Supervised learning

Susanne Still
University of Hawaii at Manoa

Least squares linear regression

- Let's be conservative. "Just fit a line".
- minimize the deviation from the actual observations; get analytical solution for the optimal weights. Can solve using numerical methods (matrix inversion).
- can do this also in a more generalized way by assuming a linear mix of basis functions
- (in this class of methods are SVD, PCA)
- What could possibly go wrong? ....

An entertaining example from *Nature* (2004)

Momentous sprint at the 2156 Olympics?

Women sprinters are closing the gap on men and may one day overtake them.

The 2004 Olympic women's 100-metre sprint champion, Yuliya Nesterenko, is assured of fame and fortune. But we show here that — if current trends continue — it is the winner of the event in the 2156 Olympics whose name will be etched in sporting history forever, because this may be the first occasion on which the race is won in a faster time than the men's event.

The Athens Olympic Games could be viewed as another giant experiment in human athletic achievement. Are women narrowing the gap with men, or falling further behind? Some argue that the gains made by women in running events between the 1930s and the 1980s are decreasing as the women's achievements plateau[1]. Others contend that there is no evidence that athletes, male or female, are reaching the limits of their potential[1,2].

In a limited test, we plot the winning times of the men's and women's Olympic finals over the past 100 years (ref. 3; for data set, see supplementary information) against the competition date (Fig. 1). A range of curve-fitting procedures were tested (for methods, see supplementary information), but there was no evidence that the addition of extra parameters improved the model fit significantly from the simple linear relationships shown here. The remarkably strong linear trends that were first highlighted over ten years ago[2] persist for the Olympic 100-metre sprints. There is no indication that a plateau has been reached by either male or female athletes in the Olympic 100-metre sprint record.

Extrapolation of these trends to the 2008 Olympiad indicates that the women's 100-metre race could be won in a time of 10.57 ± 0.232 seconds and the men's event in 9.73 ± 0.144 seconds. Should these trends continue, the projections will intersect at the 2156 Olympics, when — for the first time ever — the winning women's 100-metre sprint time of 8.079 seconds will be lower than that of the men's winning time of 8.098 seconds (Fig. 1). The 95% confidence intervals, estimated through Markov chain Monte Carlo simulation[4] (see supplementary information), indicate that this could occur as early as the 2064 or as late as the 2788 Games.

This simple analysis overlooks numerous confounding influences, such as timing accuracy, environmental variations, national boycotts and the use of legal and illegal stimulants. But it is also defended by the limited amount of variance that remains unexplained by these linear relationships.

So will these trends continue and can women really close the gap on men? Those who contend that the gender gap is widening say that drug use explains why women's times were improving faster than men's, particularly as that improvement slowed after the introduction of drug testing[1]. However, no evidence for this is found here. By contrast, those who maintain that there could be a continuing decrease in gender gap point out that only a minority of the world's female population has been given the opportunity to compete (O. Anderson, www.pponline.co.uk/encyc/0151.htm).

Whether these trends will continue at the Beijing Olympics in 2008 remains to be seen. Sports, biological and medical sciences should enable athletes to continue to improve on Olympic and world records, by fair means or foul[5]. But only time will tell whether in the 66th Olympiad the fastest human on the planet will be female.

