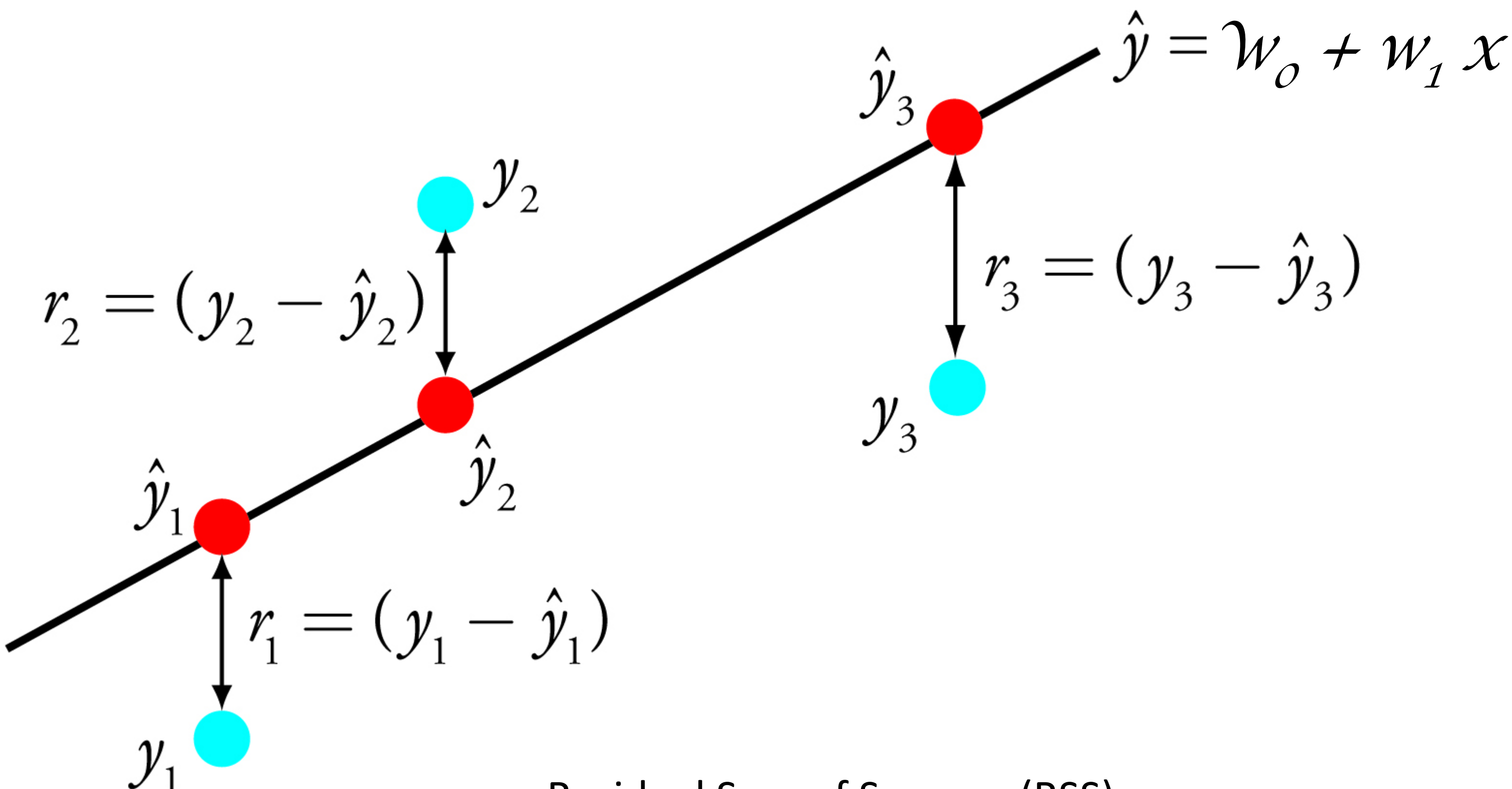
Epidemiology and medicine



Engineering

Apparent age estimation: http://gesture.chalearn.org

Least square regression



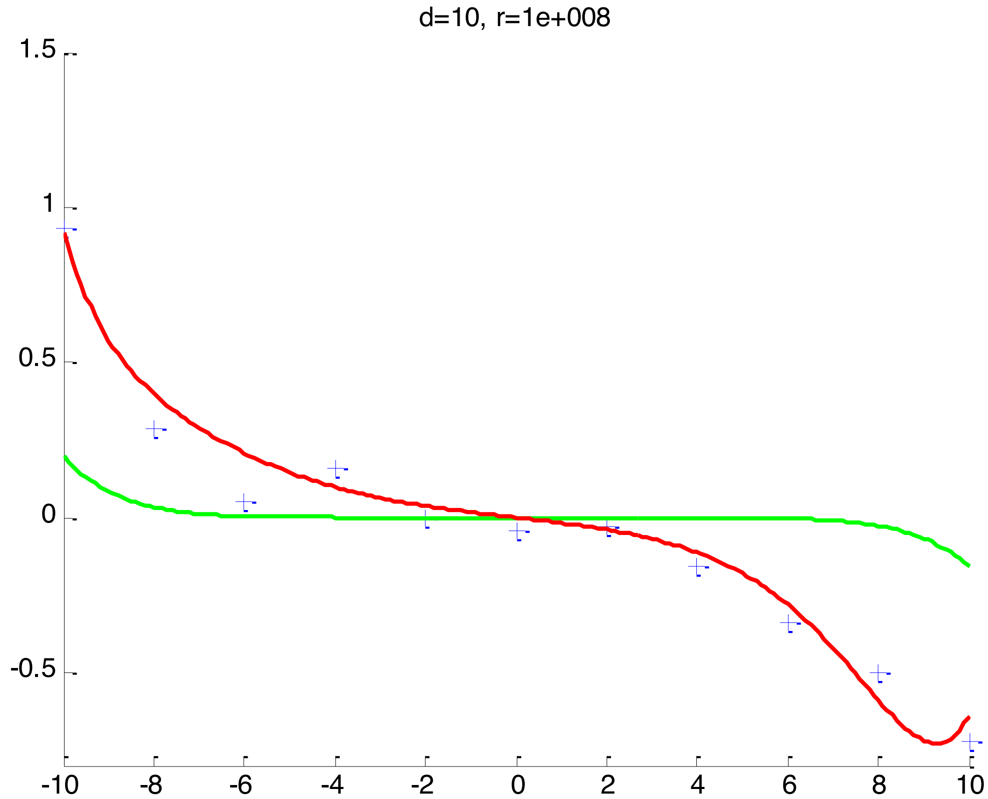
Picture: http://chemwiki.ucdavis.edu/

Residual Sum of Squares (RSS):

$$R[\mathbf{w}] = \Sigma_k (w_o + w_1 x_k - y_k)^2 = \Sigma_k (\mathbf{w}.\mathbf{x}_k - y_k)^2$$

