Supervised and Unsupervised Learning

Machine Learning II (SS 2008, TU Berlin)

Prof. Dr. Klaus-Robert Müller
Dr. Alexander Zien

15. 04. 2008

MAX-PLANCK-GESELLSCHAFT

Taxonomy of Learning Tasks

input	output	
	discrete	**continuous**
supervised given $\{(\mathbf{x}_i, y_i)\}$	• classification	• regression
unsupervised given $\{\mathbf{x}_i\}$	• clustering • anomality detection	• dimensionality reduction

Outside this scheme:

- active learning
- reinforcement learning

Taxonomy of Learning Approaches

Models

- geometrical models
- statistical learning theory
- probabilistic models

Methods

- optimization
  - regularization
  - ML / MAP
- Bayesian inference

we'll see close connections...

Regression: Geometric View (1)

Minimize squared distance of predictions to labels:

- $J(\mathbf{w}) := \sum_{i=1}^{N} \left(y_i - \mathbf{x}_i^\top \mathbf{w} \right)^2 = \left(\mathbf{y} - \mathbf{X}^\top \mathbf{w} \right)^\top \left(\mathbf{y} - \mathbf{X}^\top \mathbf{w} \right)$
  where $\mathbf{X} \in \mathbb{R}^{D \times N}$ contains N points as columns
- $\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} J(\mathbf{w})$

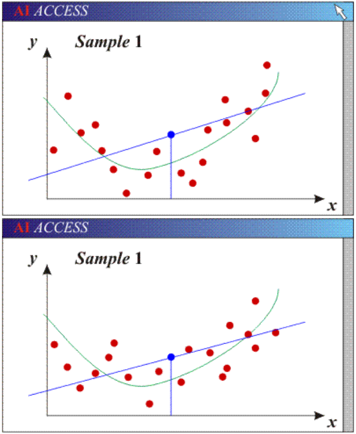
Convex $\Rightarrow$ vanishing derivative indicates global optimum $\widehat{\mathbf{w}}$.

$$0 = \left. \frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} \right|_{\mathbf{w} = \widehat{\mathbf{w}}} = \mathbf{X}\mathbf{X}^\top \widehat{\mathbf{w}} - \mathbf{X}\mathbf{y} \qquad \Rightarrow \qquad \widehat{\mathbf{w}} = \left(\mathbf{X}\mathbf{X}^\top \right)^{-1} \mathbf{X}\mathbf{y}$$

“Ordinary Least Squares”

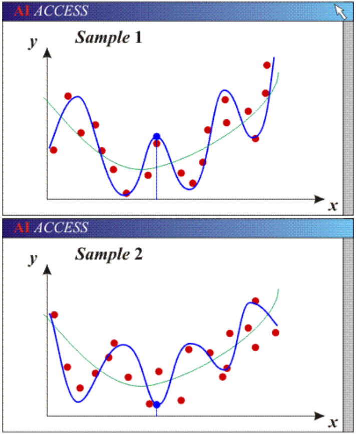
Regression: Bias-Variance Tradeoff (1)

simple model (linear)



high bias, low variance
"underfitting"

complex model



low bias, high variance
"overfitting"

images from `http://www.aiaccess.net/English/Glossaries/GlosMod/e_gm_bias_variance.htm`

Regression: Bias-Variance Tradeoff (2)

Bias-variance decomposition of approximation error:

$$\mathbb{E}\left[\|\mathbf{y}-\mathbf{X}\widehat{\mathbf{w}}\|^2\right] = \begin{cases} \mathbb{E}\left[\|\mathbf{y}-\mathbf{X}\mathbf{w}\|^2\right] & noise \\ +\mathbb{E}\left[\|\mathbf{X}\mathbf{w}-E[\mathbf{X}\widehat{\mathbf{w}}]\|^2\right] & bias^2 \\ +\mathbb{E}\left[\|\mathbb{E}\left[\mathbf{X}\widehat{\mathbf{w}}\right]-\mathbf{X}\widehat{\mathbf{w}}\|^2\right] & variance \end{cases}$$

Shrinkage

- increase bias...
- ... to reduce variance even more
- first proposed for the multivariate mean (James/Stein)

Regression: Geometric View (2)

Apply shrinkage to least squares ...