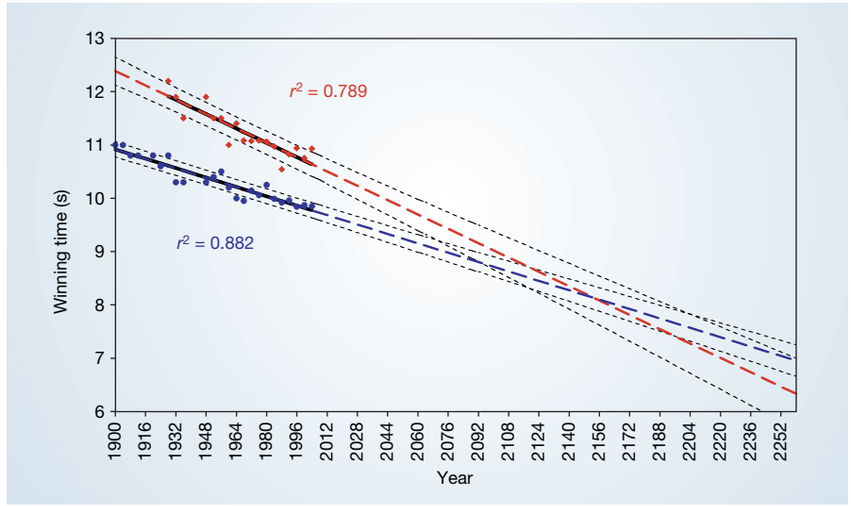


Figure 1 The winning Olympic 100-metre sprint times for men (blue points) and women (red points), with superimposed best-fit linear regression lines (solid black lines) and coefficients of determination. The regression lines are extrapolated (broken blue and red lines for men and women, respectively) and 95% confidence intervals (dotted black lines) based on the available points are superimposed. The projections intersect just before the 2156 Olympics, when the winning women's 100-metre sprint time of 8.079 s will be faster than the men's at 8.098 s.

Andrew J. Tatem*, Carlos A. Guerra*, Peter M. Atkinson†, Simon I. Hay*‡

**TALA Research Group, Department of Zoology, University of Oxford, Oxford OX1 3PS, UK*
e-mail: andy.tatem@zoology.oxford.ac.uk
†*School of Geography, University of Southampton, Highfield, Southampton SO17 1BJ, UK*
‡*Public Health Group, KEMRI/Wellcome Trust Research Laboratories, PO Box 43640, 00100 GPO, Nairobi, Kenya*

1. Holden, C. *Science* **305,** 639–640 (2004).
2. Whipp, B. J. & Ward, S. A. *Nature* **355,** 25 (1992).
3. Rendell, M. (ed.) *The Olympics: Athens to Athens 1896–2004* 338–340 (Weidenfeld and Nicolson, London, 2003).
4. Gilks, W. R., Thomas, A. & Spiegelhalter, D. J. *Statistician* **43,** 169–178 (1994).
5. Vogel, G. *Science* **305,** 632–635 (2004).

Supplementary information accompanies this communication on *Nature*'s website.
Competing financial interests: declared none.

Lung cancer

Intragenic ERBB2 kinase mutations in tumours

The protein-kinase family is the most frequently mutated gene family found in human cancer and faulty kinase enzymes are being investigated as promising targets for the design of antitumour therapies. We have sequenced the gene encoding the transmembrane protein tyrosine kinase ERBB2 (also known as HER2 or Neu) from 120 primary lung tumours and identified 4% that have mutations within the kinase domain; in the adenocarcinoma subtype of lung cancer, 10% of cases had mutations. ERBB2 inhibitors, which have so far proved to be ineffective in treating lung cancer, should now be clinically re-evaluated in the specific subset of patients with lung cancer whose tumours carry *ERBB2* mutations.

The successful treatment of chronic myelogenous leukaemia with a drug (known as imatinib, marketed as Gleevec) that inhibits a mutant protein kinase has fostered interest in the development of other kinase inhibitors[1]. Gefitinib, an inhibitor of the epidermal growth-factor receptor (EGFR), induces a marked response in a small subset of lung cancers; activating mutations have been found in the *EGFR* gene in tumours that respond to gefitinib but are rare in those that do not respond[2,3]. The response to gefitinib as a treatment for lung cancer therefore seems to be predicated upon the presence of an *EGFR* mutation in the tumour.