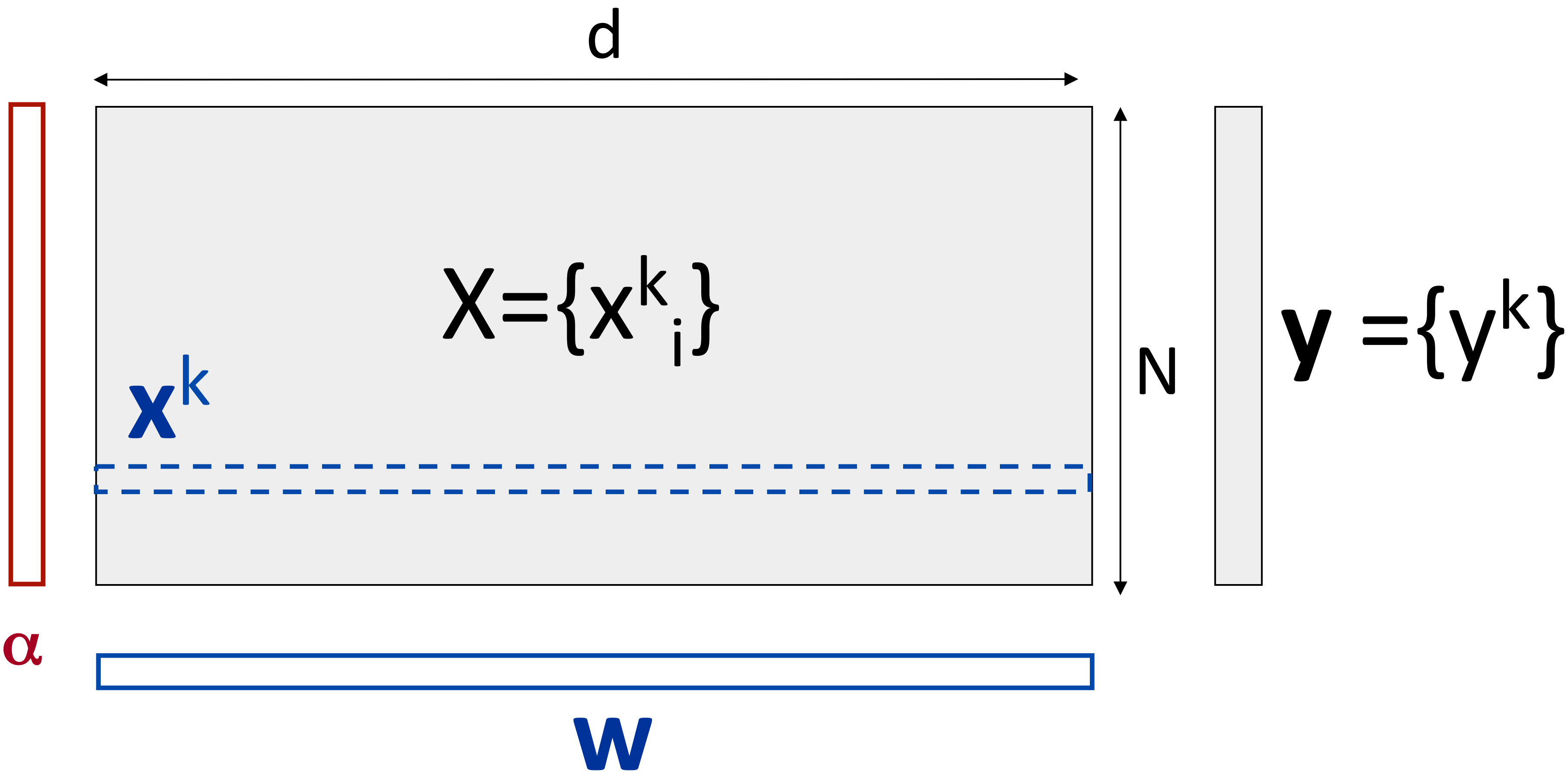
Polynomial Ridge Regression

d=polynomial degree
r=ridge (λ)



$R[\mathbf{w}] = \sum_k (\mathbf{w}.\Phi(\mathbf{x}_k) - y_k)^2 + \lambda \| \mathbf{w} \|^2$ $\quad\quad$ $\Phi(\mathbf{x})=[1, x, x^2, x^3...., x^d]$

Conventions



Matrix Notations

$w_i = \sum_k y^k x^k_i$ $\qquad$ $\mathbf{w} = \mathbf{y}^T X$ $\qquad$ $\mathbf{w}^T = X^T \mathbf{y}$

