Lecture 3: Shrinkage

Isabelle Guyon
guyoni@inf.ethz.ch

References

Structural risk minimization for character recognition
Isabelle Guyon et al.
http://clopinet.com/isabelle/Papers/srm.ps.Z

Kernel Ridge Regression
Isabelle Guyon
http://clopinet.com/isabelle/Projects/ETH/KernelRidge.pdf

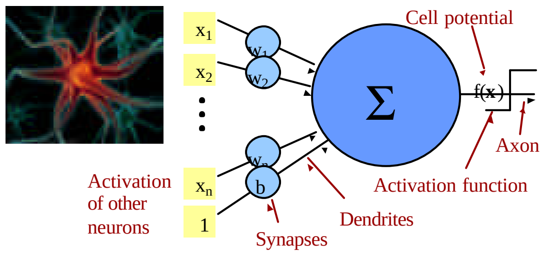
Ockham's Razor

- Principle proposed by William of Ockham in the fourteenth century: "**Pluralitas non est ponenda sine neccesitate**".
- Of two theories providing similarly good predictions, prefer the simplest one.
- Shave off unnecessary parameters of your models.

The Power of Amnesia

- The human **brain** is made out of billions of cells or Neurons, which are highly interconnected by synapses.
- Exposure to enriched environments with extra sensory and social stimulation enhances the **connectivity** of the synapses, but children and adolescents can lose them up to 20 million per day.

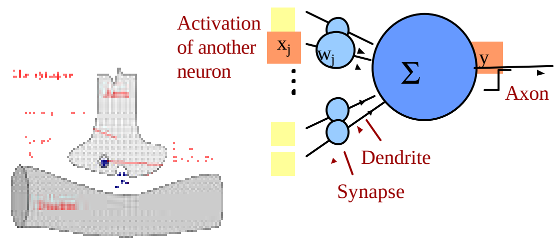
Artificial Neurons



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

Hebb's Rule

$$w_j \leftarrow w_j + y_i\, x_{ij}$$



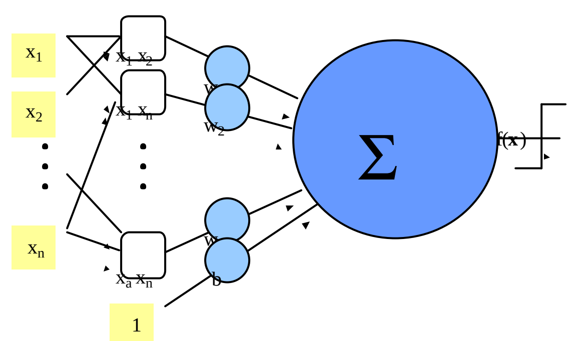
Weight Decay

$w_j \leftarrow w_j + y_i\, x_{ij}$ Hebb's rule

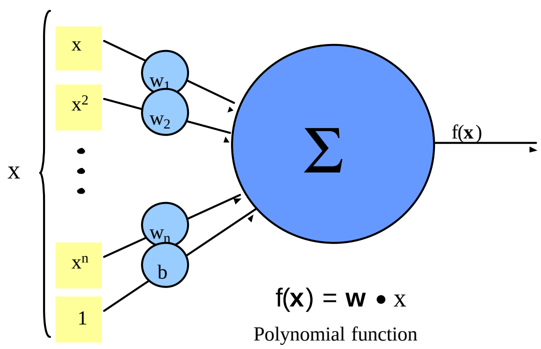
$\boxed{w_j \leftarrow (1-\lambda)\, w_j + y_i\, x_{ij}}$ Weigh decay

$\lambda \in [0, 1]$, decay parameter

Sigma-Pi Unit



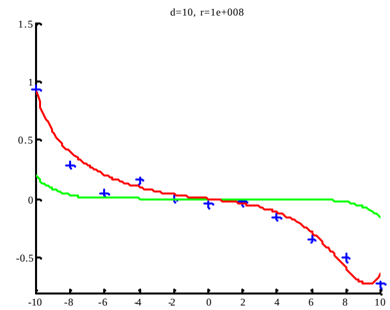
One Dimensional Example



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

Polynomial function

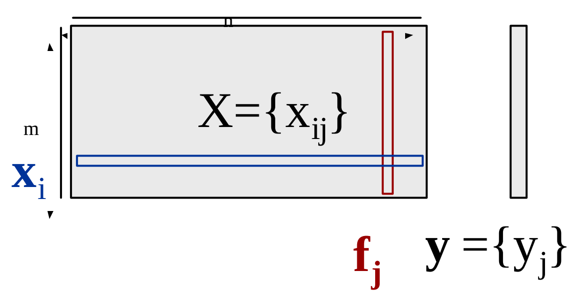
Polynomial Regression



Conventions

- $X = \{x_{ij}\}$ training data matrix, i=1:m, j=1:n
  $\mathbf{x}_i = \{x_j\}$ matrix line, training pattern i
  $\mathbf{x}$ test pattern, dim n
- y_i target value of pattern i
  y target value of test pattern
- $\mathbf{w}$ weight vector, dim n
- a weight vector, dim m

