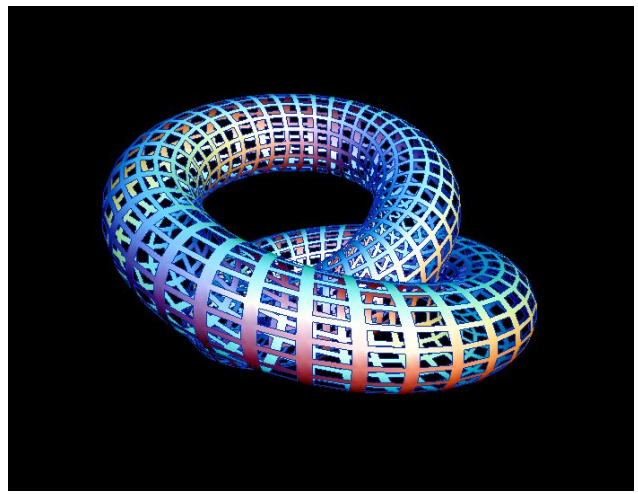
Learning and Prediction - a Survey

Kristiaan Pelckmans

IT/Uppsala University, SE
ESAT - SCD/sista, KULeuven, Leuven, Belgium

Juli 03, 2009



Overview

PART I. - A Birdseye View on ML

Different views to the same:

1. Stochastic
2. Function Approximation
3. Online Learning
4. Optimization

PART II. - Algorithms

1. Pattern Recognition
2. Regression

PART III. - Applications

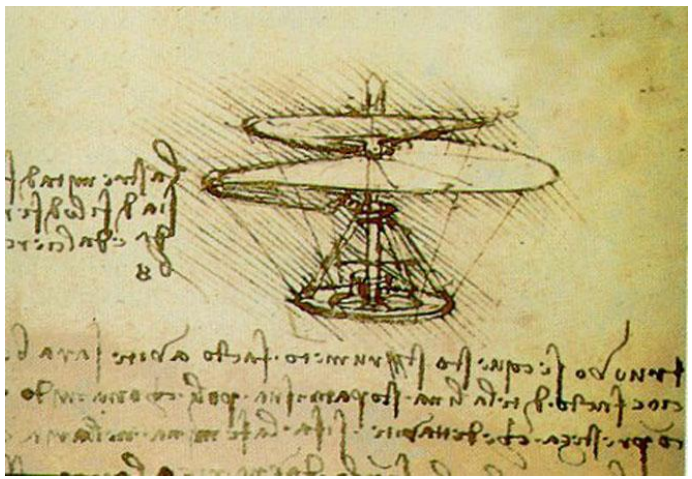
1. Reliability Analysis
2. System Identification

PART I

I. - A Birdseye View on Machine Learning

Making Models

Observations

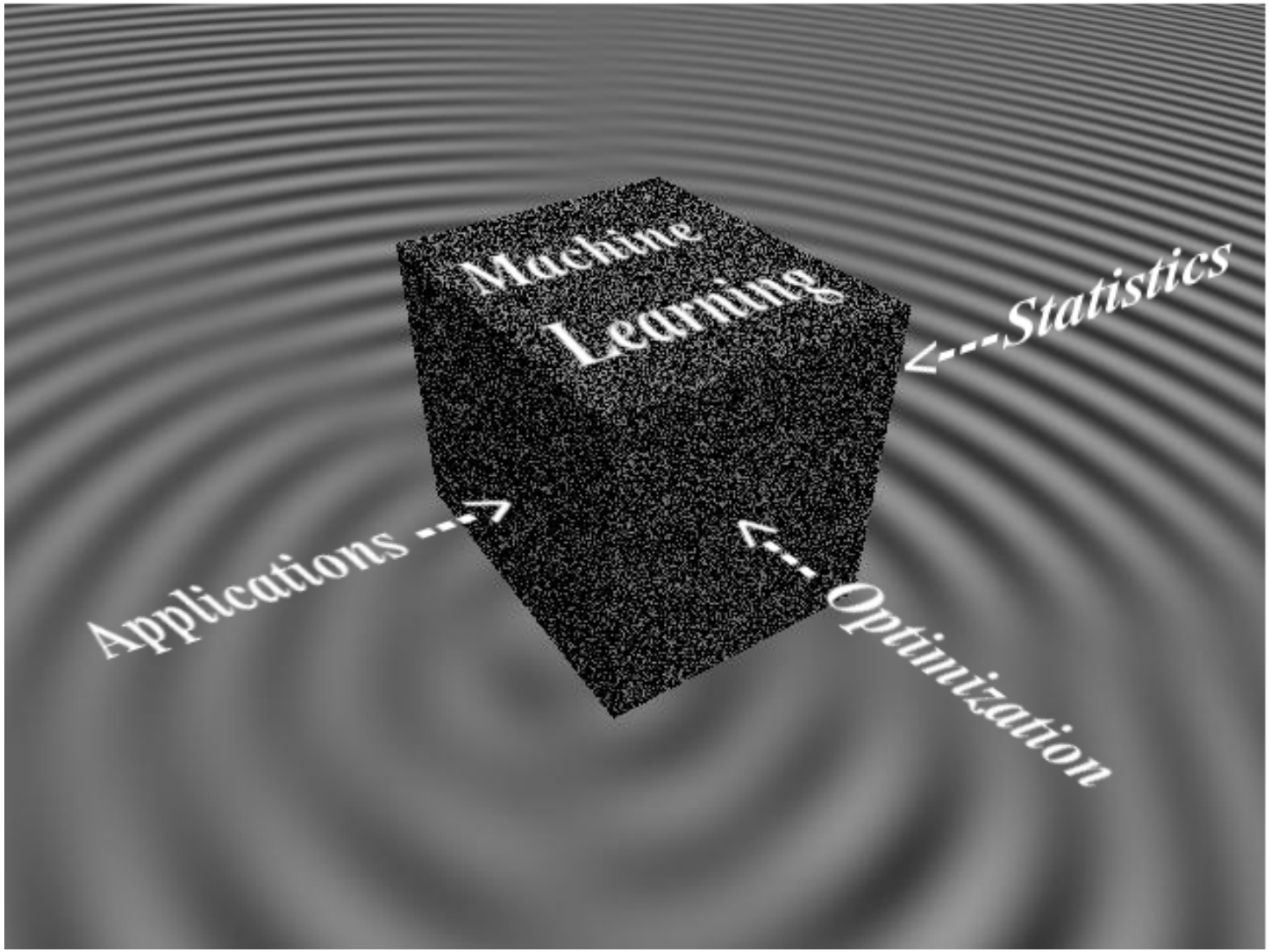


Model

Why?

1. Insight
2. Parameters
3. Predictions
4. Causal (control)

Making Models



Making Models

Relations

1. Compression and Information Theory
2. Algorithms (CS)
3. Signal Processing
4. Optimization theory/modelling
5. Inverse problems and Operators
6. Game theory

Making Models

"Whoa! *That* was a good one! Try it, Hobbs — just poke his brain right where my finger is."

Theorems:

1. Mistake Bound for Perceptron
2. Representer Theorem
3. Concentration Bound

Theorems as a sanity check & give a controlled R&D environment.

A Simple Example: The Expected Location

Given: An unordered set $\{z_i\}_{i=1}^{n} \subset \mathbb{R}$,

Task: Where will z_{n+1} lie?

Main Theme: Stochastic versus Deterministic Inference (Ct'd)

"Essentially, all models are wrong, but some are useful" - G.E.P. Box.

- In the context of closed-form (parametric) stochastic models

Main Theme: Stochastic versus Deterministic Inference

Stochastic: Associate to each set $A \subset \mathbb{R}$ a function $P(\cdot \in A) : \mathbb{R} \to [0, 1]$

Meaning: Frequency, Rational Believe, Bet, ...

IID: Independent and Identically Distributed

$$P(Z_1 \in A, \ldots, Z_n \in A) = P(Z_1 \in A). \ldots . P(Z_n \in A)$$

Exchangeable: for each permutation $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$:

$$P(Z_1 \in A, \ldots, Z_n \in A) = P(Z_{\pi(1)} \in A, \ldots, Z_{\pi(n)} \in A)$$

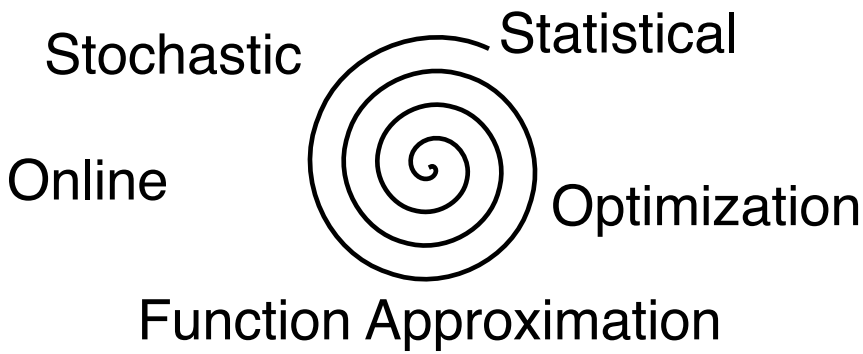
Expectation: or 'the expected number of occurences when sampling IID'

$$P(Z \in A) = E\left[I(Z \in A)\right]$$

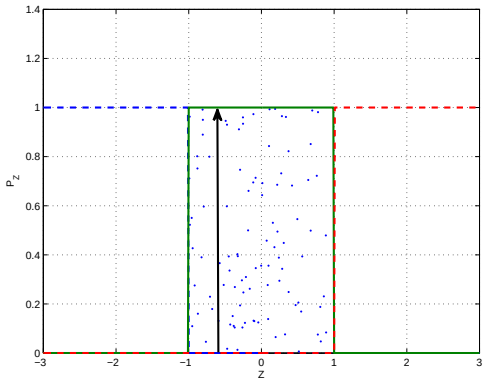
with $I(z) = 1$ if z holds true, 0 otherwise.

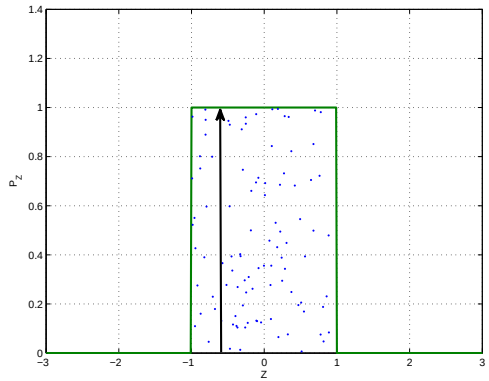
Deterministic: No assumption

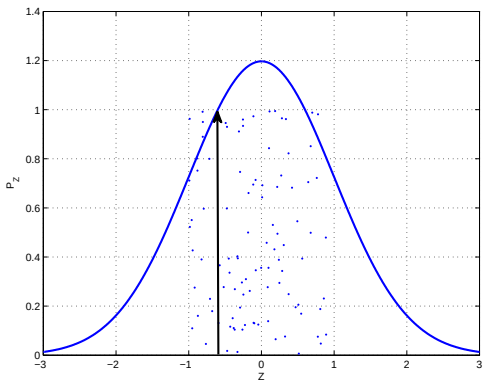
Different Views to the Same

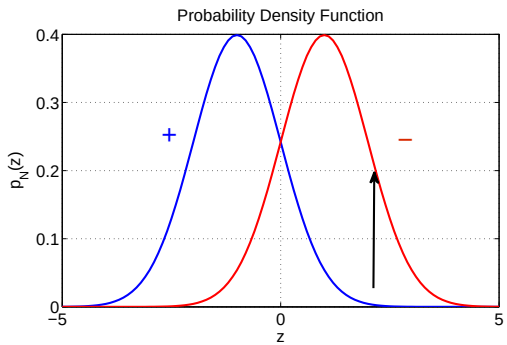


Parametric Statistical Inference









Given stochastic sample

$$\{Z_i\}_{i=1}^n$$

Probability constraint:

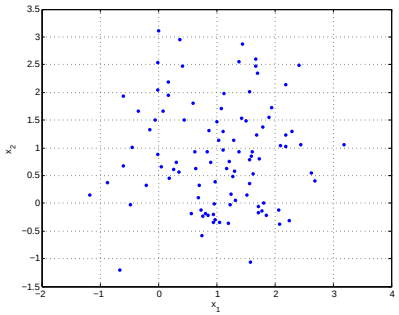
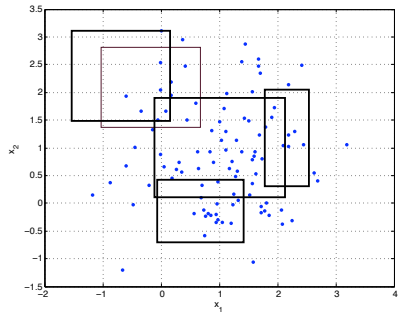
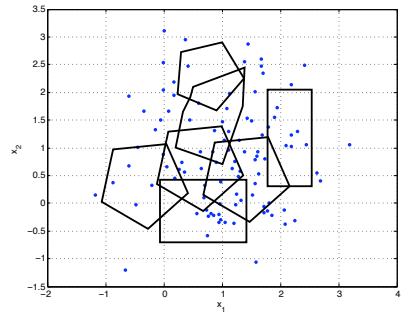
$$\int p_\theta(z)dz = 1$$

Maximum Likelihood:

$$\hat{\theta} = \arg\max_\theta \prod_i p_\theta(Z_i)$$

..it explains the data best.