- shrink $\mathbf{w}$
- $\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \left\| \mathbf{y} - \mathbf{X}^\top \mathbf{w} \right\|^2 + \lambda \left\| \mathbf{w} \right\|^2$
- $\Rightarrow \quad \widehat{\mathbf{w}} = \left(\mathbf{X}\mathbf{X}^\top + \lambda \mathbf{I} \right)^{-1} \mathbf{X}\mathbf{y}$

"Regularized Least Squares"

- aka rigde regression
- aka Tikhonov regularization

Regression: Probabilistic View (1)

Multivariate Gaussian error model:

$$P(\mathbf{y}|\mathbf{X}, \mathbf{w}) \;=\; \mathcal{N}\left(\mathbf{y}\,\middle|\,\mathbf{X}^\top\mathbf{w}, \sigma^2\mathbf{I}\right)$$

Likelihood of parameter $\mathbf{w}$ given the data $(\mathbf{X}, \mathbf{y})$:

$$\mathcal{L}(\mathbf{w}) \;=\; (2\pi)^{-\frac{N}{2}} \left|\sigma^2\mathbf{I}\right|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\left(\mathbf{y} - \mathbf{X}^\top\mathbf{w}\right)^\top \left(\sigma^2\mathbf{I}\right)^{-1} \left(\mathbf{y} - \mathbf{X}^\top\mathbf{w}\right)\right]$$

(same as $P(\mathbf{y}|\mathbf{X}, \mathbf{w})$, just seen as function of $\mathbf{w}$)

Maximum Likelihood (ML) approach:

- $\widehat{\mathbf{w}} = \arg\max_{\mathbf{w}} \log \mathcal{L}(\mathbf{w}) = \arg\min_{\mathbf{w}} \left\|\mathbf{y} - \mathbf{X}^\top\mathbf{w}\right\|^2$
- same as OLS!

Regression: Probabilistic View (2)

Maximum A Posteriori (MAP) approach:

- add normal prior: $P(\mathbf{w}) = \mathcal{N}(\mathbf{0}, \lambda\mathbf{I})$
- maximize posterior $P(\mathbf{w}|\mathbf{X}, \mathbf{y}) \propto P(\mathbf{w})P(\mathbf{y}|\mathbf{X}, \mathbf{w})$ (Bayes theorem)
- $\Rightarrow \quad \widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \left\| \mathbf{y} - \mathbf{X}^\top \mathbf{w} \right\|^2 + \lambda \|\mathbf{w}\|^2$
- same as RLS!

Remarks:

1. nice way to incorporate prior knowledge: $P(\mathbf{w}) = \mathcal{N}(\mathbf{w}_0, \lambda\mathbf{I})$
2. point estimate — not Bayesian!

Regression: Probabilistic View (3)

Bayesian approach ("inference"):

- probabilities model your own uncertainty
- follow Bayes' rule:
  1. encode uncertainty in beliefs as prior distribution
  2. update beliefs according to data
  3. this yields posterior beliefs *as distribution*
- if you believe in 3 simple axioms (Cox/Jaynes), the only way!
- "loss" only considered afterwards ($\rightarrow$ decision theory)

For our regression setting:

- posterior $P(\mathbf{w}|\mathbf{X},\mathbf{y}) = \dfrac{P(\mathbf{y}|\mathbf{X},\mathbf{w})P(\mathbf{w})}{P(\mathbf{y}|\mathbf{X})} = \mathcal{N}(\cdot,\cdot)$
- predictive distribution:
  $P(\mathbf{y}^*|\mathbf{x}^*,\mathbf{X},\mathbf{y}) = \int P(\mathbf{y}^*|\mathbf{x}^*,\mathbf{w})P(\mathbf{w}|\mathbf{X},\mathbf{y})d\mathbf{w} = \mathcal{N}(\cdot,\cdot)$