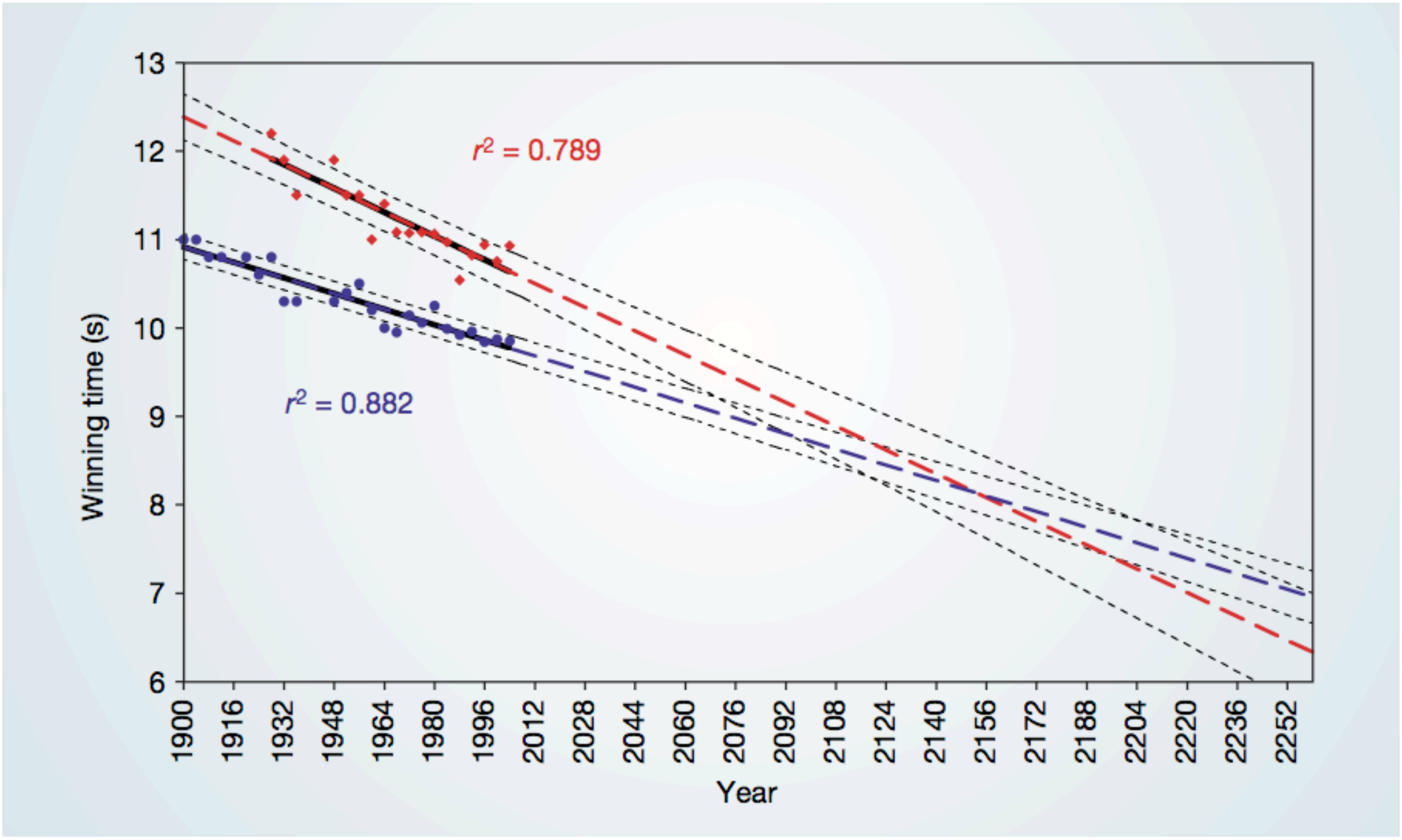


Figure 1 The winning Olympic 100-metre sprint times for men (blue points) and women (red points), with superimposed best-fit linear regression lines (solid black lines) and coefficients of determination. The regression lines are extrapolated (broken blue and red lines for men and women, respectively) and 95% confidence intervals (dotted black lines) based on the available points are superimposed. The projections intersect just before the 2156 Olympics, when the winning women's 100-metre sprint time of 8.079 s will be faster than the men's at 8.098 s.

English humor saves the day...

Sprint research runs into a credibility gap

Sir — A. J. Tatem and colleagues calculate that women may outsprint men by the middle of the twenty-second century (*Nature* **431,** 525; 2004). They omit to mention, however, that (according to their analysis) a far more interesting race should occur in about 2636, when times of less than zero seconds will be recorded.