Distribution-Free Statistical Inference

Sample $\{Z_i\}_i \in \mathbb{R}^2$ of size n

1. For all $A \in \mathcal{A}$, is $\frac{1}{n}\sum_{i=1}^{n} I(Z_i \in A) \to P(A)$?
2. $\mathcal{A}$: Set of Rectangles
3. $\mathcal{A}$: Set of Polygons

Aim:

Stat. What happens if $n \to \infty$ (limit distribution)

SLT How fast (for n)

Distribution-free Inference

Useful for

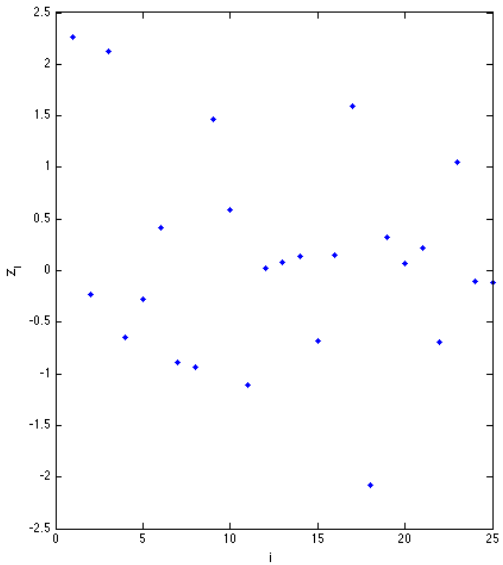
1. Multiple testing: 'Area A contains more than half the probability mass'.
2. Find the smallest set containing the largest mass
3. Test equality of P underlying $\{Z_i\}_i$ and $\{Y_i\}_i$
4. How good is A to predict Z?
5. 'How many prob. conclusions (i.e. p-values) can we extract from a finite sample [Benjamini]'

Distribution-free Inference

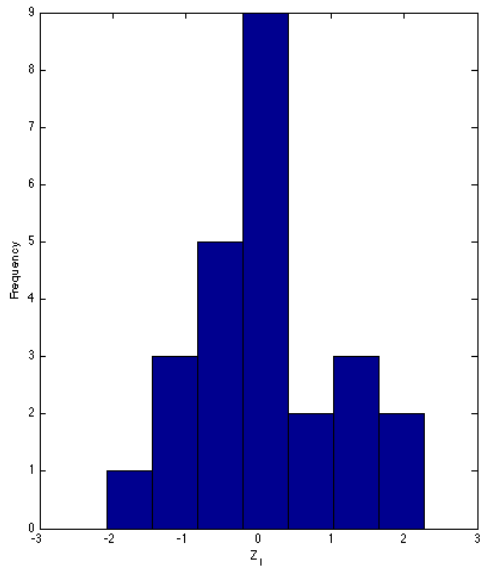
Empirical Distribution Function ECDF $F_n : \mathbb{R} \rightarrow [0, 1]$:

$$F_n(z) = \frac{1}{n} \sum_{i=1}^{n} I(Z_i \leq z)$$

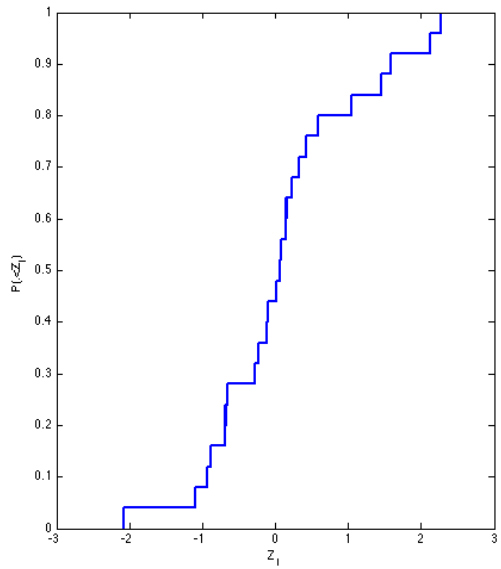
Sufficient:



Sample
$\{Z_i\}_i$

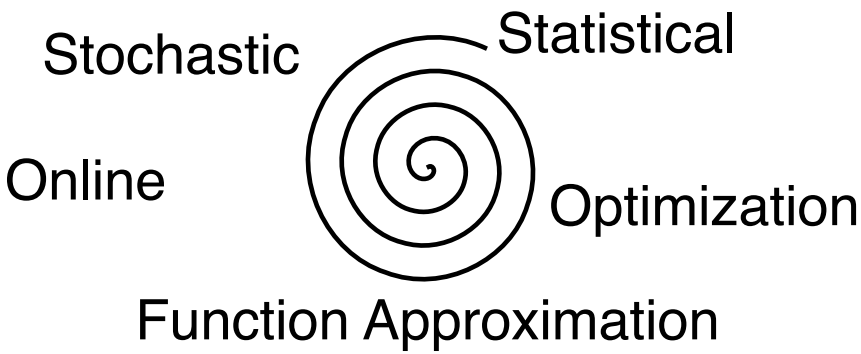


Histogram of sample
$\approx p(Z_i)$



ECDF of sample.
$\approx P(Z_i \leq z)$.

Different Views to the Same: ERM



Empirical Risk Minimization

Prediction: Assume $(X, Y) \overset{\text{i.i.d}}{\sim} F_{XY}$, ($F_{XY}$ fixed but unknown):

Task: Mapping $f : \mathbb{R}^d \to \mathbb{R}$

$$f(X) \approx Y$$

- loss ℓ, design $\mathcal{H}$, structural constraints?

Generalization: risk minimization:

$$f^* = \arg\min_{f \in \mathcal{H}} E\left[\ell(f(X), Y)\right]$$

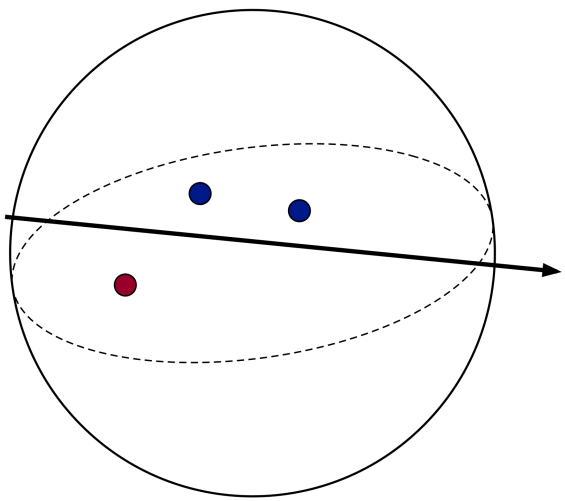
Empirical Risk Minimization: Given a sample $\{(X_i, Y_i)\}_{i=1}^n$,

$$\hat{f} = \arg\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i)$$

Algorithm: Searching in $\mathcal{H}$ for empirical minimum.

Empirical Risk Minimization (Ct'd)

ex. Classification



What: $Y \in \{-1, 1\}$, and $X \in \mathbb{R}^d$

Assumption: $(X, Y) \overset{\text{i.i.d}}{\sim} F_{XY}$,

Realizable: Assume $\exists w \in \mathbb{R}^d, \rho \in \mathbb{R}$ such that

$$Y(w^T X) \geq \rho$$

for all $(X, Y) \sim F_{XY}$

Risk: $\mathcal{R}(w) = P(Y(w^T X) < 0)$

Emp. Risk: $\mathcal{R}_n(w) = \frac{1}{n} \sum I(Y_i(w^T X_i) < 0)$

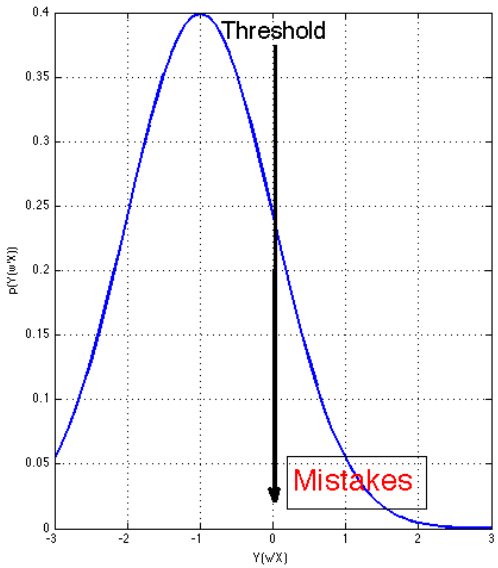
ERM: Find

$$\hat{w} = \arg\min_w \mathcal{R}_n(w)$$

Task: 'What is the probability that

$$Y_{n+1}(\hat{w}^T X_{n+1}) < 0?$$

Empirical Risk Minimization (Ct'd)



Fixed hypothesis: Fix $\bar{w} \in \mathbb{R}^d$

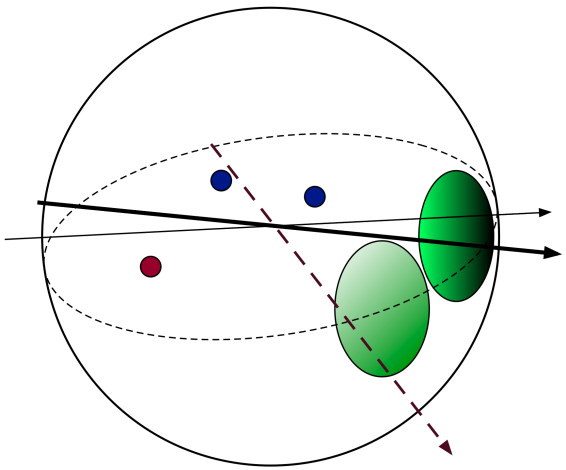
Concentration: 'Suppose there is a probability of mistake $\epsilon > 0$ for $\bar{w}$, what is the chance that a set of n i.i.d. samples will contain no such error?'

$$P\left(\forall i : Y_i(w^T X_i) > 0 \mid \epsilon\right) = (1-\epsilon)^n \leq \exp(-\epsilon n)$$

(Binomial bound)

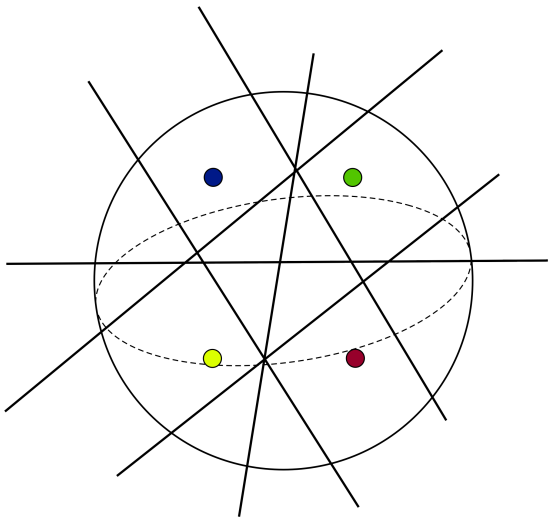
Union Bound : Let this inequality be satisfied for any w representing an equivalence class:

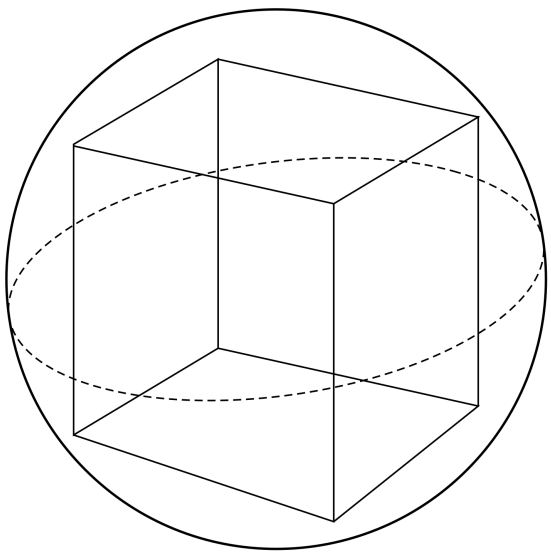
Empirical Risk Minimization (Ct'd)



Structure:	Structure by assumption:
Equivalence Classes:	Instead of considering the infinite set $\{w : \|w\|_2 = 1\}$, choose a representative of each equivalence class
Problem:	'Without the data we do not know which equivalence class to study.'
Solution:	'Guarantee result for all of them.'