Conventions



Matrix Notations

$w_j = \sum_i y_i x_{ij}$ $\quad$ $\mathbf{w} = \mathbf{y}^T X$ $\quad$ $\mathbf{w}^T = X^T \mathbf{y}$

$(1,n)=(1,m)(m,n)$ $\quad$ $(n,1)=(n,m)(m,1)$

$f(\mathbf{x}) = \sum_j w_j x_j$ $\quad$ $f(\mathbf{x}) = \mathbf{x}\,\mathbf{w}^T = \mathbf{w}\,\mathbf{x}^T$

Linear Regression

- What we want:

  $\sum_j w_j x_{ij} = y_i$ for all examples i=1…m $\quad (b=w_0)$

  or for classification, $y_i = \pm 1$, $\mathrm{sign}(\sum_j w_j x_{ij}) = y_i$

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

  $(m,n)(n,1)=(m,1)$

Regression: m>n

- Solve:

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

  $(m,n)(n,1) = (m,1)$

- Normal equations

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

  $(n,m)(m,n)(n,1) = (n,m)(m,1)$

- Solution:

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

n

m X

rank(X)≤min(n,m)
assume rank(X)=n
implies rank($X^T X$)=n
$X^T X$ is invertible

Pseudo-Inverse

- Solution:

  $\mathbf{w}^T = (X^T X)^{-1} X^T \ \mathbf{y}$ $\quad$ X^+ pseudo-inverse, $X^+ X = I$

  $(n,1)$ $\quad (n,m)(m,n)(n,m)(m,1)$ $\quad (n,m)$

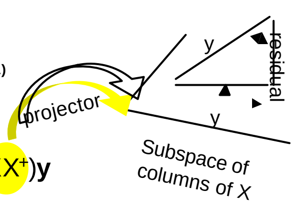
- Predictor:

  $f(\mathbf{x}) = \mathbf{x}\ \mathbf{w}^T = \mathbf{x}\ \ X^+ \mathbf{y}$

  $(1,1)$ $\quad (1,n)(n,1)$ $\quad (1,n)(n,n)(n,1)$

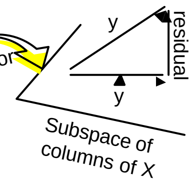
- Residual:

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



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:

  $L_i = (\mathbf{x}_i\, \mathbf{w}^T - y_i)^2$
- Sum of squares:

  $R = \Sigma_i\, (\mathbf{x}_i\, \mathbf{w}^T - y_i)^2$

  $= \|\, X\mathbf{w}^T - \mathbf{y} \,\|^2$

  $= \mathbf{w}X^TX\mathbf{w}^T - 2\mathbf{w}X^T\mathbf{y} + \mathbf{y}^T\mathbf{y}$
- Gradient:

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

Normal Equations

- At the optimum:

  $\tilde{N}_{\mathbf{w}}R = \mathbf{0}$

  $2\,(X^TX\mathbf{w}^T - X^T\mathbf{y}) = \mathbf{0}$
- Normal equations (again):

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

  Solve by inverting X^TX, if regular.

- What if X^TX, is singular?

Regularization

- Normal equations:

  $\underset{(n,m)}{X^T}\ \underset{(m,n)}{X}\ \underset{(n,1)}{\mathbf{w}^T} = \underset{(n,m)}{X^t}\ \underset{(m,1)}{\mathbf{y}}$
- Case m<n (interpolation), rank(X)≤m<n, matrix X^TX singular.
- Replace X^TX by $(X^TX+\lambda I)$ $\lambda>0$
- Solution:

  $\underset{(n,1)}{\mathbf{w}^T} = \underset{(n,m)}{(X^T}\ \underset{(m,n)}{X}\ \underset{(n,n)}{+\lambda I)^{-1}}\ \underset{(n,m)}{X^T}\ \underset{(m,1)}{\mathbf{y}}$

  Regularized inverse

Why it works

- Diagonalization:

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

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

  D diagonal matrix of eigenvalues

  Singularity: some eigenvalues are zero.
- Regularization:

  $X^T X + \lambda I = U\,(D + \lambda I)\,U^T \qquad \lambda > 0$

  no more zero eigenvalue.