In the intervening 600 years, the authors may wish to address the obvious challenges raised for both time-keeping and the teaching of basic statistics.

Kenneth Rice

MRC Biostatistics Unit, Institute of Public Health, Forvie Site, Robinson Way, Cambridge CB2 2SR, UK

Another motivating example

- Portfolio selection: how to invest your money?
- Combination of N assets with different relative weights.
- Question: How to spread the money across the different assets.
- If you know their returns, then can you find an optimal portfolio?!

Too lazy to worry about it?

- Give your money to the bank
- Then they do this for you (not alone for you, but they collect all the money and then decide how to invest it)
- So, you can't really get away from the problem
- We are interested the **risk in the context of large portfolios** (banks, insurances, etc.)

Mathematically

- Measure returns x_i^k for assets i=1, ...,N at k=1, ...,T time points.
- Portfolio is a linear combination $\vec{w} \cdot \vec{x}$ with weights w_i that fulfill the budget constraint $\sum_i w_i = 1$
- Given a risk functional $F(\vec{w} \cdot \vec{x})$
- Portfolio weights are chosen such that the risk $R(\vec{w}) = \langle F(\vec{w} \cdot \vec{x}) \rangle_{p(\vec{x})}$ is minimized.
- Simplify: to study risk, *set return to zero.*

Example: Markowitz

- Risk = variance: $F = \frac{1}{2}(\vec{w}\vec{x})^2$

$$\min_{\vec{w}} \left\langle \frac{1}{2}(\vec{w}\vec{x})^2 \right\rangle_{p(\vec{x})} \quad \text{s.t.} \quad \sum_i w_i = 1$$

- Optimal solution: $w_i^* = \frac{\sum_j \sigma_{ij}^{-1}}{\sum_{j,k} \sigma_{kj}^{-1}}$

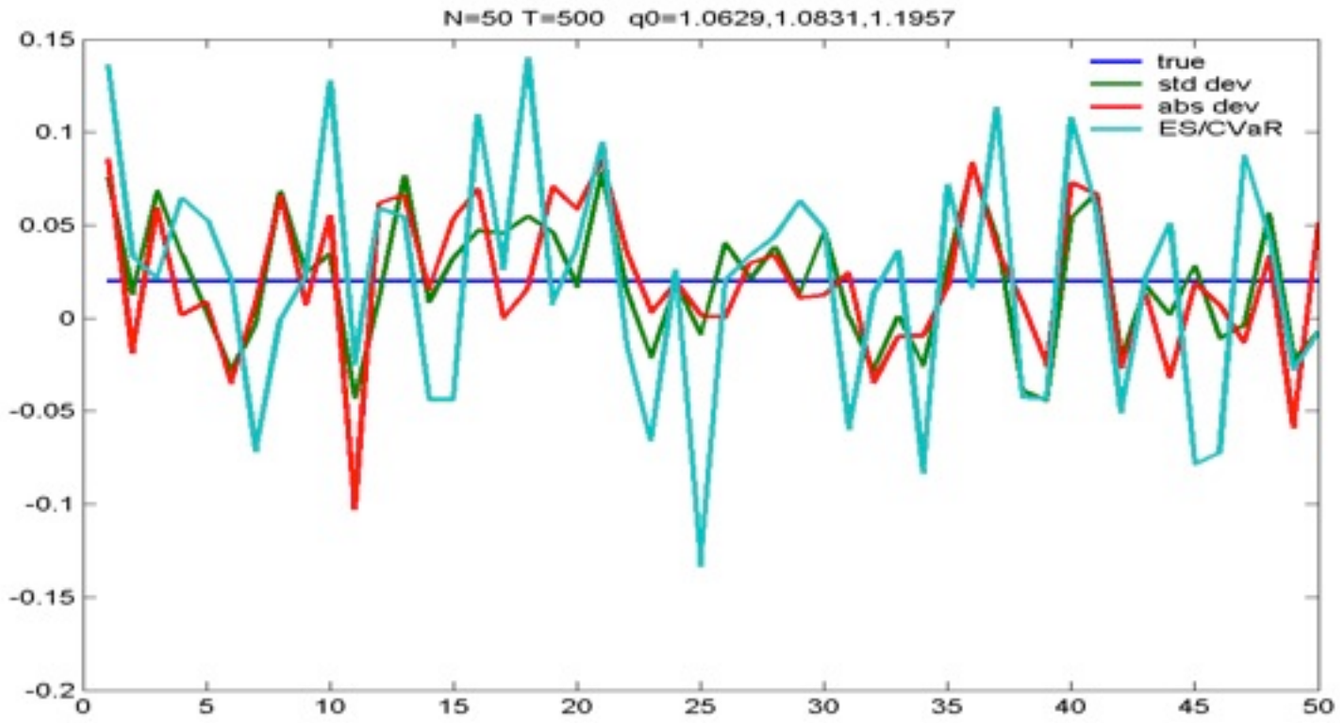
- with covariance $\sigma_{ij} = \langle x_i x_j \rangle$
- In practice, have to use: $\hat{\sigma}_{ij} = \frac{1}{T} \sum_k x_i^{(k)} x_j^{(k)}$