Empirical Risk Minimization (Ct'd)





Growth Function: All possible Dichotomies

VC: One can always find a 'covering' with d elements where

$$d \leq \left(\frac{R}{\rho}\right)^2$$

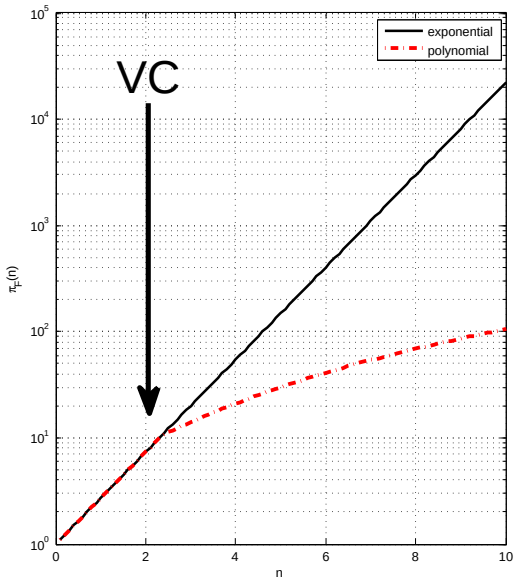
Lemma (Sauer's Lemma)

Number of different equivalence sets of n samples with sets of VC dimension d

$$\sum_{i=0}^{d} \binom{i}{n} \leq (en)^d$$

:

Empirical Risk Minimization (Ct'd)



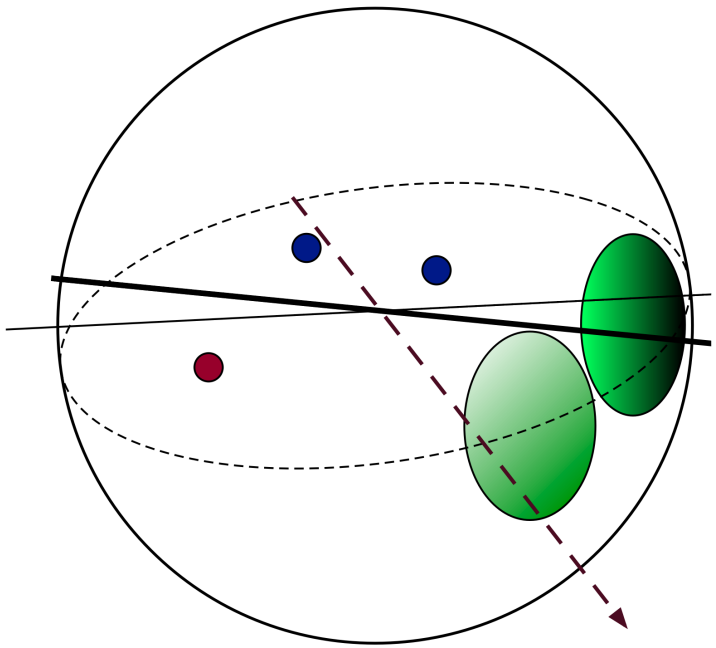
Sauer's lemma - intuition

1. Number of equivalence classes only polynomial
2. Concentration exponential

Hence universal GC: learning!

Empirical Risk Minimization (Ct'd)

Realizable case



Lemma (Probabilistic Guarantee)

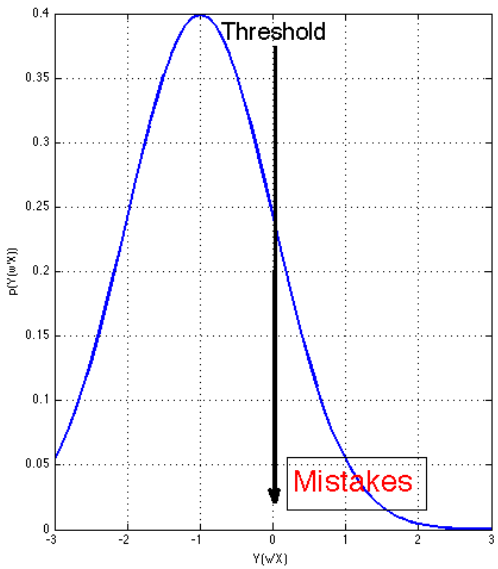
Given $\delta > 0$ *Then with probability exceeding* $1 - \delta$, *and* $\hat{w} \in \mathbb{R}^d$ *such that*

$$\hat{\mathcal{R}}(\hat{w}) = 0$$

we have

$$\mathcal{R}(\hat{w}) \leq O\left(\frac{d\log(n) - \log(\delta)}{n}\right)$$

Empirical Risk Minimization (Ct'd)



Fixed hypothesis: Fix $\bar{w} \in \mathbb{R}^d$ (and NOT $\rho > 0$ where $P(Y(\bar{w}^T X) \geq \rho) = 1$)

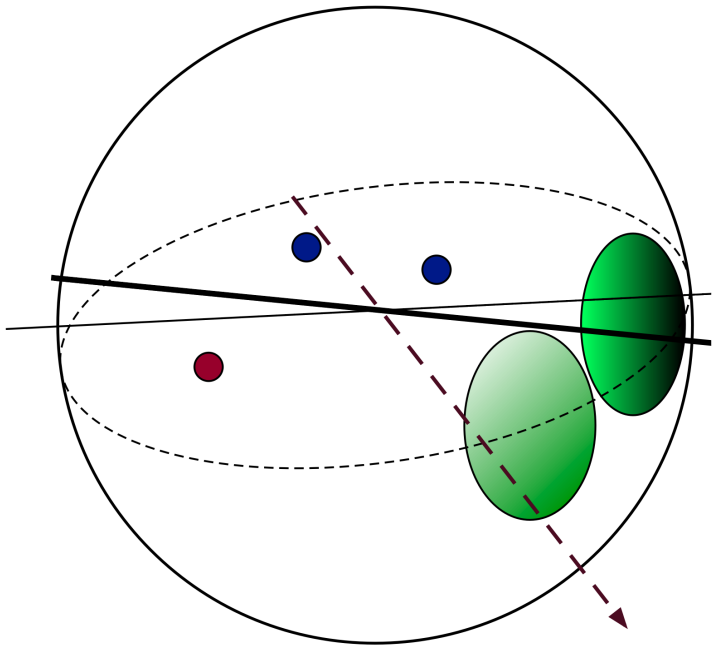
Concentration: Hoeffding: given IID $\{Z_i\}_i \sim Z \in \mathbb{R}$, and let $P(0 \leq Z \leq 1) = 1$, then

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - E[Z]\right| \geq \epsilon\right) \leq 2\exp -\frac{1}{2}\epsilon n^2$$

Union Bound: Let this inequality be satisfied for any w representing an equivalence class...

Empirical Risk Minimization (Ct'd)

Agnostic case

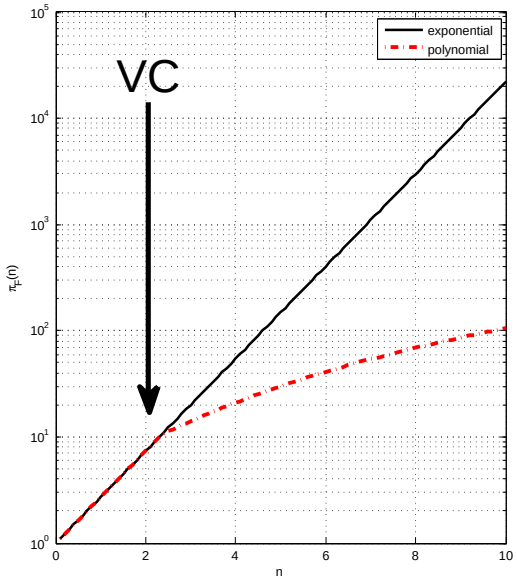


Lemma (Probabilistic Guarantee)

Given $\delta > 0$ *Then with probability exceeding* $1 - \delta$, *and* $\hat{w} \in \mathbb{R}^d$, *we have*

$$\mathcal{R}(\hat{w}) \leq \mathcal{R}(\hat{w}) + O\left(\sqrt{\frac{d\log(n) - \log(\delta)}{n}}\right)$$

Empirical Risk Minimization (Ct'd)



Rademacher variables (where $\{\sigma_i\} \in \pm^n$, $P(\sigma_i = 1) = P(\sigma_i = -1) = 0.5$)

$$R_n(\mathcal{A}) = E\left[\sup_{A \in \mathcal{A}} \frac{2}{n} \sum_{i=1}^{n} \sigma_i I(x_i \in A)\right]$$

1. Technical tool (symmetrisation)
2. Alternative complexity measure to VC
3. Fitting of noise
4. For finite classes $|\mathcal{A}| < \infty$

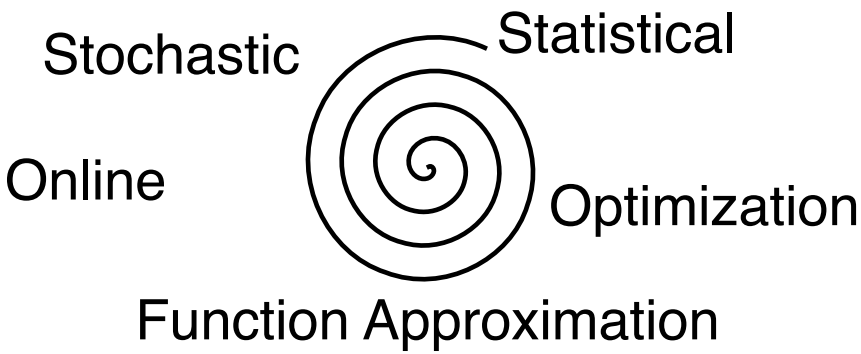
$$R_n(\mathcal{A}) \leq \sqrt{\frac{2 \ln |\mathcal{A}|}{n}}$$

5. For finite VC classes $VC(\mathcal{A}) < \infty$

$$R_n(\mathcal{A}) \leq \sqrt{\frac{2 VC(\mathcal{A}) \ln(n)}{n}}$$

Hence universal GC: learning!

Different Views to the Same: Online Learning



Pattern Recognition as a Game

Algorithm (Perceptron)

1. *Nature* presents $x_t \in \mathbb{R}^d$
2. *Algorithm* predicts

$$\hat{y}_t = \text{sign}(w_{t-1}^T x_t)$$

3. *Nature* returns $y_t \in \{-1, 1\}$
4. *Algorithm* incurs loss $l_t = I(y_t \neq \hat{y}_t) \in \{0, 1\}$
   If $l_t = 1$, *then* $\mathcal{M}_t = \mathcal{M}_{t-1} \cup \{t\}$ *and*

$$w_t = w_{t-1} + y_t x_t,$$

   else $\mathcal{M}_t = \mathcal{M}_{t-1}$ *and* $w_t = w_{t-1}$.