$w_i = \sum_k \alpha^k x^k_i$ $\qquad$ $\mathbf{w} = \alpha^T X$ $\qquad$ $\mathbf{w}^T = X^T \alpha$

(1,d) = (1,N)(N,d) $\qquad$ **(d,1) = (d,N)(N,1)**

$f(\mathbf{x}) = \sum_i w_i x_i$ $\qquad$ $f(\mathbf{x}) = \mathbf{x}\,\mathbf{w}^T = \mathbf{w}\,\mathbf{x}^T$

(1,d)(d,1) $\qquad$ **(1,d)(d,1)**

Linear Regression

- What we want:

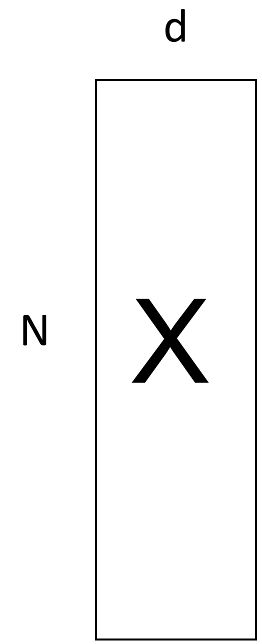
$\sum_i w_i \, x^k_i \;\; = y^k$ for all examples k=1…m (b=w_0)

or for classification, $y^k=\pm 1$, $\mathrm{sign}(\sum_i w_i \, x^k_i) = y^k$

- Solve: $\boxed{X\mathbf{w}^T = \mathbf{y}}$

(N,d)(d,1)=(N,1)

Regression: N>d



- Solve:

$$X \quad \mathbf{w}^T = \mathbf{y}$$

(N,d)(d,1) = (N,1)

- Normal equations

$$X^T X \, \mathbf{w}^T = X^T \mathbf{y}$$

(d,N)(N,d)(d,1) = (d,N)(N,1)

rank(X)≤min(d,N)
assume rank(X)=d
implies rank($X^T X$)=d
$X^T X$ is invertible

- Solution:

$$\mathbf{w}^T = (X^T X)^{-1} X^T \mathbf{y}$$

Pseudo-Inverse

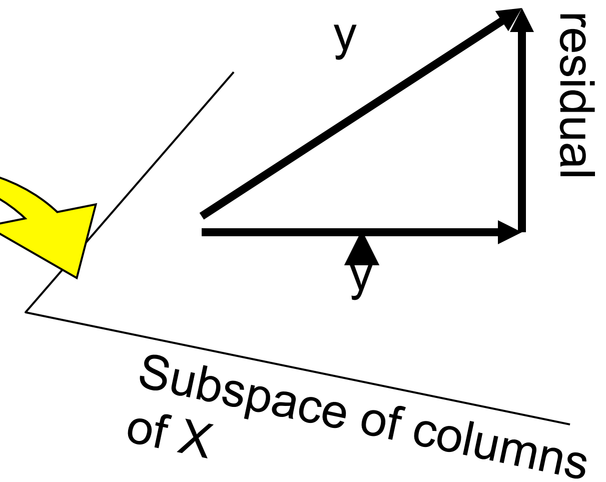
- Solution:

$$\underset{(d,1)}{\mathbf{w}^T} = \underset{(d,N)(N,d)}{(X^T X)^{-1}}\ \underset{(d,N)(N,1)}{X^T \mathbf{y}} \qquad \underset{(d,N)}{X^+} \text{ pseudo-inverse},\ X^+X=I$$

- Predictor:

$$\underset{(1,1)}{f(\mathbf{x})} = \underset{(1,d)(d,1)}{\mathbf{x}\,\mathbf{w}^T} = \underset{(1,d)(d,N)(N,1)}{\mathbf{x}\ X^+\ \mathbf{y}}$$