Penalized Risk

- Sum of squares:

  $R = \Sigma_i\,(\mathbf{x}_i\,\mathbf{w}^T - y_i)^2$

  $= \|\,X\mathbf{w}^T - \mathbf{y}\,\|^2$
- Add "regularizer":

  $R = \|\,X\mathbf{w}^T - \mathbf{y}\,\|^2 + \lambda \|\,\mathbf{w}\,\|^2$
- Gradient:

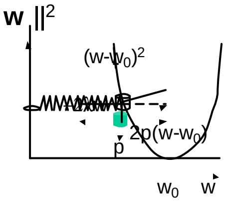
  $\nabla_{\mathbf{w}} R = 2\,((X^T X + \lambda\,I)\,\mathbf{w}^T - X^T \mathbf{y})$

Mechanical Interpretation

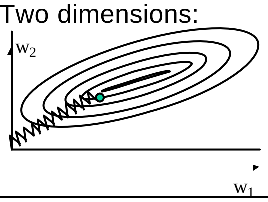
- Quadratic form:

  $R = \|\,X\mathbf{w}^T - \mathbf{y}\,\|^2 + \lambda \|\,\mathbf{w}\,\|^2$
- One dimension:

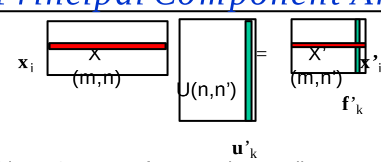
  $R = p\,(w - w_0)^2 + \lambda\,w^2$



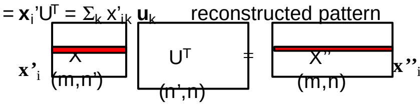
- Two dimensions:



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.
- $\mathbf{f}'_k = X\,\mathbf{u}_k$ linear combinations of columns of X
- $\mathbf{x}_i'' = \mathbf{x}_i' U^T = \Sigma_k\, x'_{ik}\,\mathbf{u}_k$ reconstructed pattern



PCA Solution

- $X' = X U$
- $X'' = X' U^T$
- $X'' = X UU^T$
- $\min_U \| X - X UU^T \|^2$
- Can be brought back to solving and eigenvalue problem: $X^TX = UDU^T$ i.e. $X'^TX'=D$
- Compare:
  Regularization $X^TX + \lambda I = U(D + \lambda I)U^T$
  PCA: Remove the dimensions with smallest eigenvalues.

Kernel "Trick" (m<n)

- Solve: $X\mathbf{w}^T = \mathbf{y}$
- Assume: $\mathbf{w} = \Sigma_i \alpha_i \mathbf{x}_i = a^T X$

  $(1,n) \quad (1,m) \quad (m,n)$
- Solve instead: $X X^T a = \mathbf{y}$

  $(m,n)(n,m)(m,1)=(m,1)$

  Full rank (m,m) matrix
- Solution: $a = (X X^T)^{-1} \mathbf{y}$

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

Kernel Ridge Regression

- $\Xi = \Phi(X)$
- Solve: $\Xi\mathbf{w}^T = \mathbf{y}$
- Assume: $\mathbf{w} = \Sigma_i \alpha_i x_i = a^T \Xi$

  $(1,N) \quad (1,m) \quad (m,N)$
- Solve instead: $\Xi \Xi^T a = \mathbf{y}$

  $(m,n)(n,m)(m,1)=(m,1)$

  (m,m) kernel matrix K
- Solution: $a = K^{-1} \mathbf{y}$
- Regularization: replace K by $K+\lambda I$

Regularization and PI

- Case m>n and $\mathrm{rank}(X^TX)=n$

  $X^+ = (X^T X)^{-1} X^T$
- Case m<n and $\mathrm{rank}(X^TX)=m$

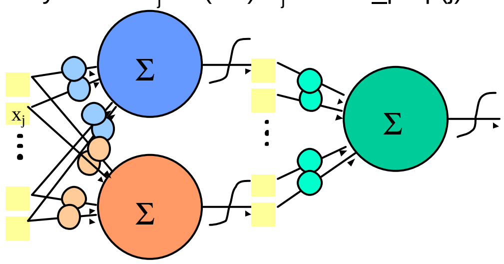
  $X^+ = X^T(X X^T)^{-1}$
- Either case:

  $X^+ = \lim_{\lambda\to 0} (X^TX + \lambda I)\ X^T$

  $= \lim_{\lambda\to 0} X^T(X X^T + \lambda I)^{-1}$

Weight Decay for MLP

Replace: $w_j \leftarrow w_j$ + back_prop(j)
by: $w_j \leftarrow (1-\lambda)\, w_j$ + back_prop(j)



Priors and Bayesian Learning

- Double random process:
  - Draw a target function f in a family of functions {f}
  - Draw the data pairs ($\mathbf{x}_i$, $y_i = f(\mathbf{x}_i)$+noise)
- The distribution of f is called the "prior" P(f).
- Our revised opinion about f once we see the data is the "posterior" P(f|D).
- Bayesian "learning":

  $P(y|x,D) \;\alpha\; \int P(y|x,D,f)\, dP(f|D)$
- MAP:

  $f = \text{argmax } P(f|D)$

  $= \text{argmax } P(D|f)\, P(f)$

MAP = RRM

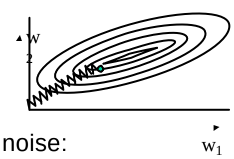
- Maximum A Posteriori (MAP):

  $f = \text{argmax } P(D|f)\, P(f)$

  $= \text{argmin } \underbrace{-\log P(D|f)}_{\substack{\text{Negative log likelihood} \\ = \text{Empirical risk } R[f]}} \; \underbrace{-\log P(f)}_{\substack{\text{Negative log prior} \\ = \text{Regularizer } \Omega[f]}}$
- Regularized Risk Minimization (RRM):

  $f = \text{argmin } R[f] + \Omega[f]$

Example: Gaussian Prior



- Linear model:

  $f(\mathbf{x}) = \mathbf{x}\, \mathbf{w}^T$
- Square loss ⇔ Gaussian noise:

  $P(D|f) = \exp -||X\mathbf{w}^T - \mathbf{y}||^2 / \sigma^2$

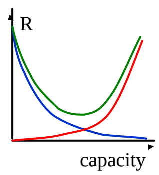
  $R[f] = -\log P(D|f) \sim ||X\mathbf{w}^T - \mathbf{y}||^2$
- Weight decay ⇔ Gaussian prior:

  $P(f) = \exp -\lambda||\mathbf{w}||^2$

  $\Omega[f] = -\log P(f) = \lambda||\mathbf{w}||^2$

Structural Risk Minimization

- Nested subsets of models, increasing complexity/capacity:
  $S_1 \subset S_2 \subset \ldots S_N$
- Example, rank with $||\mathbf{w}||^2$
  $S_k = \{ w \mid ||\mathbf{w}||^2 < A_k \}, A_1 < A_2 < \ldots < A_k$
- Minimization under constraint:
  min $R_{emp}[f]$ s.t. $||\mathbf{w}||^2 < A_k$
- Lagrangian:
  $R_{reg}[f] = R_{emp}[f] + \lambda ||\mathbf{w}||^2$



Conclusion

- Weight decay is a means of avoiding "overfitting" that is justifiable from many perspectives:
  - Ockham's razor
  - Synaptic decay
  - Regularization
  - Gaussian prior on the weights
  - Structural risk minimization
- It works for linear models, kernel methods, and neural networks.
- It can be combined with various loss functions.

Practical Work

Homework 3:

1) Same data and software as homework 2.
2) Create a heatmap of the 100 top ranking features you selected.
3) Make a scatter plot of the 3 top ranking features you selected.
4) Email the result zip file with the figures to guyoni@inf.ethz.ch with subject "Homework3" no later than:
   Tuesday November 15th.

Risk Minimization

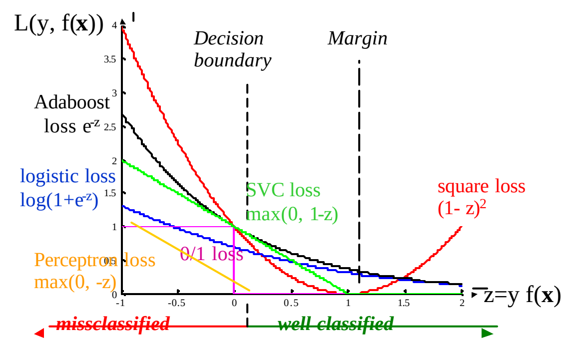
- **Learning problem:** find the best function f(**x**; a) minimizing the risk functional

$$R[f] = \int \underbrace{L(f(\mathbf{x};\ a),\ y)}_{\text{loss function}}\ d\underbrace{P(\mathbf{x},\ y)}_{\text{unknown distribution}}$$