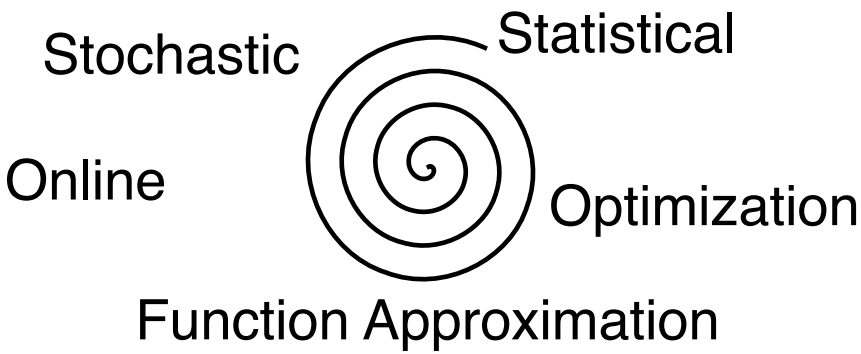
Online Learning

$$\mathcal{R}(w_1, \ldots, w_T) \leq \frac{1}{T} \left(\frac{R_x}{\rho} \right)^2 = O\left(\frac{1}{T} \right)$$

MIIMAX

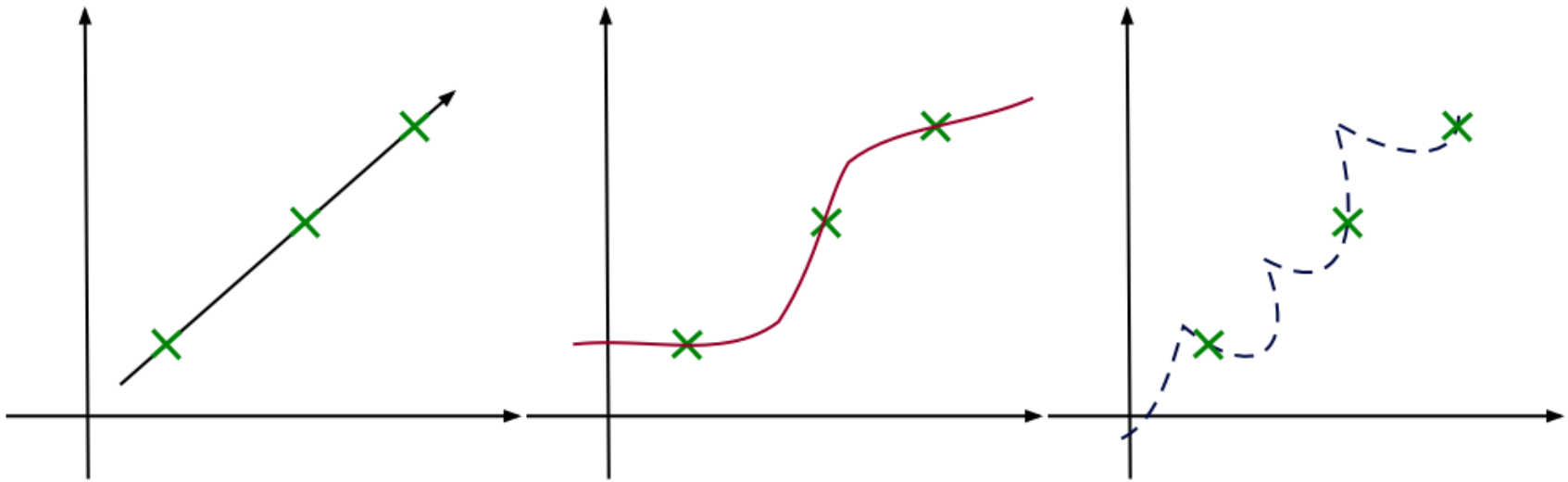
$$\mathcal{R}(w_1, \ldots, w_T) \geq \Omega\left(\frac{1}{T} \right)$$

Different Views to the Same: Function Approximation



Function Approximation

Interpolate points:



Function Approximation (Ct'd)

Interpolate points:

1. Linear Model

$$w : \forall i = 1, \ldots, n : x_i w = y_i$$

2. Smoothing Spline

$$\min_{f : f(x_i) = y_i} \int_z f''(z)^2 dz$$

3. Kernels

$$\min_{f : f(x_i) = y_i} \|f\|_{\mathcal{H}}$$

Function Approximation (Ct'd)

Theorem (Representer Theorem)

Let $f \in \mathcal{H}$ (Hibert space), and $\{f(x_i)\}_i$ linearly independent. Then

$$\hat{f} = \arg\min_{f: f(x_i)=y_i} \|f\|_{\mathcal{H}},$$

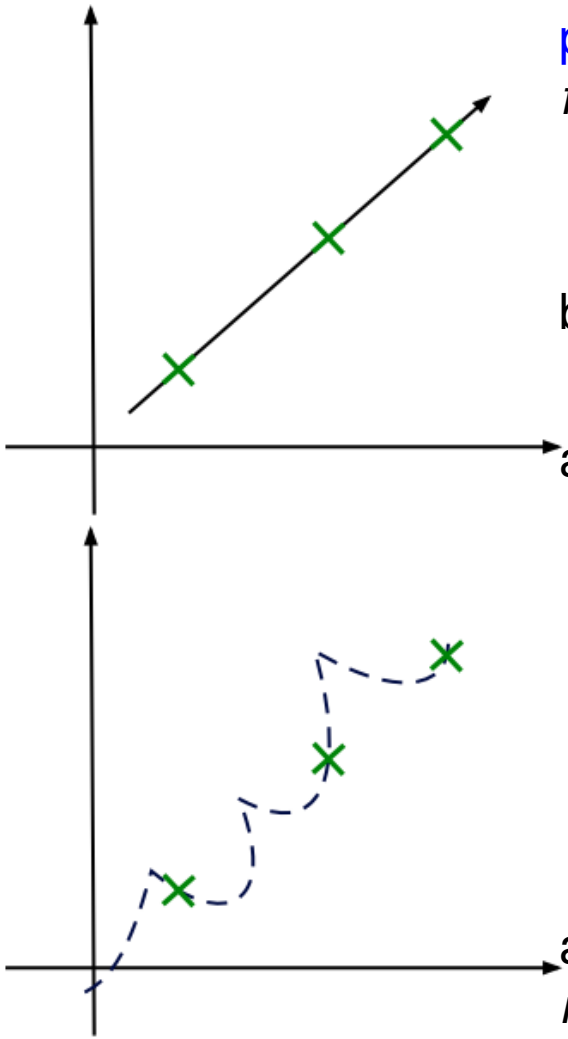
then $\exists \alpha \in \mathbb{R}^n$ such that

$$f(\cdot) = \sum_{i=1}^{n} \alpha_i K(x_i, \cdot)$$

with K the RK associated to $\mathcal{H}$, or $\|f\|_{\mathcal{H}} = K(f, f)$.

(see e.g. [M.Pontil and C. Micchelli, 'on Learning Vector Valued Functions', NC, 2005])

Function Approximation (Ct'd)



proof: Let $f \in \mathcal{H}$ and $f = \hat{f} + g$ with $f \in \mathcal{H}$ and $f(x_i) = y_i$, then

$$\|f\|^2_{\mathcal{H}} = \|\hat{f}\|^2_{\mathcal{H}} + 2 < \hat{f}, g > + \|g\|^2_{\mathcal{H}}$$

but since

$$g(x_i) = 0$$

and hence

$$< \hat{f}, g > = \sum_{i=1}^{n} < K_i \alpha, g > = \sum_{i=1}^{n} < \alpha_i, K_i g >$$

$$= \sum_{i=1}^{n} \alpha_i g(x_i) = 0$$

and hence $\hat{f}$ is always smaller (or equal) in norm than a *nonrepresentable* function f.

Function Approximation (Ct'd)

Kernel:

1. Interpretation as inner product in feature space:

$$K(x, x') = \varphi(x)^T \varphi(x')$$

with a (implicitly defined) mapping $\varphi : \mathbb{R}^d \to \mathbb{R}^{d_\phi}$.

2. As a similarity measure

$$x \approx x' \to K(x, x') \text{ Large}$$

3. as an 'alignment' measure

$$(x, y), (x', y') : \quad y \approx y' \to K(x, x') \text{ Large}$$

Function Approximation (Ct'd)

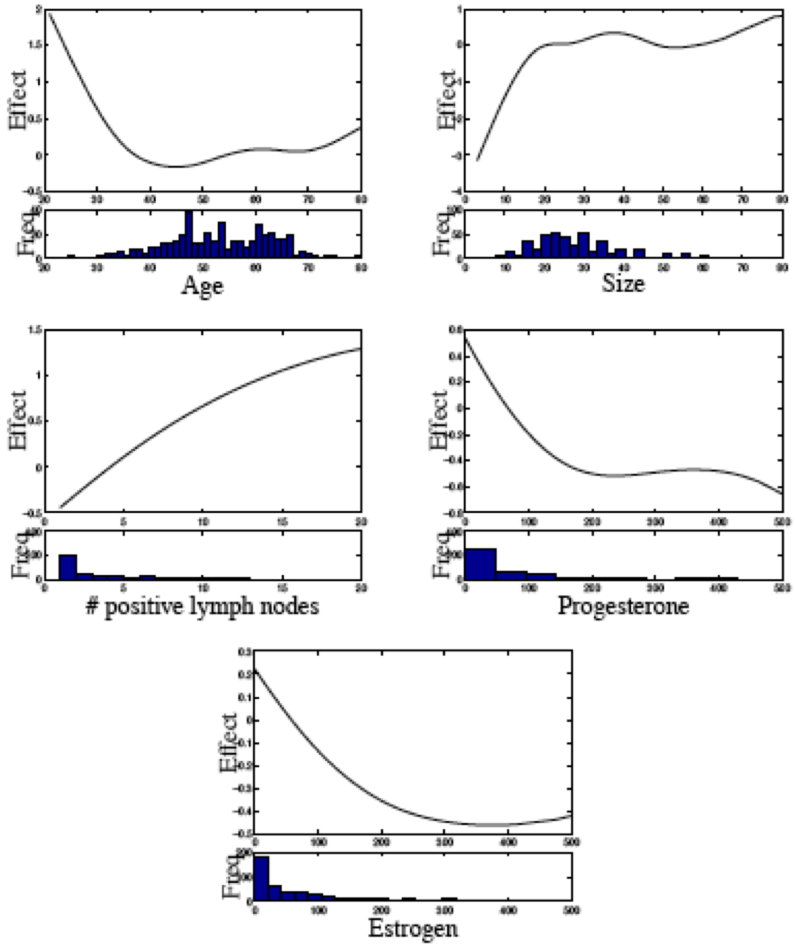


Figure 7: Estimation of the covariate effects on the risk of relapse (remark the difference with Figure 8) with smoothing splines within Cox' proportional hazard model and histograms of the variables. The estimated effects are inversely related with the survival time. The model estimates a lower chance for relapse for older patients up to the age of 40, whereafter an the risk increases again, albeit slowly. The chance for relapse increases for larger tumors until a size of 20mm, whereafter the chance remains fairly constant. For common values of the number of positive lymph nodes and receptors, the risk increases for larger/lower values respectively. Conclusions drawn by the model agree with what is known from literature.

$$f(x) = f_1(x) + \cdots + f_d(x)$$

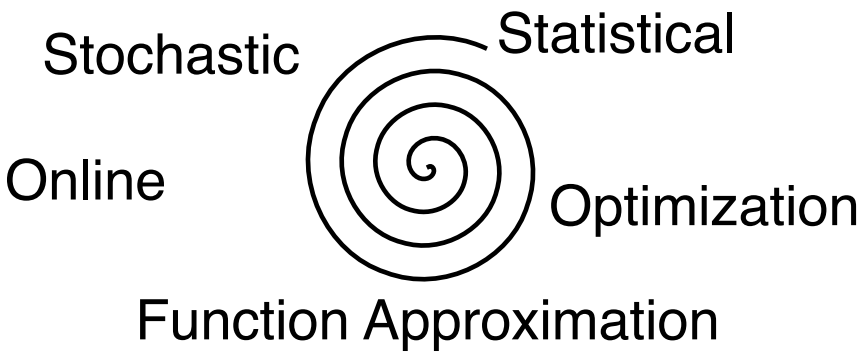
Empirical Risk Minimization (Ct'd)

Functonal ANOVA model $f = \sum_d f_d$

1. Multile Kernel Learning
2. COSSO $\rightarrow \hat{f}_d = 0$
3. LASSO
4. Concurvity (RIP)

[Wahba, Gu, ...], [Signoretto, Pelckmans, 2007-...]