Other risk measures

- Mean Absolute Deviation (MAD)
- Value at Risk (VaR): high quantile - threshold below which a given percentage of the weight of the profit-loss distribution resides. NOT CONVEX.
- Expected Shortfall (ES): the conditional average over a high quantile.
- Maximal Loss (ML): the extreme case of ES, the optimal combination of the worst outcomes.
- ***Coherent risk measures***: monotonic, sub-additive, positive homogeneous, and translationally invariant.
- *ES and ML are coherent.*
  VaR, ES, and ML are downside risk measures (not bounded from below).

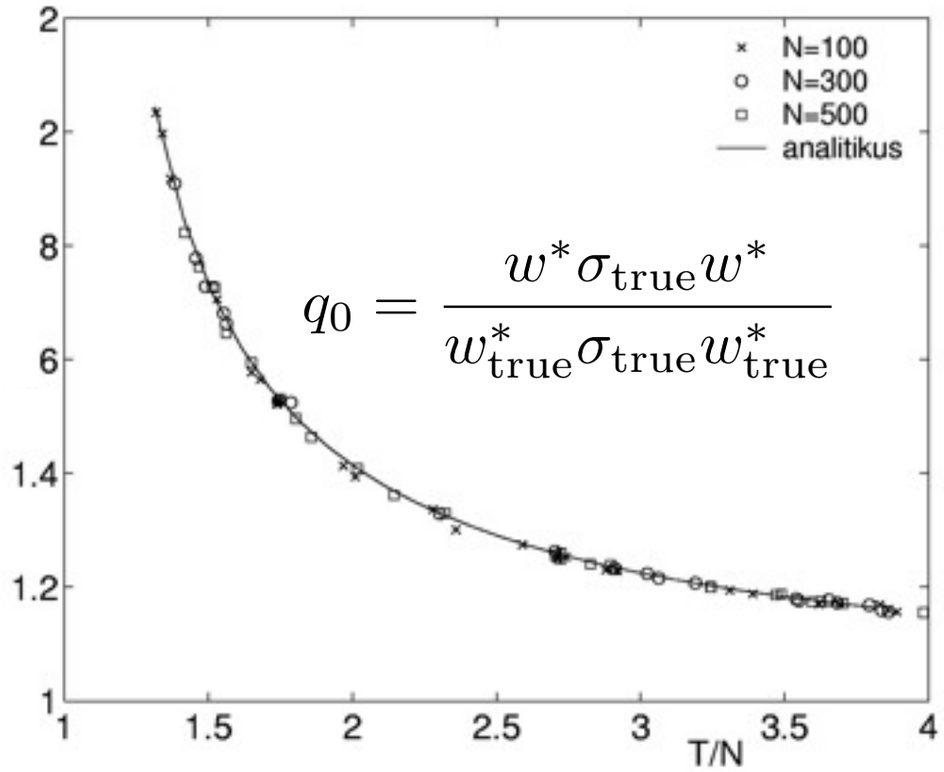
Instability

- For large portfolios the weights are far from optimal.
- Weights fluctuate due to sampling errors - have T samples to estimate risk functional.