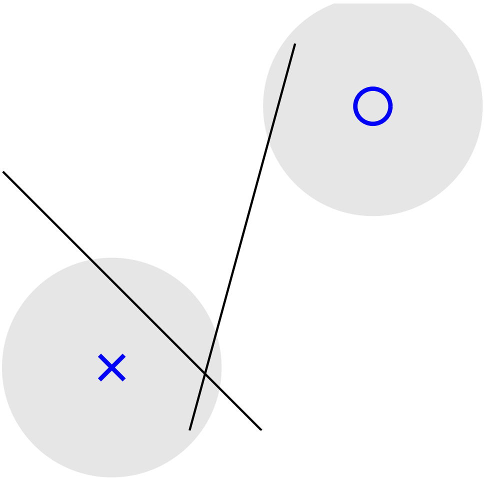
Classification

Generative Learning (Sampling Paradigm)

- **model** $P(y, \mathbf{x})$, often as $P(y)P(\mathbf{x}|y)$
- predict via Bayes theorem: $P(y|\mathbf{x}) = \dfrac{P(y)P(\mathbf{x}|y)}{\sum_{y'} P(y')P(\mathbf{x}|y')}$
- naive Bayes: assume $P(\mathbf{x}|y) = \prod_{i=1}^{D} P(x_{[i]}|y)$
- in general: prior knowledge explicitly built in

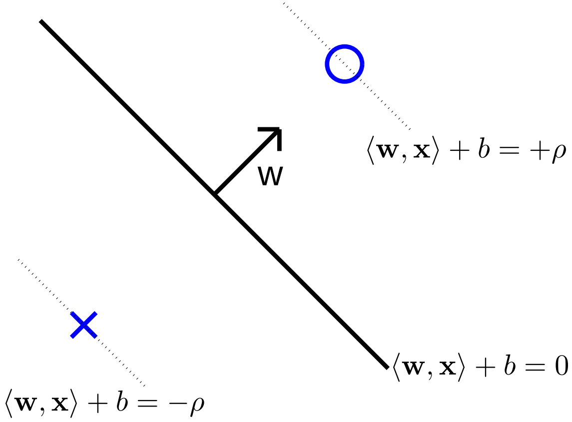
Discriminative Learning (Diagnostic Paradigm)

- **model** $p(y|\mathbf{x})$ (or just boundary: $\left\{\mathbf{x} \,\middle|\, p(y|\mathbf{x}) = \frac{1}{2}\right\}$)
- no naive independence assumption
- in general: less prior knowledge (lower bias, higher variance)
- examples: **SVM**, **Logistic Regression**



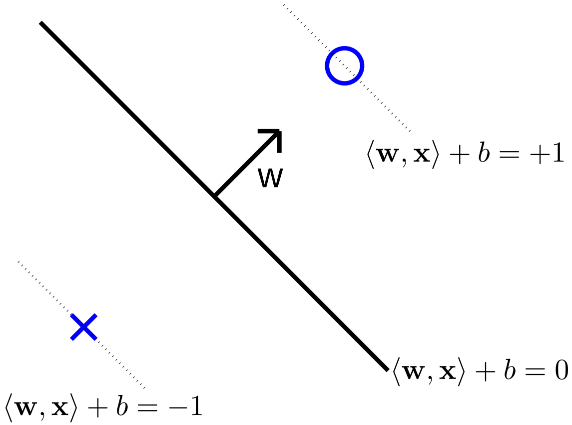
not robust wrt input noise!