Different Views to the Same: Commodities

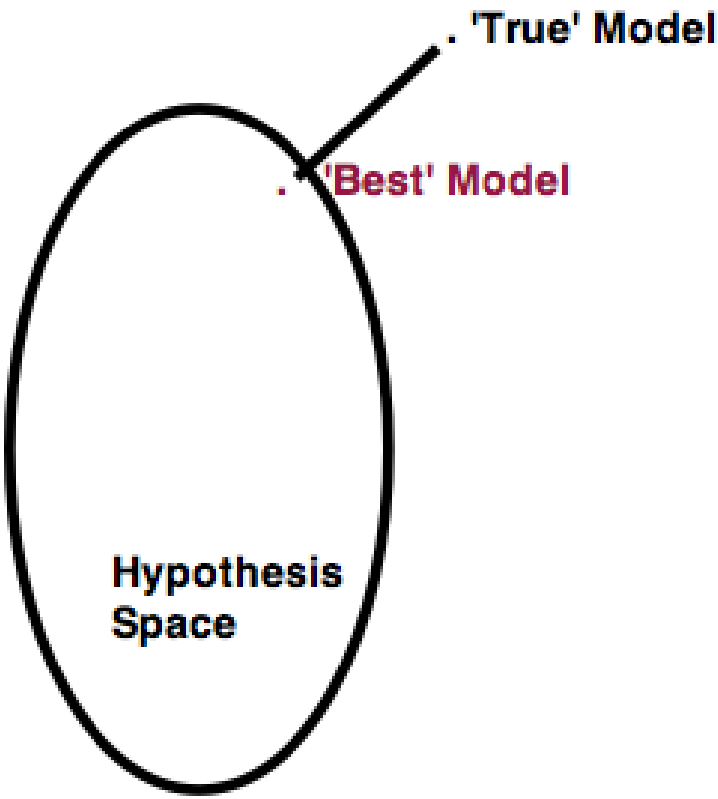


Complexity Control

Common denominator

1. Maximum Likelihood $|\theta| = O(1)$, and AIC
2. Functional Classes $\{f\}$:
   1. Cardinality
   2. VC and growth function
   3. Covering numbers, Baracketting and Metric Entropy
   4. Rademacher complexity
   5. Stability

Bias-Variance Decomposition



Let $\mathcal{R}_*$ the Bayes risk, and let

$$\hat{f} = \inf_{f \in \mathcal{H}} \hat{\mathcal{R}}(f)$$

Decompose error

$$\mathcal{R}_* - \mathcal{R}(\hat{f})$$

in

Bias: Let $f_* = \inf_{f \in \mathcal{H}} \mathcal{R}(f)$, then

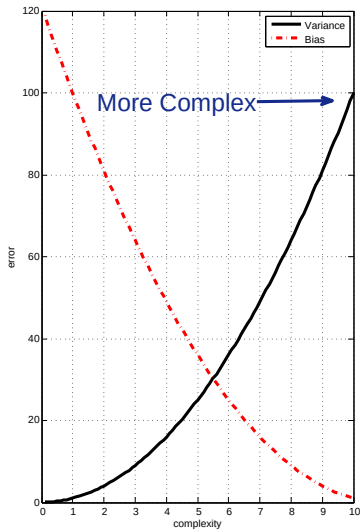
$$\mathcal{R}_* - \mathcal{R}(f_*)$$

Variance: based on sample of size n

$$\mathcal{R}(f_*) - \mathcal{R}(\hat{f})$$

Adding bias for reducing variance: 'less wrong'

Model Selection



1. How does the model perform for its purpose?
2. Mathematically: independent test data
3. Is P(test) $\approx$ P(train)?
4. Cross-validation, Generalized Cross-validation, Goodness-of-fit, Information Criterion
5. Theoretical (sanity check) or heuristic?

Optimization Modeling

Maximum Likelihood (ML):

$$\hat{w} = \arg\max_{w} p_w\big(\{(X_t, Y_t)\}_t\big).$$

Maximum A Posteriori (MAP):

$$\hat{W} = \arg\max_{W} p(\{(X_t, Y_t)\}_t, W\big).$$

Empirical Risk min. (ERM):

$$\hat{w} = \arg\min_{w} \sum_t \mathcal{R}(\{(X_t, Y_t)\}_t; w).$$

Function Approximation (FA):

$$\hat{w} = \arg\min_{w} \sum_t \mathcal{R}(\{(x_t, y_t)\}_t; w).$$

Online Learning (OL):

$$w_t = w_{t-1} - \mu \frac{\partial \mathcal{R}((x_t, y_t); w)}{\partial w}.$$

Optimization Modeling (Ct'd)

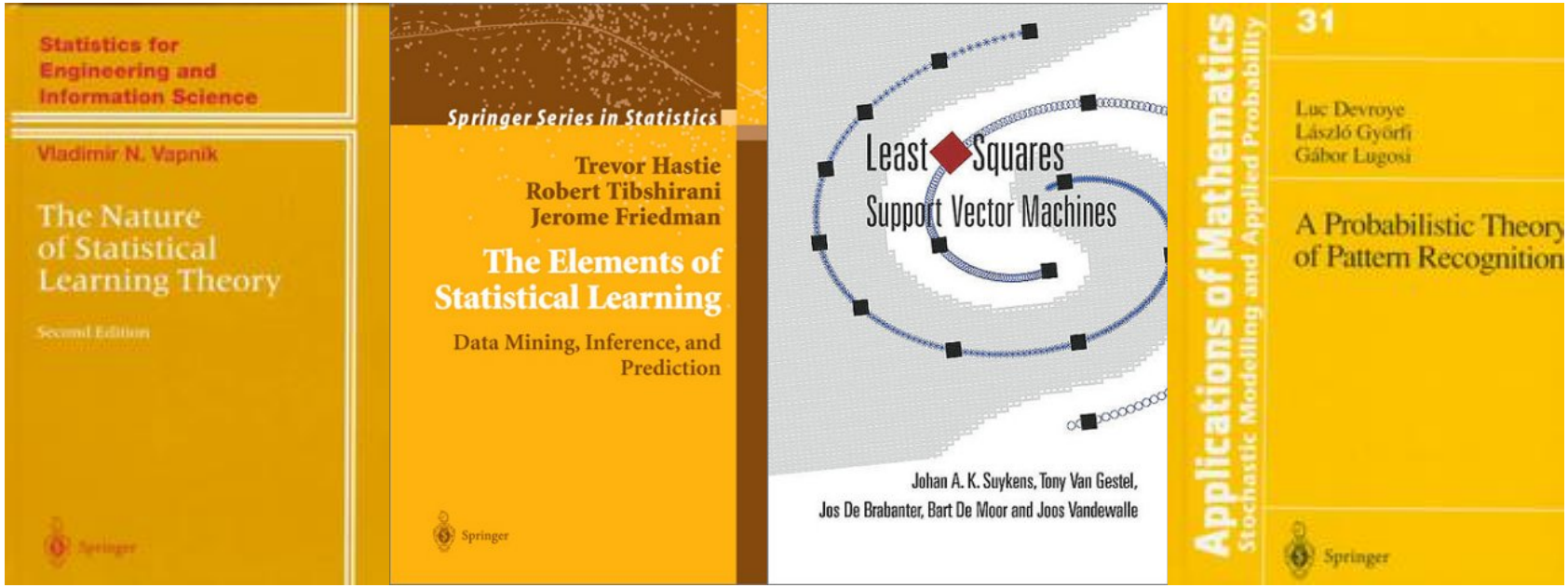
Mean:

$$\hat{\theta} = \arg\max_{\theta} \prod_{i} \exp(-(Z_i - \theta)^2) = \arg\min_{\theta} \sum_{t=1}^{T} (Z_t - \theta)^2 = \frac{1}{T} \sum_{t=1}^{T} Z_t$$

Median:

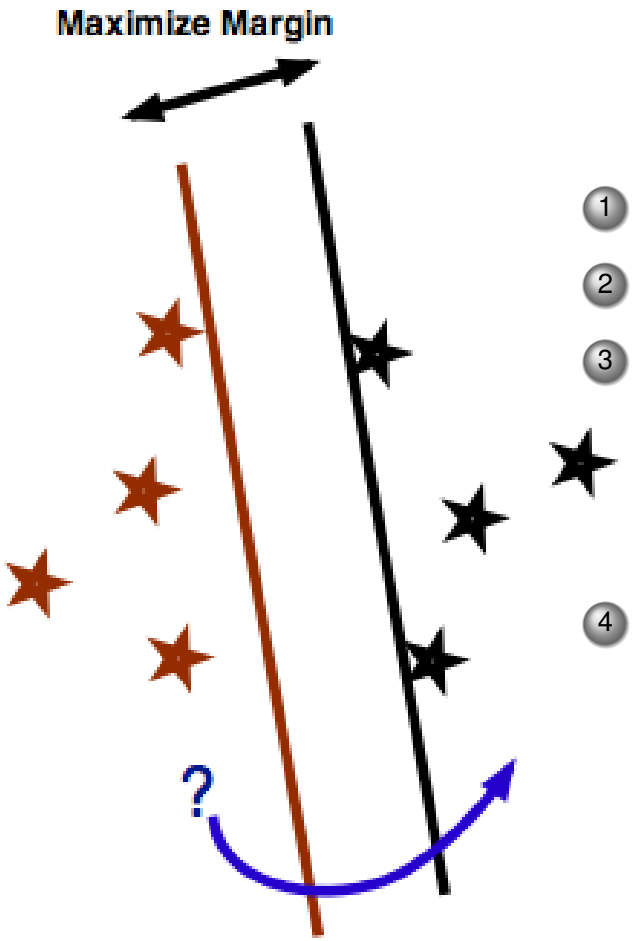
$$\hat{\theta} = \arg\max_{\theta} \prod_{i} \exp(-|Z_i - \theta|) = \arg\min_{\theta} \sum_{t=1}^{T} |Z_t - \theta| = \frac{1}{2}(Z_{(\lfloor T/2 \rfloor)} + Z_{(\lceil T/2 \rceil)})$$

Machine Learning



II. - Algorithms

Pattern recognition: Support Vector Machines



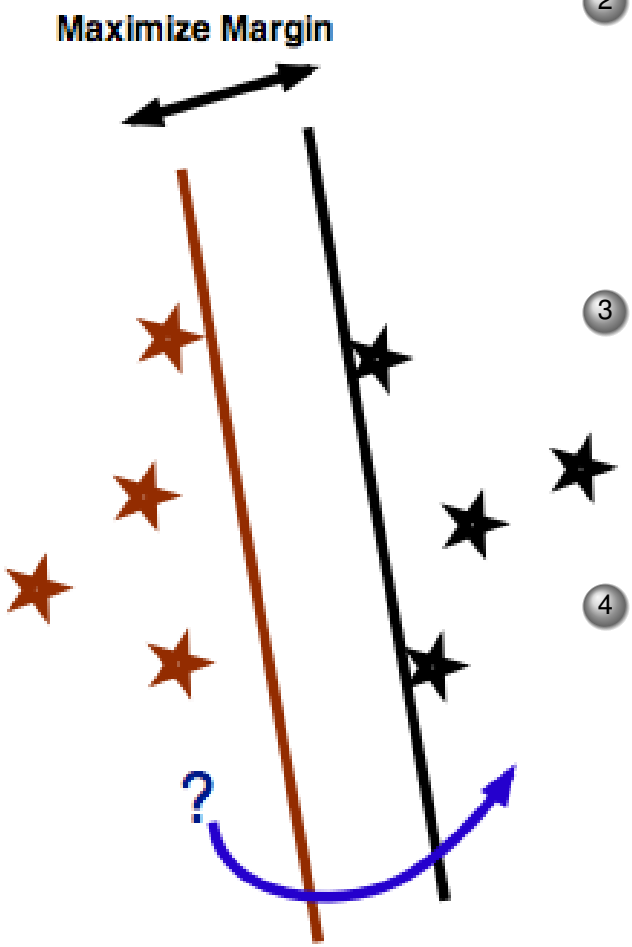
1. Maximize $Y_i(w^T X_i)$ such that $w^T w = 1$
2. Include a bias term b
3. Optimization problem

$$\max \frac{Y_i(w^T X_i + b)}{w^T w}$$

4. Reformulate as

$$\min_{w,b} w^T w \text{ s.t. } Y_i(w^T X_i + b) \geq 1, \forall i = 1, \ldots, n$$

Pattern recognition: Support Vector Machines (Ct'd)



1. Convex Quadratic Program
2. Lagrange dual:

$$\min_{\alpha} \alpha^T \mathbf{YKY} - 1^T \alpha \text{ s.t. } \alpha \geq 0, \forall i = 1, \ldots, n.$$

with $\mathbf{K} = \mathbf{X}\mathbf{X}^T \in \mathbb{R}^{n \times n}$ and $\mathbf{Y} = \text{diag}(y_1, \ldots, y_n)$.

3. Prediction rule for new $x \in \mathbb{R}^d$

$$\hat{w}^T x + \hat{b} = \sum_{i=1}^{n} K(X_i, x)\alpha_i + \hat{b}$$

4. Add slack variables $\{e_i\}_i$ and trde-off with $C > 0$:

$$\min_{w,b,e} \frac{1}{2} w^T w + C \sum_i e_i$$

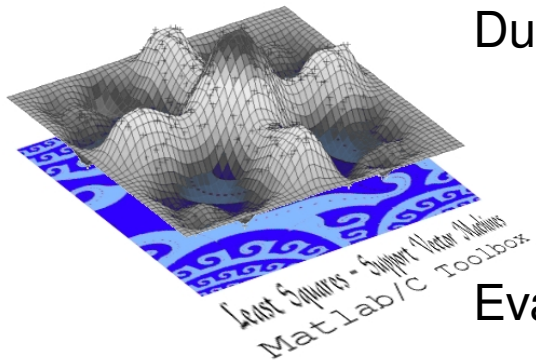
s.t. $Y_i(w^T X_i + b) \geq 1 - e_i, e_i \geq 0, \forall i = 1, \ldots, n.$

5. $\sim$ Ridge Regression, Regularization networks, Gaussian Processes, Smoothing Splines, ...