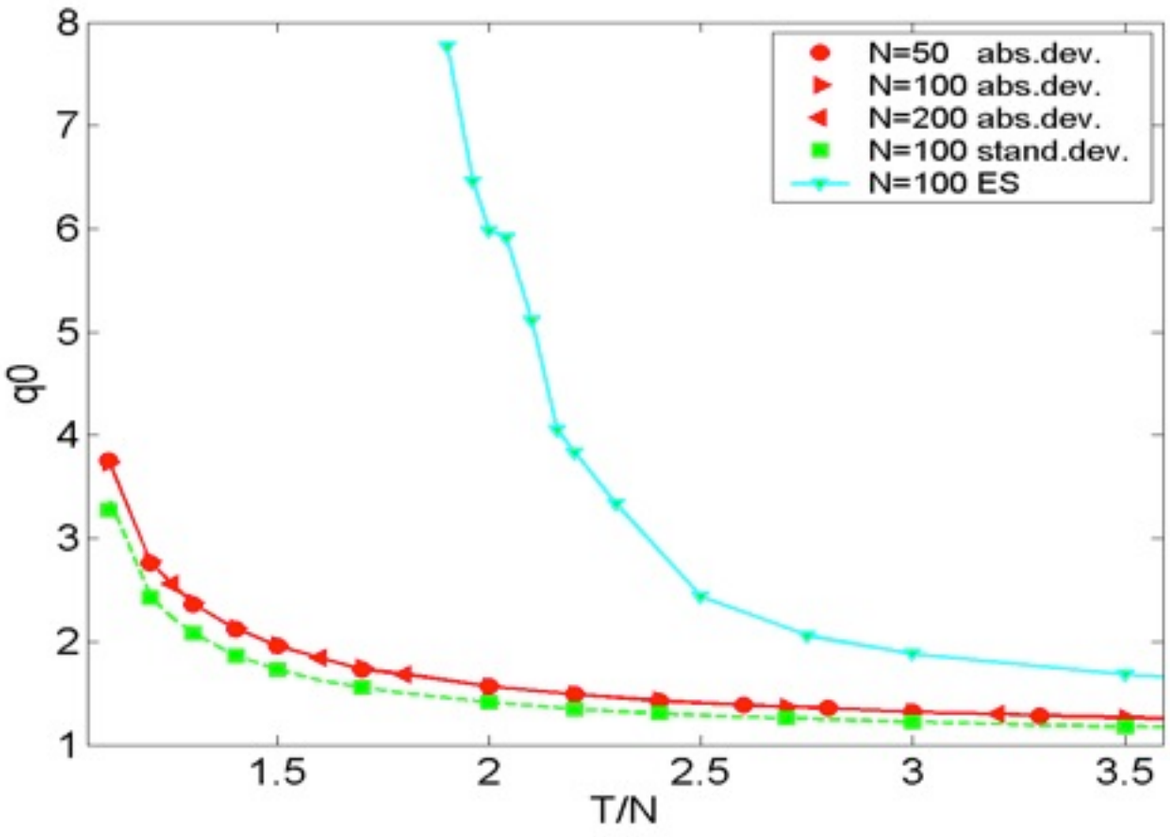
Figures courtesy of I. Kondor

- The resulting estimation error diverges at a critical value of the ratio N/T.
  Measure: q_0 =ratio of estimated vs true



Instability persists

- Found across ***all tested risk measures***
- Persists when linear constraints are added



Kondor and coworkers
2003-2008

Figure courtesy of I. Kondor

What went wrong?