SVM:
maximum margin
classifier



$$\max_{\mathbf{w}, b, \rho} \quad \underbrace{\rho}_{\text{margin}} \quad s.t. \quad \underbrace{y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq \rho}_{\text{data fitting}}, \quad \underbrace{\|\mathbf{w}\| = 1}_{\text{normalization}}$$

SVM:
regularized
data fitting



$$\min_{\mathbf{w},b} \quad \underbrace{\frac{1}{2}\langle \mathbf{w}, \mathbf{w} \rangle}_{\text{regularizer}} \quad s.t. \quad \underbrace{y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1}_{\text{data fitting}}$$

Equivalent Reformulation of the SVM

$$\max_{\mathbf{w},b,\rho} \quad \rho \qquad s.t. \quad y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq \rho, \quad \|\mathbf{w}\| = 1$$

$$\Leftrightarrow \quad \max_{\mathbf{w}',b,\rho} \quad \rho^2 \qquad s.t. \quad y_i \left(\left\langle \frac{\mathbf{w}'}{\|\mathbf{w}'\|}, \mathbf{x}_i \right\rangle + b \right) \geq \rho, \quad \rho \geq 0$$

$$\Leftrightarrow \quad \max_{\mathbf{w}',b,\rho} \quad \rho^2 \qquad s.t. \quad y_i \left(\left\langle \underbrace{\frac{\mathbf{w}'}{\|\mathbf{w}'\|\,\rho}}_{\mathbf{w}''}, \mathbf{x}_i \right\rangle + \underbrace{\frac{b}{\rho}}_{b''} \right) \geq 1, \quad \rho \geq 0$$

$$\Leftrightarrow \quad \max_{\mathbf{w}'',b''} \quad \frac{1}{\|\mathbf{w}''\|^2} \quad s.t. \quad y_i \left(\langle \mathbf{w}'', \mathbf{x}_i \rangle + b'' \right) \geq 1,$$

using $\|\mathbf{w}''\| = \left\| \frac{\mathbf{w}'}{\|\mathbf{w}'\|\,\rho} \right\| = \left| \frac{1}{\rho} \right| \cdot \left\| \frac{\mathbf{w}'}{\|\mathbf{w}'\|} \right\| = \frac{1}{\rho}$

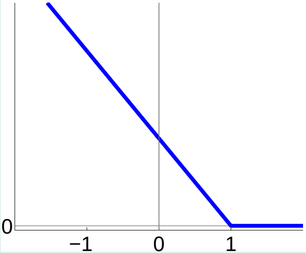
Soft-Margin SVM Loss

$$\min_{\mathbf{w},b,(\xi_k)} \quad \frac{1}{2}\langle \mathbf{w}, \mathbf{w}\rangle + C\sum_i \xi_i$$

$$s.t. \quad y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) \geq 1 - \xi_i, \quad \xi_i \geq 0$$

Effective Loss Function

$$\xi_i = \max\{1 - y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b), 0\}$$



$y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b)$

Logistic Regression (1)

- **log-linear** likelihood ratio:

$$f(\mathbf{x}) := \log\left(\frac{p(y=+1|\mathbf{x})}{p(y=-1|\mathbf{x})}\right) = \mathbf{w}^\top \Phi(\mathbf{x}) + b$$

$\Rightarrow$ prediction for $\mathbf{x}$: $sign\left(\mathbf{w}^\top \Phi(\mathbf{x}) + b\right)$

- implied **likelihood**:

$$p(y=+1|\mathbf{x}) = \frac{1}{1+\exp(-f(\mathbf{x}))}$$

$$p(y=-1|\mathbf{x}) = \frac{1}{1+\exp(+f(\mathbf{x}))}$$

- possibly **Gaussian prior**

$$\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$$

Logistic Regression (2)

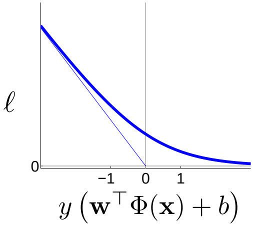
- **maximum likelihood (ML)** estimation: (convex)

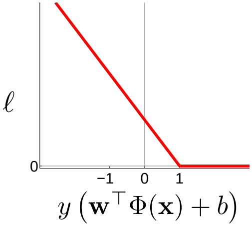
$$\min_{\mathbf{w},b} \quad \sum_i \underbrace{\log\Big(1+\exp\big(-y_i(\mathbf{w}^\top\Phi(\mathbf{x}_i)+b)\big)\Big)}_{=:\ell_{\mathbf{w},b}(\mathbf{x}_i,y_i)}$$