Least squares Support Vector Machines

Objective function:

$$\min \frac{1}{2} w^T w + \frac{\gamma}{2} \sum_{i=1}^{n} (Y_i (w^T X_i - 1))^2$$



Dual system ($\mathbf{K} = \mathbf{X}\mathbf{X}^T \in \mathbb{R}^{n \times n}$)

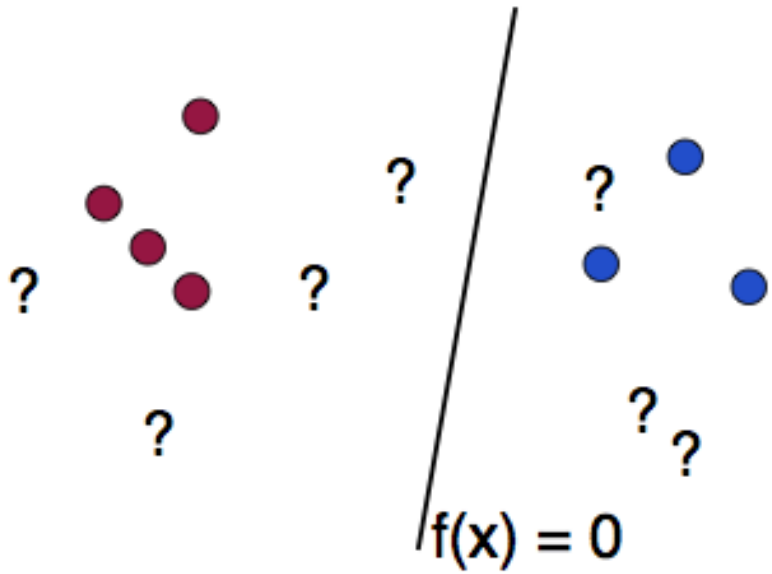
$$\left(\mathbf{K} + \frac{1}{\gamma} I_n \right) \alpha = 1_n$$

Evaluate estimate $\hat{w}$ in new point $x \in \mathbb{R}^d$ as

$$\hat{w}^T x_i = \sum_{i=1}^{n} K(X_i, x) Y_i \alpha_i$$

Solve with Conjugate Gradients

Support Vector Machines: Variations on the Theme



1. Other loss functions
2. Multi-class and Structured Outputs
3. Regression and Unsupervised Learning
4. Semi-Supervised Learning
5. Active Learning and Selective Sampling
6. Missing values (inputs)
7. Time-dependence
8. Other Loss functions

Support Vector Machines: Success stories

svm	lda	rpart	qda	nn	multi	f.mars	f.bruto	nnet
3.14	3.56	5.51		24.27	5.86	4.09		4.49
2.58	2.62	2.62		13.75	0.00	2.62		0.00
15.87	13.67	18.52		19.58	14.70	15.98		14.50
5.93	12.99	12.80		13.62	11.77	9.66		12.13
12.48	13.47	13.38		11.44	13.03	13.62		12.67
3.94	5.05	12.50		0.00	0.00	3.61		0.48
0.14	1.69	8.24		33.97	1.80	1.69		2.57
0.49	5.73	3.20		39.73	2.43	5.88		1.14
1.07	3.57	1.07		28.65	1.97	3.57		1.74
10.96	28.34	30.89		23.44	12.31	16.73		14.98
21.16	22.16	21.48		66.90	22.16	22.16		21.09
15.44	25.20	29.42	25.97	12.69	27.10	22.30	30.08	21.96
23.53	22.60	25.38	25.94	29.74	22.37	22.67	34.90	23.73
29.11	32.03	32.77	39.45	37.12	31.75	29.56	57.96	34.25
5.98	19.48	22.48	33.08	11.69	17.39	14.65	43.26	14.14
23.65	23.12	25.12	28.45	33.13	23.24	24.10		27.59
2.66	49.48	10.65	10.94	5.88	49.49	5.61	4.05	4.17
0.81	49.99	3.70	50.00	0.17	49.99	9.49	7.90	3.37
2.82	3.16	25.98	5.32	7.27	6.95	6.34	4.13	5.32
15.76	18.20	34.36	20.86	25.29	18.58	24.29	21.69	22.01
3.58	38.75	25.28	6.53	40.87	39.07	9.34	9.06	30.47

['The Support Vector Machine Under Test ', Neucom, 2003]

1. Text Classifications and Information Retrieval
2. Computational Biology and Micro-array studies
3. Medical diagnosis
4. Intrusion detection
5. Software: LIBSVM, WEGA, ...

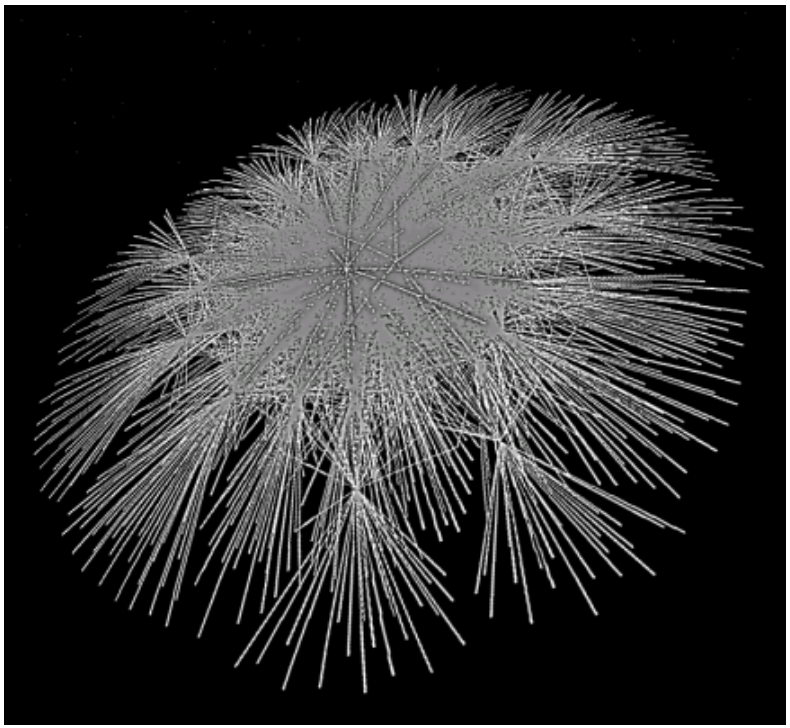
Support Vector Machines: Large

Table 1.1. *Huber's Classification of Dataset Sizes*

Size	Description	Bytes
Tiny	Can be written on a blackboard	10^2
Small	Fits on a few pages	10^4
Medium	Fills a floppy disk	10^6
Large	Fills a tape	10^8
Huge	Needs many tapes	10^{10}

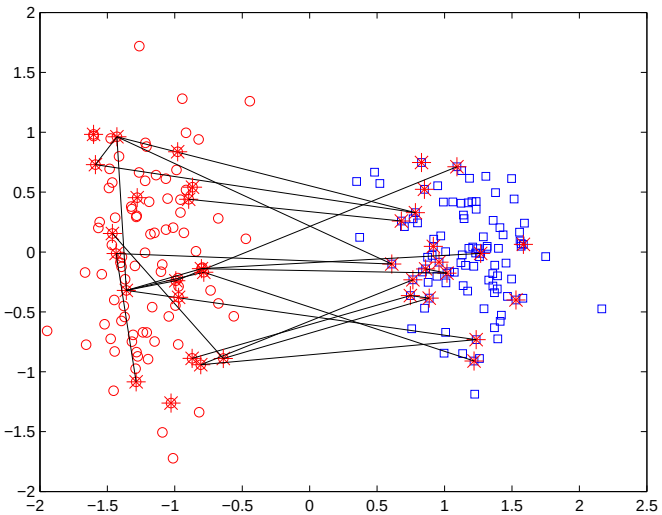
1. Memory
2. Offline - Online
3. Convergence versus approximation
4. Back to the perceptron
5. Ensemble methods and Learning with Experts

Pattern recognition: MINCUT



1. Nodes $V = \{v_i\}_i$ and positively weighted edges $\{a_{ij} \geq 0\}_{ij}$
2. Organization of entries in a graph
3. Rather than explicit representations
4. Interactomics, web, social graphs, kernels, ...
5. Learn a mapping $f : V \rightarrow \{-1, 1\}$, or $q \in \{-1, 1\}^n$

Pattern recognition: MINCUT



1. 'true' labeling $y \in \{\pm 1\}^n$
2. Hypothesis $q \in \{\pm 1\}^n$
3. graph cut

$$\sum_{i,j} I(q_i \neq q_j) a_{ij} = \frac{1}{4} q^T \mathbf{L} q$$

4. Class of hypothesis for given $B \geq 0$

$$\mathcal{H}_B = \left\{ q : \; q^T \mathbf{L} q \leq B \right\}$$

5. Given sample $\mathcal{M} \subset \{1, \ldots, n\}$, then ERM

$$\hat{q} = \arg\min_q \sum_{i \in \mathcal{M}} q_i y_i + \gamma q^T \mathbf{L} q.$$

6. VC dim. of $\mathcal{H}_B$ finite [Pelckmans,2007]
7. Algorithm! [Blum et al.,2002]
8. Graphtron [Pelckmans et al., 2008]

Regression: Least Mean Squares

or Gradient Descent

Algorithm (LMS)

Given learning rate $\nu > 0$.

1. *Nature presents* $x_t \in \mathbb{R}^d$
2. *Algorithm predicts*

$$\hat{y}_t = w_{t-1}^T x_t$$

3. *Nature returns* $y_t \in \mathbb{R}$
4. *Algorithm computes loss* $e_t = (y_t - \hat{y}_t)$

$$w_t = w_{t-1} - \nu e_t x_t,$$

Regression: Least Mean Squares (Ct'd)

Lemma (Shuffling)

For all $z > 0$, vectors $v, w_t, x \in \mathbb{R}^d$

$$2z(w_t^T x - v^T x) = \|v - w_t\|_2^2 - \|v - w_{t+1}\|_2^2 + z^2\|x\|_2^2$$

where $w_{t+1} = w_t - zx$.

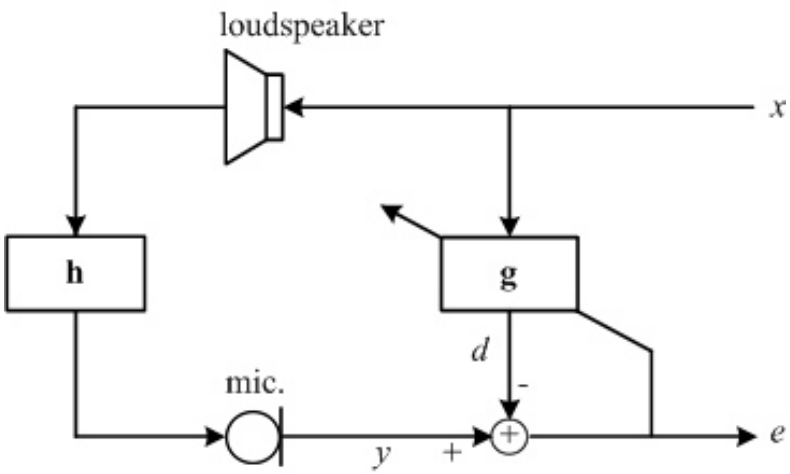
Lemma (Competitive Learning Bound)

Let the LMS algorithm run for T iterations, with $\nu = U/(R_x Z\sqrt{T})$. ($R_x \geq \max_t \|x_t\|_2$ and $Z \geq \max_t |e_t|$)

$$\sum_{t=1}^{n}(w_{t-1}^T x_t - y_t)^2 \leq \min_v \sum_{t=1}^{n}(v^T x_t - y_t)^2 + R_x Z\|v\|_2\sqrt{T}$$

Regression: Least Mean Squares (Ct'd)