- **Examples are given**:

$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots (\mathbf{x}_m, y_m)$

Loss Functions



Kernel "Trick"

- $f(\mathbf{x}) = \sum_i \alpha_i\, k(\mathbf{x}_i, \mathbf{x})$
- $k(\mathbf{x}_i, \mathbf{x}) = F(\mathbf{x}_i) \bullet F(\mathbf{x})$

Dual forms

- $f(\mathbf{x}) = \mathbf{w} \bullet F(\mathbf{x})$
- $\mathbf{w} = \sum_i \alpha_i F(\mathbf{x}_i)$

What is a Kernel?

A kernel is a dot product in *some* feature space: $k(\mathbf{s}, \mathbf{t}) = F(\mathbf{s}) \bullet F(\mathbf{t})$

- Examples:
- $k(\mathbf{s}, \mathbf{t}) = \exp(-\|\mathbf{s}-\mathbf{t}\|^2/\sigma^2)$ Gaussian kernel
- $k(\mathbf{s}, \mathbf{t}) = 1/\|\mathbf{s}-\mathbf{t}\|$ Potential function
- $k(\mathbf{s}, \mathbf{t}) = (\mathbf{s} \bullet \mathbf{t})^q$ Polynomial kernel

$$\underbrace{([s_1,s_2]\bullet[t_1,t_2])^2}_{k(\mathbf{s},\ \mathbf{t})} = \underbrace{[s_1^2,s_2^2,\sqrt{2}s_1s_2]}_{F(\mathbf{s})}\,.\underbrace{[t_1^2,t_2^2,\sqrt{2}t_1t_2]}_{F(\mathbf{t})}$$

Simple Kernel Methods

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

$\mathbf{w} = \sum_i \alpha_i \Phi(\mathbf{x}_i)$

Perceptron algorithm
$\mathbf{w} \leftarrow \mathbf{w} + y_i \Phi(\mathbf{x}_i)$ if $y_i f(\mathbf{x}_i) < 0$
(Rosenblatt 1958)

Minover (optimum margin)
$\mathbf{w} \leftarrow \mathbf{w} + y_i \Phi(\mathbf{x}_i)$ for min $y_i f(\mathbf{x}_i)$
(Krauth-Mézard 1987)

LMS regression
$\mathbf{w} \leftarrow \mathbf{w} + \eta (y_i - f(\mathbf{x}_i)) \Phi(\mathbf{x}_i)$

$f(\mathbf{x}) = \sum_i \alpha_i K(\mathbf{x}_i, \mathbf{x})$

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

Potential Function algorithm
$\alpha_i \leftarrow \alpha_i + y_i$ if $y_i f(\mathbf{x}_i) < 0$
(Aizerman et al 1964)

Dual minover
$\alpha_i \leftarrow \alpha_i + y_i$ for min $y_i f(\mathbf{x}_i)$
(ancestor of SVM 1992, similar to kernel Adatron, 1998, and SMO, 1999)

Dual LMS
$\alpha_i \leftarrow \alpha_i + \eta (y_i - f(\mathbf{x}_i))$

Exercise: Gradient Descent

- Linear discriminant $f(\mathbf{x}) = \Sigma_j w_j x_j$
- Functional margin $z = y f(\mathbf{x})$, $y = \pm 1$
- Compute $\partial z / \partial w_j$
- Derive the learning rules $\Delta w_j = -\eta\, \partial L / \partial w_j$ corresponding to the following loss functions:

square loss $(1-z)^2$

SVC loss $\max(0, 1-z)$

Adaboost loss e^{-z}

Perceptron loss $\max(0, -z)$

logistic loss $\log(1+e^{-z})$

Exercise: Dual Algorithms

- From the derive the Δw_j derive the $\Delta \mathbf{w}$
- $\mathbf{w} = \Sigma_i \alpha_i \mathbf{x}_i$
- From the $\Delta \mathbf{w}$, derive the $\Delta \alpha_i$ of the dual algorithms.

Exercise: Linear Algebra

- Prove that if X is of rank r, $X^T X$ and $X X^T$ have the same rank.
- Show that $X^T X$ and $X X^T$ have only positive eigenvalues.