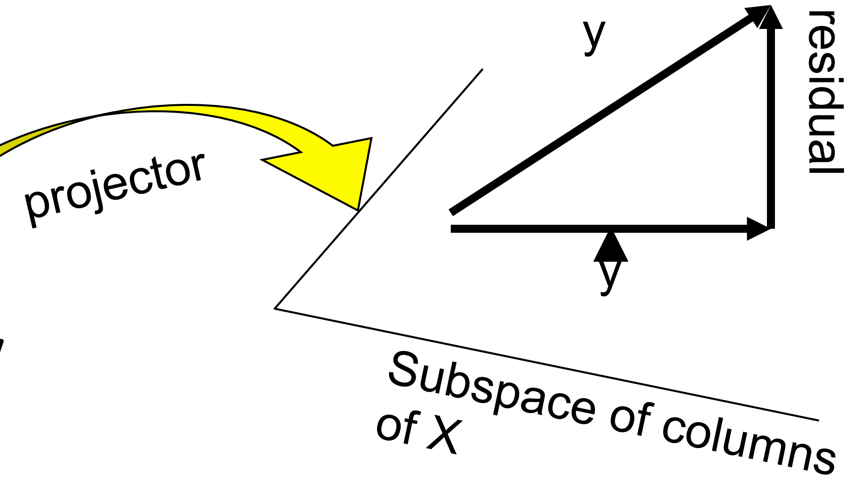
projector



- Residual:

$$\mathbf{y}-\hat{\mathbf{y}} = \mathbf{y} - X\,\mathbf{w}^T = (I-XX^+)\mathbf{y}$$

Least-Squares



$$\mathbf{y}-\hat{\mathbf{y}} = \mathbf{y} - X\,\mathbf{w}^T = (I-XX^+)\mathbf{y}$$

The pseudo-inverse solution is optimal in the least-square sense:

$$\min_{\mathbf{w}} \| \mathbf{y} - X\,\mathbf{w}^T \|^2 = \| (I-XX^+)\mathbf{y} \|^2$$

Gradient Descent

- Square loss:

$$L_k = (\mathbf{x}^k \, \mathbf{w}^T - y^k)^2$$

Risk = Residual Sum of Squares (RSS):

$$R = \Sigma_k \, (\mathbf{x}^k \, \mathbf{w}^T - y^k)^2$$
$$= \| X\mathbf{w}^T - \mathbf{y} \|^2$$
$$= \mathbf{w}X^TX\mathbf{w}^T - 2\mathbf{w}X^T\mathbf{y} + \mathbf{y}^T\mathbf{y}$$

- Gradient:

$$\nabla_{\mathbf{w}} R = 2 \, (X^TX\mathbf{w}^T - X^T\mathbf{y})$$

Normal Equations

- At the optimum:

$$\nabla_{\mathbf{w}} R = \mathbf{0}$$

$$2\,(X^T X \mathbf{w}^T - X^T \mathbf{y}) = \mathbf{0}$$

- Normal equations (again):

$$X^T X \mathbf{w}^T = X^T \mathbf{y}$$

Solve by inverting $X^T X$, if regular.

- What if $X^T X$, is singular?

Regularization

- Normal equations:
- Normal equations:

$$\underset{(d,N)}{X^T}\,\underset{(N,d)}{X}\,\underset{(d,1)}{\mathbf{w}} = \underset{(d,N)}{X^T}\,\underset{(N,1)}{\mathbf{y}}$$

- $X^T X$ may be singular

Regularized inverse

- Solution: $\underset{(d,1)}{\mathbf{w}} = (X^T X + \underset{(d,d)}{\lambda I})^{-1}\,\underset{(d,N)}{X^T}\,\mathbf{y}$ $\quad \lambda>0$

Why it works

- Diagonalization:

$$X^T X = U\,D\,U^T$$

- Diagonalization:

U orthogonal matrix of eigenvectors ($UU^T=I$)

D diagonal matrix of eigenvalues

Singularity: some eigenvalues are zero.

$$X^T X = U\,D\,U^T$$

U orthogonal matrix of eigenvectors ($UU^T=I$)

no more zero eigenvalue.