Example: Acoustic Echo Cancellation

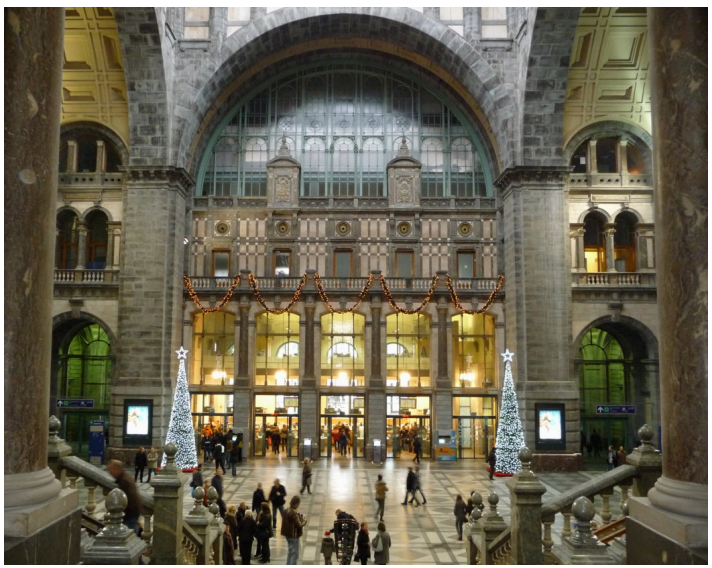


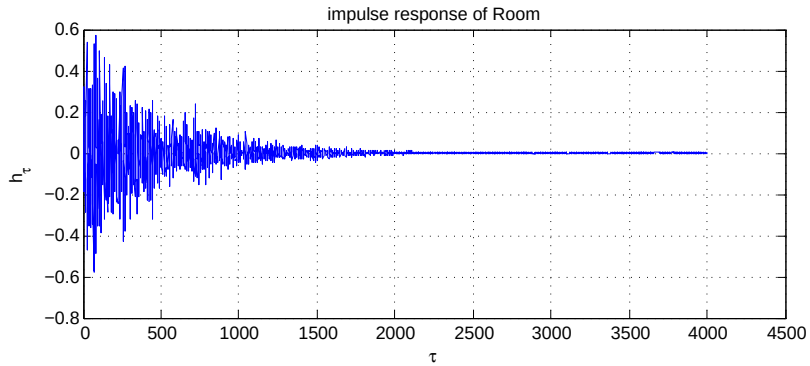
Acoustic Echo Cancellation

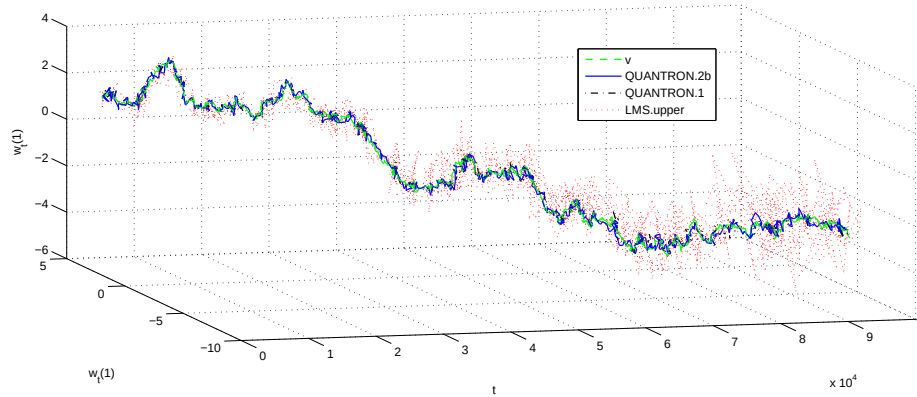
1. u: Signal of Speaker
2. x: Signal coming out of speaker
3. $h(x)$: x bouncing back in room
4. $y = h(x) + u$: Signal picked up by mic.
5. $e = y - \hat{h}(x)$: Echo-cancelled

Regression: Least Mean Squares (Ct'd)

Example: Acoustic Echo Cancellation







Regression: Least Mean Squares (Ct'd)

Example: Acoustic Echo Cancellation

[LMS]

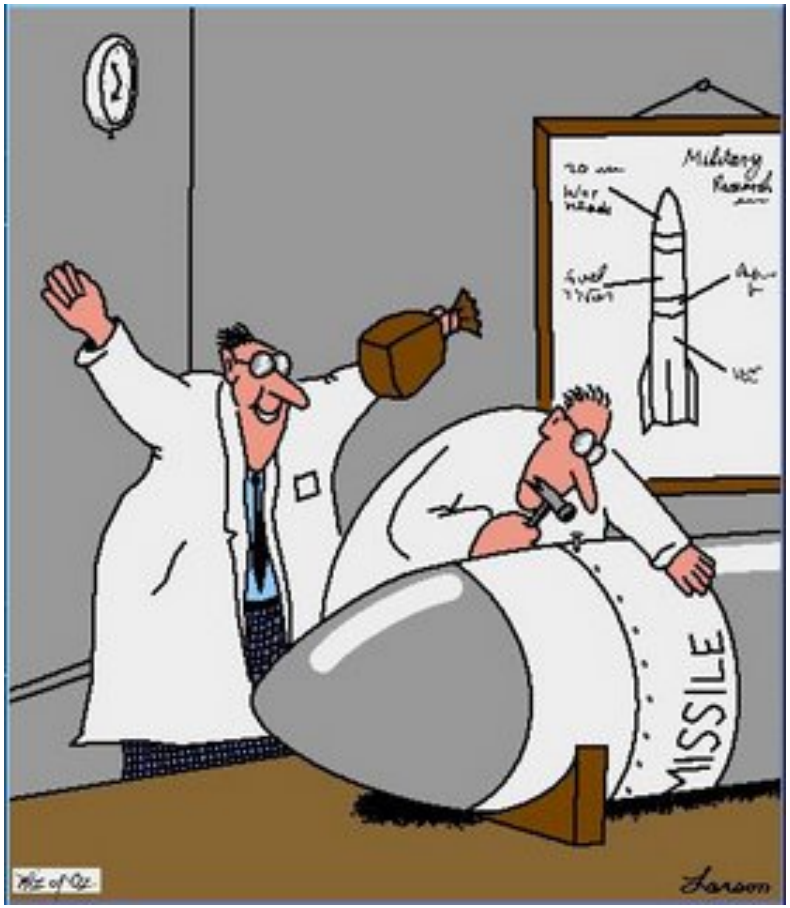
Aim to filter off noise:

1. 'Residual Signal' = information
2. Long filters
3. Properties of speech signal
4.

PART III

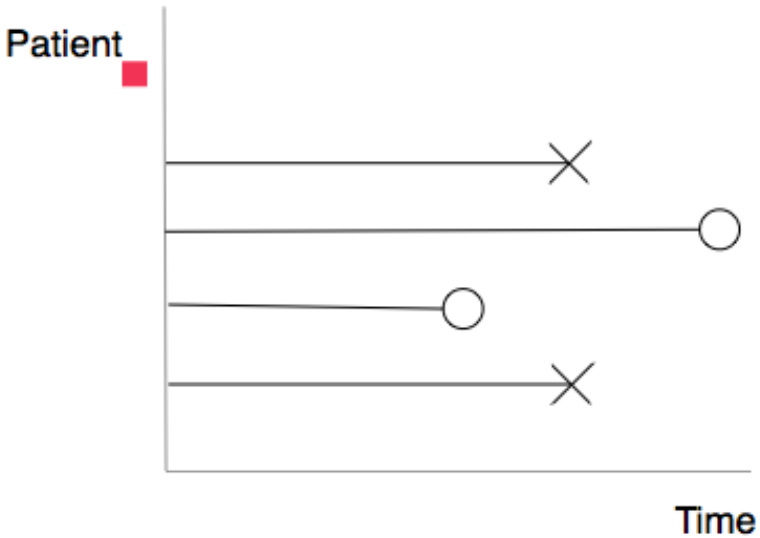
III. - Applications

Reliability Analysis



An ERM approach to modeling Survival [Learning Transformation Models for Ranking and Survival Analysis, Van Belle, Pelckmans et al., 2009]

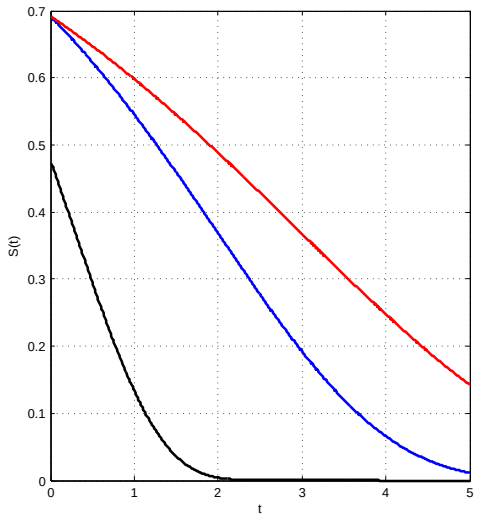
Reliability Analysis (Ct'd)



Events or failure-times:

- '... when patient relapse after surgery'
- '... when mother gives birth'
- '... when next lot with structural deficiencies'
- '... when optimal time for maintenance'
- '... when alert of structural change'

Reliability Analysis (Ct'd)



Survival distributions of $\{T_i\}_i$:

EV: Extreme Value Distribution

$$S_{\mu,\sigma}(t) = 1 - \exp\left(-\exp\left(\frac{-(t-\mu)}{\sigma}\right)\right)$$

Conditional (given covariates) $\{(X_i, T_i)\}_i$

PH: Cox' Proportional Hazard

$$S_w(t|x) = S_0(t)^{\exp(-w^T x)}$$

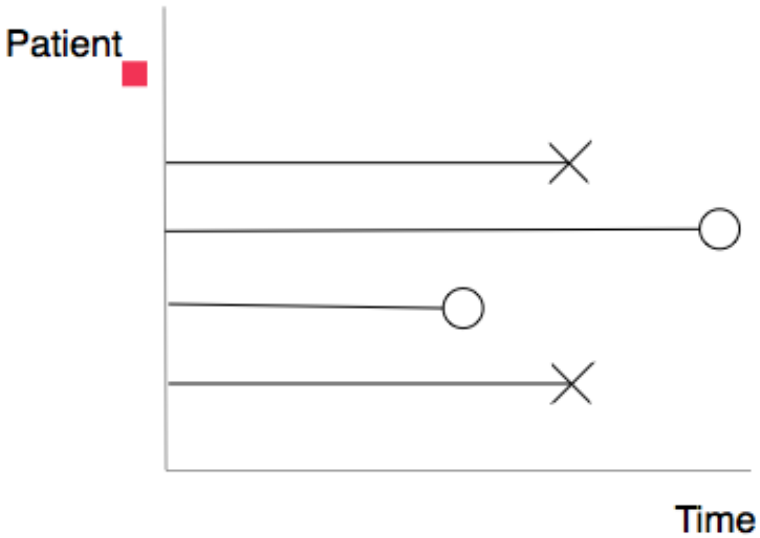
with $S_0 : \mathbb{R}_+ \to [0,1]$ the *baseline survival function*

AFT: Accelerated Failure Time model:

$$\ln\left(\frac{1 - S(t|x)}{S(t|x)}\right) = w^T x$$

and w fitted with maximum likelihood

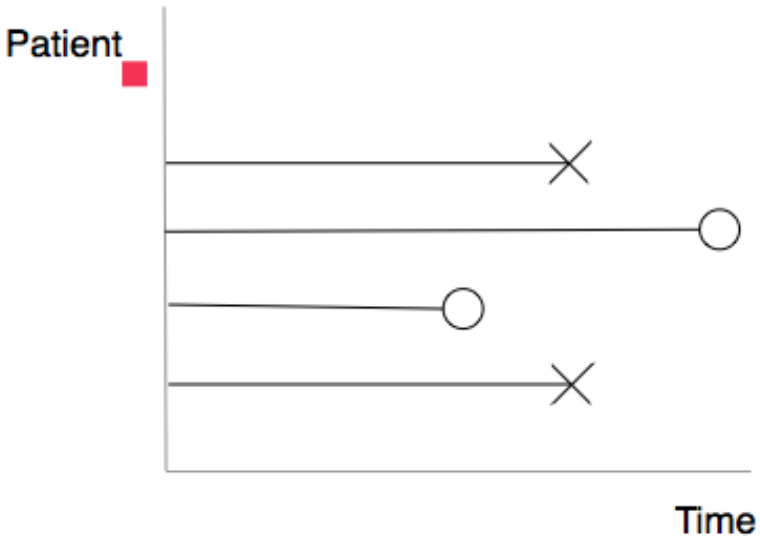
Reliability Analysis (Ct'd)



Censoring:

- '... no observations of events happening in future.'
- '... only possible to make observations on regular timestamps'
- '... uncertain what the begin time of subject is"
- '... uncertain if all (even short time-to-events) are included in study.'