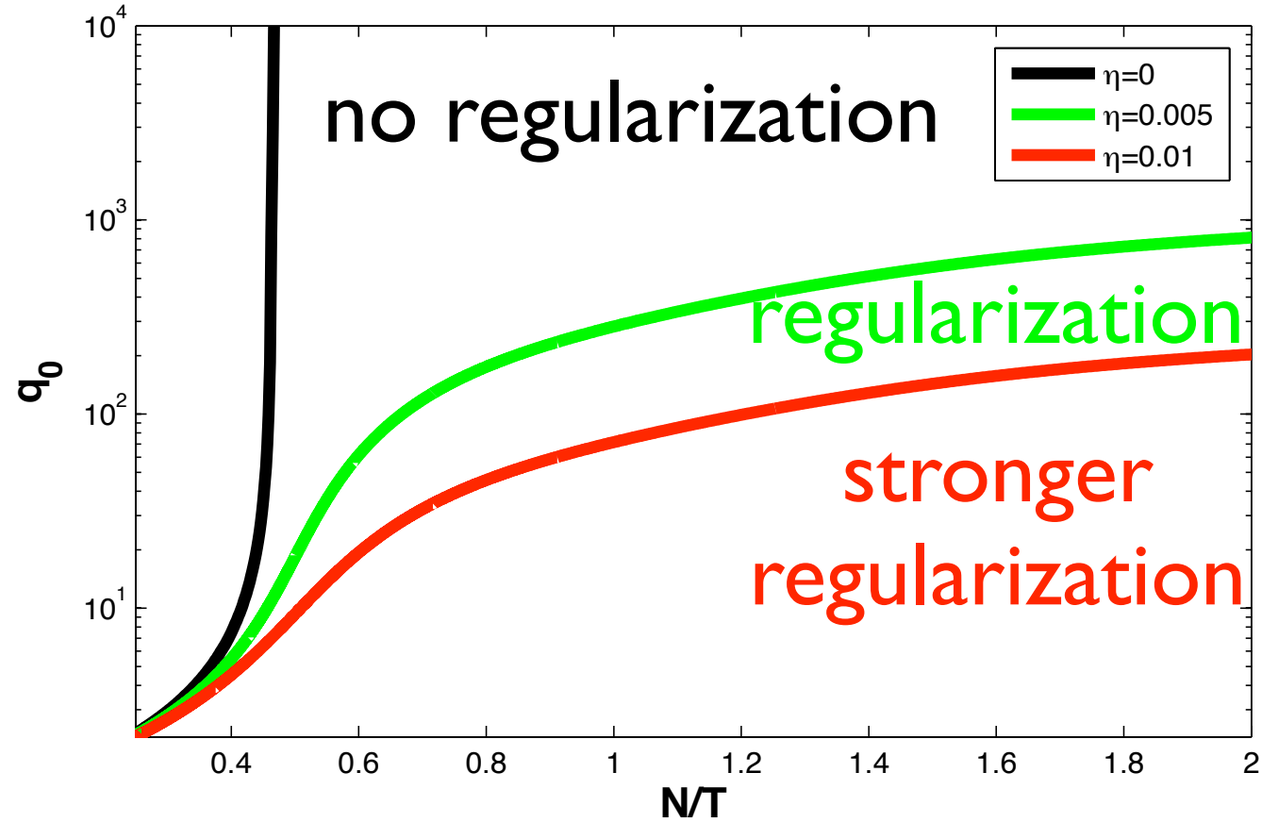
- We based our estimate of the risk that a portfolio of M assets has on the measurement of N data points.
- We then minimized the empirical risk.
- Problem: Small empirical risk guarantees small actual risk only in the limit in which N >> M. But we are not in that limit.
- We are over-fitting the data!

How to fix the problem?

- Statistical learning theory finds bounds on the difference between actual risk and empirical risk. Then minimizes the bound. Bound is monotonic in capacity.
  => limiting the capacity leads to better generalization. (= regularization)
- Bayesian inference uses the prior to achieve regularization.
- For the finance problem: Regularize portfolio optimization!