D diagonal matrix of eigenvalues

Singularity: some eigenvalues are zero.

Penalized Risk

$$\Sigma_k (\mathbf{x}^k \mathbf{w}^T - y^k)^2 = \| X\mathbf{w}^T - \mathbf{y} \|^2$$

- Add "regularizer":

$$R = \| X\mathbf{w}^T - \mathbf{y} \|^2 + \lambda \| \mathbf{w} \|^2 \qquad \lambda>0$$

- Gradient:

$$\nabla_{\mathbf{w}} R = 2\left((X^T X + \lambda I)\, \mathbf{w}^T - X^T \mathbf{y}\right)$$

Mechanical Interpretation

- Quadratic form:
- Quadratic form:

$$R = \| Xw - \mathbf{y} \|^2 + \lambda \| \mathbf{w} \|^2$$

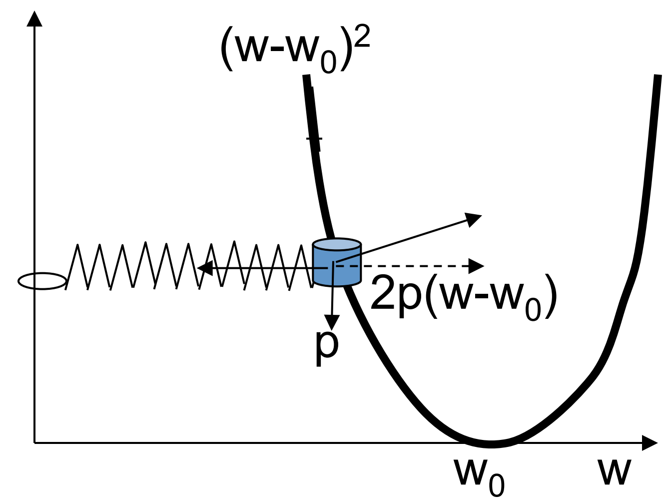
$X\mathbf{w}^T$

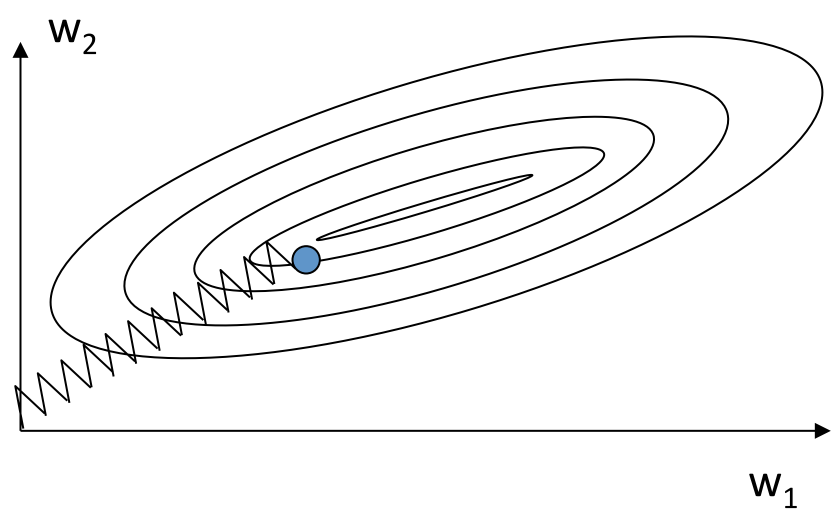
- One dimension:

$$R = p\,(w-w_0)^2 + \lambda\, w^2$$

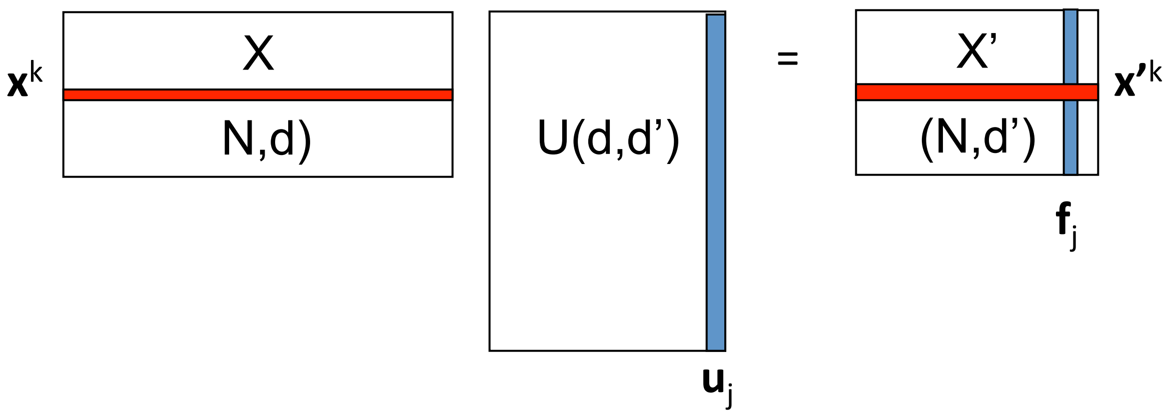
- Two dimensions:

$$-\mathbf{y} \|^2 + \lambda \| \mathbf{w} \|^2$$





Principal Component Analysis

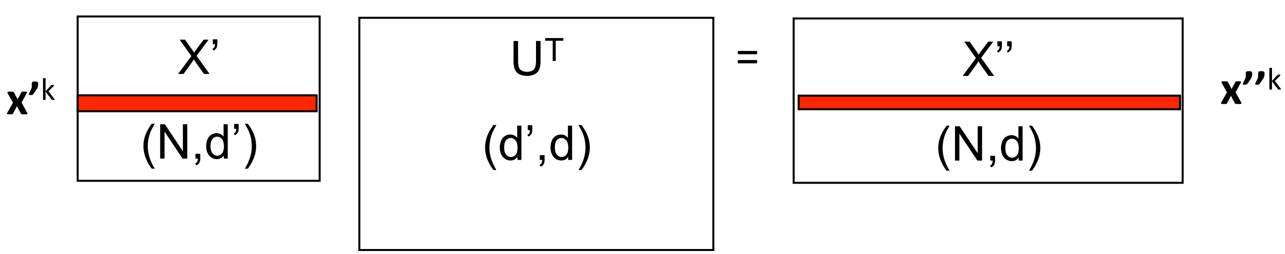


- Problem: Construct features that are linear combinations of the original features, such that the reconstructed patterns are as close as possible to the original in the least square sense.

- $\mathbf{f}_j = X\,\mathbf{u}_j$ linear combinations of columns of X

Problem: Construct features that are linear combinations of the

- $\mathbf{x}''^k = \mathbf{x}'^k U^T = \Sigma_j x'^k_j \mathbf{u}_j$



PCA Solution

- $X' = X U$
- $X'' = X' U^T$
- $X' = X U$
- $X'' = X' U^T$
- $\min_U \| X - X UU^T \|^2$
- $X'' = X UU^T$
- Can be brought back to solving and eigenvalue

- Compare:

  Regularization $X^TX + \lambda I = U(D+ \lambda I)U^T$

  PCA: Remove the dimensions with smallest

Kernel "Trick" (N<d)

- Solve: $X\mathbf{w}^T = \mathbf{y}$
- Assume: $\mathbf{w} = \Sigma_K \alpha^k \mathbf{x}^k = \alpha^T X$

  (1,d) (1,N) (N,d)

  $\alpha_K \mathbf{x}_K$

- Solve instead: $X X^T \alpha = \mathbf{y}$

  ( )(d,N)(N,1)=(N,1)

  Full rank (N,N) matrix

- Solution: $\alpha = (X X^T)^{-1} \mathbf{y}$

  $\mathbf{w}^T = X^T (X X^T)^{-1} \mathbf{y}$

  X^+

Kernel Ridge Regression

- $\Xi = \Phi(X)$
- Solve: $\xi = \Phi(x)$

$w = \Sigma$!