- **maximum a posteriori (MAP)** estimation:

$$\min_{\mathbf{w},b} \quad \lambda\|\mathbf{w}\|^2 + \sum_i \ell_{\mathbf{w},b}(\mathbf{x}_i, y_i)$$

- comparing **LogReg likelihood** $\ell_{\mathbf{w},b}$ to **SVM loss** $\ell_{\mathbf{w},b}$





Representer Theorem

Objective: $J(\mathbf{w}) = \|\mathbf{w}\|^2 + \sum_i \ell_i\left(\mathbf{w}^\top \mathbf{x}_i\right)$.

Representer Theorem:

$\mathbf{w}^\star := \arg\min_{\mathbf{w}} J(\mathbf{w})$ is in the span of the data $(\mathbf{x}_i)$, ie

$$\mathbf{w}^\star = \sum_i \alpha_i \mathbf{x}_i \ .$$

Proof: Let $\mathbf{w}^\star = \underbrace{\sum_i \alpha_i \mathbf{x}_i}_{=:\mathbf{w}_\parallel} + \mathbf{w}_\perp$ with $\mathbf{w}_\perp \perp \mathbf{w}_\parallel$. Then

$$J(\mathbf{w}^\star) = \left\|\mathbf{w}_\parallel\right\|^2 + \|\mathbf{w}_\perp\|^2 + \sum_i \ell_i\left(\mathbf{w}_\parallel^\top \mathbf{x}_i + \mathbf{w}_\perp^\top \mathbf{x}_i\right) = J(\mathbf{w}_\parallel) + \|\mathbf{w}_\perp\|^2$$

■

Non-Linearity via Kernels

Kernel Functions

Use feature map $\Phi(\mathbf{x})$, **kernel** $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle$.

- Intuitively, kernel measures similarity of two objects $\mathbf{x}$.
- Fct is kernel $\Leftrightarrow$ fct is *positive semi-definite*.

Kernelization possible if data access only through dot products:

- requires l_2-regularization: $\|\mathbf{w}\|_2 = \langle \mathbf{w}, \mathbf{w} \rangle$
- SVMs, LogReg, LS-Regression, GPs, ...

Three Routes to Kernelization

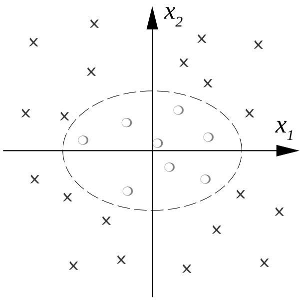
1. Dualization (the classic):
   eg SVM: $\min_{\alpha} \alpha^\top \mathbf{H}\alpha - \mathbf{1}^\top \alpha$ with $H_{ij} = y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)$
2. Plug in Representer Theorem:
   $\mathbf{w} = \sum_i \alpha_i \Phi(\mathbf{x}_i)$; now optimize α instead of $\mathbf{w}$
3. Re-represent data: $E := span\{\Phi(\mathbf{x}_i)\} \triangleq \mathbb{R}^N$ (Repr. Thrm.)

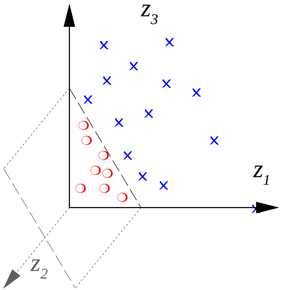
   1. expand basis vectors $\mathbf{v}_i$ of E: $\mathbf{v}_i = \sum_k A_{ik} \Phi(\mathbf{x}_k)$
   2. orthonormality gives: $(A^\top A)^{-1} = K$
      solve for A, eg by KPCA or Choleski decomposition
   3. project data $\Phi(\mathbf{x}_i)$ on basis $V = (\mathbf{v}_j)_j$:
      $\tilde{\mathbf{x}}_i = V^\top \Phi(\mathbf{x}_i) = (A)_i$