Reliability Analysis (Ct'd)



Censored variables
$\{Z_i = (X_i, T_i, \delta_i)\}_{i=1}^n$. When are two Z_i, Z_j samples comparable?

$$\Delta(Z_i, Z_j) = \begin{cases} 1 & \text{if comparable} \\ 0 & \text{if not} \end{cases}$$

Reliability Analysis (Ct'd)

Risk for suvival data

1. No censoring (AUC):

$$A_n(w) = \frac{2}{n(n-1)} \sum_{i<j} I\left((T_i - T_j)(w^T X_i - w^T X_j) > 0\right)$$

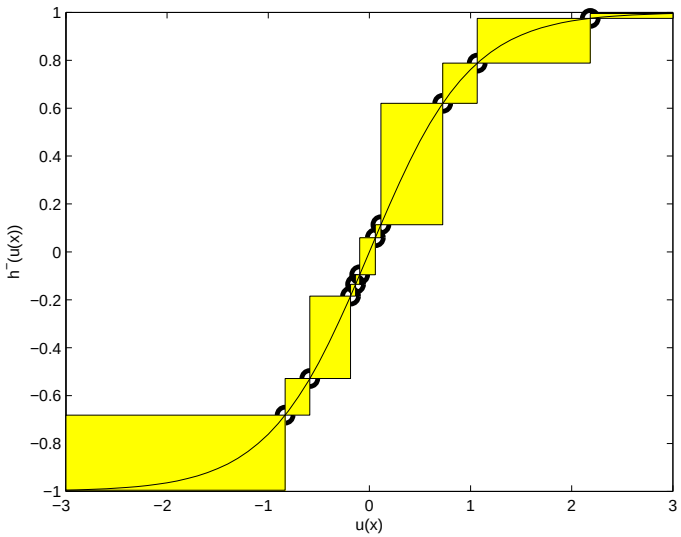
2. With censoring (Harrell's c):

$$C_n(w) = \frac{\sum_{i<j} I\left((T_i - T_j)(w^T X_i - w^T X_j) > 0\right) \Delta(Z_i, Z_j)}{\sum_{i<j} \Delta(Z_i, Z_j)}$$

3. Empirical Risk Minimization

$$\hat{w} = \arg\min_w -C_n(w)$$

Reliability Analysis (Ct'd)



Problem with ERM

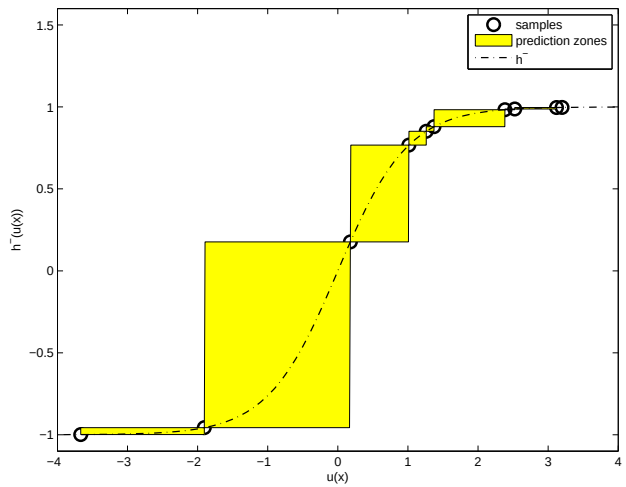
- Combinatorial - algorithm?
- Non-unique solution
- Model - predictions?
- Nonstable
- Interpretation

Realization: if $C_n(w) = 1$, then monotone (transformation) function

Reliability Analysis

Transformation Model

$$X \in \mathbb{R}^d \longrightarrow \boxed{u : \mathbb{R}^d \longrightarrow \mathbb{R}} \longrightarrow$$



$$= Y$$

Examples

- Toy:
$$y = \tanh(w^T x)$$
- Pattern Recognition
$$y = \text{sign}(w^T x + b)$$
- Ordinal Regression (monotonically increasing h)
$$y = \lfloor h(w^T x) \rfloor$$
- Conditional Probability with $g : \mathbb{R} \rightarrow [0, 1]$
$$P(T > t | w^T X) = g(w^T X + b)$$

Complexity control

1. Equivalent to 'the margin' ?
2. Smoothness of transformation function
3. f $h(x) = x$, easiest; if $h(x) = \text{sign}(x)$ hard
4. Formally for all x, x' where $w^T x \geq w^T x'$,

$$h(w^T x) - h(w^T x') \leq L(w^T x - w^T x')$$

   then L is the Lipschitz constant.
5. Necessary condition

$$h(w^T X_{(i)}) - h(w^T X_{(j)}) \leq L(w^T X_{(i)} - w^T X_{(j)})$$

   where $Y_{(i)} \geq Y_{(j)}$ and $h(w^T X_{(i)}) = Y_{(i)}$.

Reliability Analysis (Ct'd)

Transformation Model

MINLIP

1. 'Find Maximally smooth model obeying the observed ordinal relations.'
2. Formulate as a QP

$$\min_{w,L} L \;\text{ s.t. }\; h(w^T X_{(i)}) - h(w^T X_{(j)}) \leq L(w^T X_{(i)} - w^T X_{(j)}), w^T w = 1$$

3. equivalently (since $Y_i = h(w^T X_i)$)

$$\min_{w} w^T w \;\text{ s.t. }\; Y_{(i)} - Y_{(j)} \leq w^T (X_{(i)} - X_{(j)}), \forall Y_{(i)} \geq Y_{(j)}$$

4. By transitivity:

$$\min_{w} w^T w \;\text{ s.t. }\; Y_{(i+1)} - Y_{(i)} \leq w^T (X_{(i+1)} - X_{(i)}), \forall = 1, \ldots, n-1.$$

Reliability Analysis (Ct'd)

Transformation Model

MINLIP

1. 'Find Maximally smooth model obeying the observed ordinal relations.'
2. Formulate as a QP

$$\min_w \frac{1}{2} w^T w \text{ s.t. } \mathbf{D}\mathbf{X}w \geq \mathbf{D}Y$$

   with $\mathbf{D} \in \{-1, 0, 1\}^{n-1 \times n}$ and $(\mathbf{D}Y)_i = y_{(i+1)} - y_{(i)}$ for all $i = 1, \ldots, n-1$.
3. Add slack variables $e = (e_1, \ldots, e_n)^T \in \mathbb{R}^n$ with tradeoff $C > 0$:

$$\min_w \frac{1}{2} w^T w + \frac{C}{2} e^T e \text{ s.t. } \mathbf{D}\mathbf{X}w + \mathbf{D}e \geq \mathbf{D}Y$$

4. Dual problem:

$$\min_\alpha \frac{1}{2} \alpha^T (\mathbf{D}\mathbf{K}\mathbf{D}^T)\alpha - \alpha^T \mathbf{D}Y \text{ s.t. } \alpha_i \geq 0$$

5. and evaluate in $x \in \mathbb{R}^d$ as

$$\hat{w}^T x = \sum_{i=1}^{n-1} (\alpha_{i+1} - \alpha_i) K(X_i, x)$$

Reliability Analysis (Ct'd)

Transformation Model

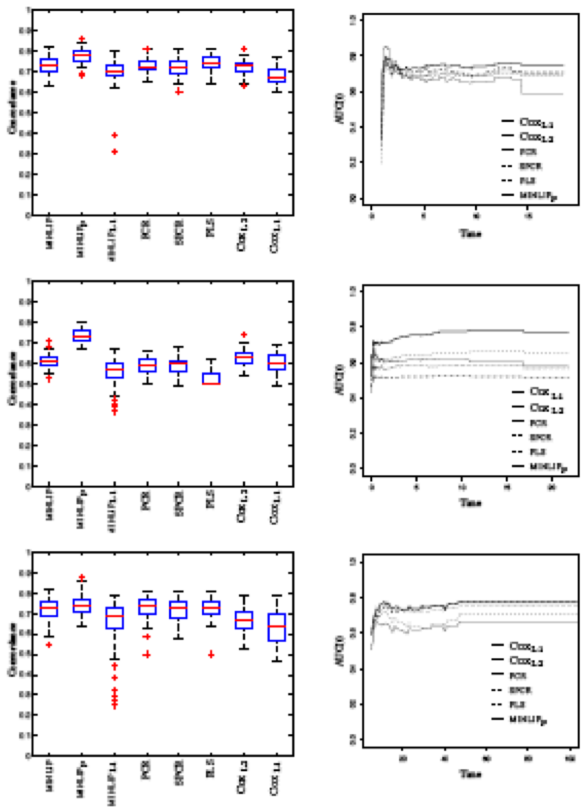


Figure 6: Concordance (left) and time dependent receiver operating characteristic curve (TDROC) (right) on the test set for three micro-array survival data sets (top: DBCD, middle: DLBCL, bottom: NSBCD). The MINLIP_p model obtains a significantly higher performance than all other models. The TDROC for the MINLIP_p model is higher than for the other models indicating that this model is better in distinguishing patients with low and high risk for relapse.

- Microarray datasets [DBCD, DLBCL, NSBCD]
- $d = O(1e3)$ expression levels
- Failure times (breast cancer)
- $n = O(1e2)$ patients
- Performance is measured in AUC (C-index) on testdata, or

$$A^v(\hat{w}) \propto \sum_{i<j} I\left((T_i^v - T_j^v)\hat{w}^T(X_i^v - X_j^v) > 0\right)$$

System Identification

Flavor: Many samples, low dimensions

Purpose: Design Control Strategies

Use: Open- and close loop

Regime: Stochastic (IID?) and Deterministic

mportant: Robustness

System Identification

Aim: Model capturing behavior between $(u_t)_t$ and $(y_t)_t$

Models: Backshift operator $q^{-1}u_t = u_{t-1}$, or

$$y_t = \frac{A(q^{-1})}{B(q^{-1})}u_t$$

with $A(\cdot)$ and $B(\cdot)$ polynomials.

Model Structure: State-Space model for MIMO

$$\begin{cases} x_t = Au_t + Be_t \\ y_t = Cx_{t-1} + D\nu_t \end{cases}$$

Parameters: Linear (FIR,ARX), nonlinear (Box-Jenkins)

Least Squares: Minimal prediction error

$$(\hat{A}, \hat{B}) = \arg\min \sum_t \left(y_t - \frac{A(q^{-1})}{B(q^{-1})}u_t \right)^2$$

System Identification

Air Pollution Control

Prediction:

Regression: RANKTRON

Algorithm (RANKTRON)

Given learning rate $\nu > 0$.

1. *Nature presents* $x_t, x'_t \in \mathbb{R}^d$
2. *Algorithm predicts*

$$\text{sign}(w_{t-1}^T(x_t - x'_t))$$

3. *Nature returns* $\text{sign}(y_t - y'_t)$
4. *Algorithm computes loss*
   $e_t = \text{sign}(w_{t-1}^T(x_t - x'_t)(y_t - y'_t)$
   if $e_t < 0$ *then*

$$w_t = \nu w_{t-1} + (y_t - y'_t)(x_t - x'_t),$$

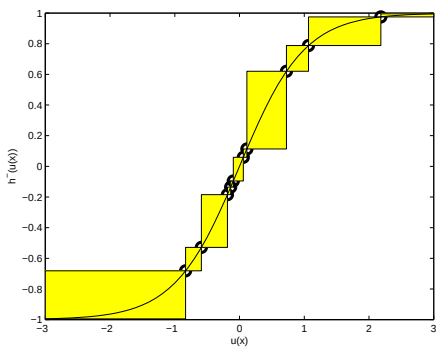
   else $w_t = w_{t-1}$

Online Learning

Lemma (Mistake Bound)

Given a set $\{(x_t, y_t), (x'_t, y'_t)\}_{t=1}^T$. *Let* R_x *be such that* $\|w\|_2 = 1$, $\|x_i\| \leq R_x$ *and* $y = h(w^T x)$ *with* h *L-Lipschitz. Then*

$$\sum_{t \in \mathcal{M}_T} (y_t - y'_t)^2 \leq (2R_x L)^2$$



Proof: Observe that $w_t = \sum_{i \in \mathcal{M}_t} (y_i - y'_i)(x_i - x'_i) \in \mathbb{R}^d$. The proof follows from Cauchy-Schwarz' inequality:

$$\begin{aligned}
\frac{1}{L} \sum_{i \in \mathcal{M}_t} (y_i - y'_i)^2 &\leq \sum_{i \in \mathcal{M}_t} (y_i - y'_i) w^T (x_i - x'_i)) \\
&\leq w^T w_t \leq \|w\|_2 \|w_t\|_2 \\
&\leq \|w\|_2 \sqrt{\sum_{i \in \mathcal{M}_t} \|(y_i - y'_i)(x_i - x'_i)\|_2^2} \\
&\leq \|w\|_2 2R_x \sqrt{\sum_{i \in \mathcal{M}_t} (y_i - y'_i)^2}.
\end{aligned}$$

Regression: RANKTRON

1. Extensions for Agnostic case:

$$\sum_{t \in \mathcal{M}_T} (y_t - y'_t)^2 \leq 8 \inf_v \sum_{t \in \mathcal{M}_v} (y_t - y'_t)^2 + \left(R_x^2 \|v\|_2^2 \right)$$

   Hence competitive!
2. Extension to drifting target
3. No gradient of the loss $\ell(y_t - w_{t-1}^T x_t)$
4. No updates when steady state

Discussion

And

- Visualization, Clustering and EDA
- knowledge Representation
- Learning Logical statements
- Partial Feedback (Active Learning, ...)

Discussion

- What has ML to offer?
- Learning from mistakes
- Learning with Experts
- Opportunities for AeroSpace