$\Xi \mathbf{w}^T = \mathbf{y}$

$= \mathbf{y}\alpha^T$

α_i

$\xi_i =$

α^T

(N,N) kernel matrix K Ξ

(1,D) (1,N) (N,D)

(N,D)(D,N)(N,1)=(N,1)

- Regularization: replace $\mathbf{K}$ by
- Solve: $\Xi\,\Xi\,\alpha = \mathbf{y}$

Regularization and PI

- Case N>d and rank(X^TX)=d

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

- Case N<d and rank(X^TX)=N
- Case N>d and rank(X^TX)=d

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

- Either case: $X^+ = (X^T X)^{-1} X^T$
- Case N<d and rank(X^TX)=N $X^+ = \lim_{\lambda \to 0} (X^T X + \lambda I)^{-1} X^T$ $X^+ = X^T (X X^T)^{-1}$
- Either case: $X^+ = \lim_{\lambda \to 0} (X^T X + \lambda I)^{-1}$

λI

Summary

- Least square regression for models linear in their parameters can be achieved with:
  - Stochastic gradient (suited for big data, see next lesson)
  - Pseudo-inverse (requires matrix inversion):
    - If $d<N$, invert X^TX, a (d,d) matrix.
    - If $N<d$, invert $X X^T$, a (N,N) matrix.
- Kernelization is easy $\Xi = \Phi(X)$
  - Invert $\Xi \Xi^T$, the kernel matrix K, a (N,N) matrix.
- Shrinkage for robustness:
  - Add a “ridge” = a small positive value to the diagonal.

Come to my office hours…
Wed 2:30-4:30 Soda 329

Next time

Kernel machines

PARAMETRIC (Perceptons)

$f(\mathbf{x}) = \mathbf{w} \bullet \Phi(\mathbf{x})$

$\mathbf{w} = \sum_k \alpha_k \Phi(\mathbf{x}^k)$

(Large margin) Perceptron

$\Delta\mathbf{w} \sim y_k \Phi(\mathbf{x}^k)$ if $y_k f(\mathbf{x}^k)<1$

$\sim \mathbf{1}(1-z_k)\, y_k \Phi(\mathbf{x}^k)$ $z_k = y_k f(\mathbf{x}^k)$

(Rosenblatt 1958)

Logistic regression

$\Delta\mathbf{w} \sim S(-z_k)\, y_k \Phi(\mathbf{x}^k)$

(Cox 1958)

LMS regression or classification

$\Delta\mathbf{w} \sim (y_k - f(\mathbf{x}^k))\, \Phi(\mathbf{x}^k) \sim (1- z_k)\, y_k \Phi(\mathbf{x}^k)$

(Widrow-Hoff, 1960)

NON PARAMETRIC (Kernel machines)

$f(\mathbf{x}) = \sum_k \alpha_k k(\mathbf{x}^k, x)$

$k(\mathbf{x}^k, \mathbf{x}) = \Phi(\mathbf{x}^k).\Phi(\mathbf{x})$

Potential Function algorithm

$\Delta\alpha_k \sim y_k$ if $y_k f(\mathbf{x}^k)<1$

$\sim \mathbf{1}(1-z_k)\, y_k$

(Aizerman et al 1964)

Dual logistic regression

$\Delta\alpha_k \sim S(-z_k)\, y_k$

Dual LMS

$\Delta\alpha_k \sim (y_k - f(\mathbf{x}^k)) \sim (1- z_k)\, y_k$

4