Non-Linear Mappings

Example: All Degree 2 Monomials

$$\begin{aligned} \Phi : \mathbb{R}^2 &\rightarrow \mathbb{R}^3 =: \mathcal{H} \quad (\text{"Feature Space"}) \\ (x_1, x_2) &\mapsto (z_1, z_2, z_3) := (x_1^2, \sqrt{2}\, x_1 x_2, x_2^2) \end{aligned}$$





Kernel Trick

Example: All Degree 2 Monomials for a 2D Input

$$\begin{aligned}
\left\langle \Phi(x), \Phi(x') \right\rangle &= \left\langle (x_1^2, \sqrt{2}\, x_1 x_2, x_2^2), ({x'}_1^2, \sqrt{2}\, x'_1 x'_2, {x'}_2^2) \right\rangle \\
&= \left(x_1 x'_1 + x_2 x'_2 \right)^2 \\
&= \left\langle x, x' \right\rangle^2 \\
&=: k(x, x')
\end{aligned}$$

$\longrightarrow$ the dot product in $\mathcal{H}$ can be computed in $\mathbb{R}^2$

Popular Discriminative (Kernel-) Classifiers

method	**SVM**	**Logistic Regression**	**Fisher Linear Discriminant**
models	$p(y\|x) = 0.5$	$p(y\|x)$	$p(y\|x)$
probabilistic	no	yes	yes
coefficients α	sparse $\Rightarrow$ efficient optimization	full	full
difference to SVM	—	uses logistic loss fct.	maximizes average margin

Parametric vs Non-Parametric

Two alternative views (depending on kernel):

- linear kernel: **parametric** method

  fixed number of parameters

  $\mathbf{w} \in \mathbb{R}^D$

- non-linear kernel: **non-parametric** method

  number of parameters α_i increases with number of data points

  $\alpha \in \mathbb{R}^N$

Spectral Clustering

Slides by Ulrike von Luxburg (MPI biol. Kybernetik, Tübingen)

* deferred to next week *

Further Reading

- Matrix calculus: `http://www.cs.toronto.edu/~roweis/notes.html`
- Multivariate normal distribution: `http://en.wikipedia.org/wiki/Multivariate_normal_distribution`
- Shrinkage: **Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution.** *Charles Stein.* Proc. Third Berkeley Symp. on Math. Statist. and Prob., Vol. 1 (Univ. of Calif. Press, 1956), 197-206. `http://projecteuclid.org/euclid.bsmsp/1200501656`
- Least Squares and Logistic Regression: **The Elements of Statistical Learning.** *Hastie, Tibshirani and Friedman.* Springer-Verlag, 2001.
- GPs: `http://www.gaussianprocess.org/`
- SVMs: `http://www.svms.org/tutorials/`
- Kernels and Kernel Machines:
  - **Learning with Kernels.** *Bernhard Schölkopf and Alex Smola.* MIT Press, Cambridge, MA, 2002.
  - `http://www.kernel-machines.org/`
- Spectral Clustering: **A Tutorial on Spectral Clustering.** *Ulrike von Luxburg.* Statistics and Computing 17(4): 395-416, 2007.
- Any statistical term: `http://en.wikipedia.org/`
  (eg Shrinkage estimation of covariance matrices:
  `http://en.wikipedia.org/wiki/Estimation_of_covariance_matrices`)

Schedule "Machine Learning II" SS'08

- 22.04. Semi-Supervised Learning
- 29.04. Kernels for Structured Data
- 06.05. Applications in Intrusion Detection
- 13.05. Text Mining
- 20.05. Bioinformatics
- 27.05. Optimization for SVMs and Math Programs
- 03.06. Large Scale Optimization
- 10.06. Relevant Dimensionality Estimation
- 17.06. Boosting and Ensemble Methods
- 24.06. Boosting and SVMs
- 01.07. Hidden Markov Models
- 08.07. Structured Output SVMs, Conditional Random Fields
- 15.07. Graphical Models