Regularized Portfolio Optimization

- Does this change things?
- Yes! The divergence disappears
- (analytical calculation: gaussian returns; replica trick) more in:



Regularizing Portfolio Optimization, S. Still and I. Kondor, *New Journal of Physics*, 12, 075034, 2010 **Optimal liquidation strategies regularize portfolio selection**, *F. Caccioli*, S. Still, M. Marsili, and I. Kondor, *The European Journal of Finance*, DOI:10.1080/1351847X.2011.601661, 2011
Liquidity risk and instabilities in portfolio optimisation, F. Caccioli, I. Kondor, M. Marsili, and S. Still, *International Journal of Theoretical and Applied Finance* (2016)

Introduction to statistical learning theory and support vector machines

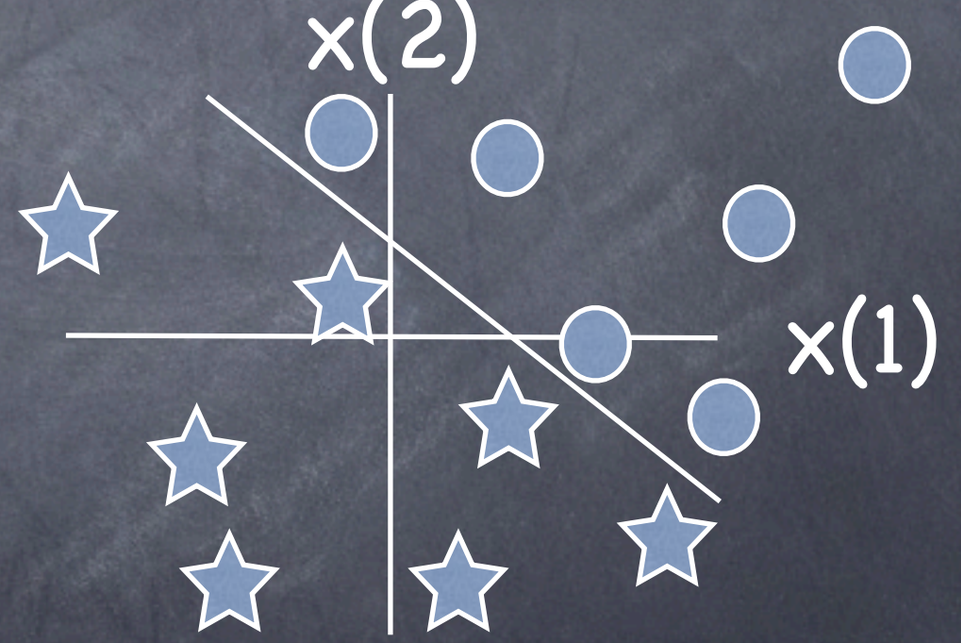
Support Vector Learning

This lecture follows, and the figures are from: Advances in Kernel Methods-Support Vector Learning (Chapter 1) B. Schoelkopf, C. Burges, A. Smolla (eds.), MIT Press, Cambridge, MA1999

- VC dimension: A measure for the capacity of a learning machine
- Structural Risk Minimization
- Linear Classifiers
- Feature Space and Kernel functions
- Support Vector Machines
- Noisy Data
- Support Vector Regression

Recall: Binary Classification

- given N inputs $\{\vec{x}_i, y_i\}_{i=1,\ldots,N}$
- with $\vec{x}_i \in R^d$
- and $y_i \in \{-1; 1\}$
- we are looking for a function f, which correctly classifies new inputs:
  i.e. $f(\vec{x}) = y$



Recall: The Perceptron

- Linear classifier
- Limitations:
  - are we finding the best separating decision boundary? (the one that gives best generalization)
  - how do we deal with data that is not linearly separable?

Addressing the limitations:

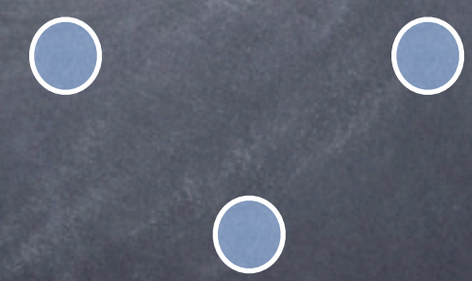
- Find best separating hyper plane using a large margin classifier.
- Deal with non-linear data via a feature map; implicit in the kernel function.
- Decide on the function class using structural risk minimization (SRM)

Capacity of a learning machine

- Learning machine: implements a function class, or hypothesis class with a certain capacity, and learns the particular function out of that class that best fits the data.
- Statistical learning theory: Measure the capacity via the Vapnik-Chervonenkis (VC) dimension.

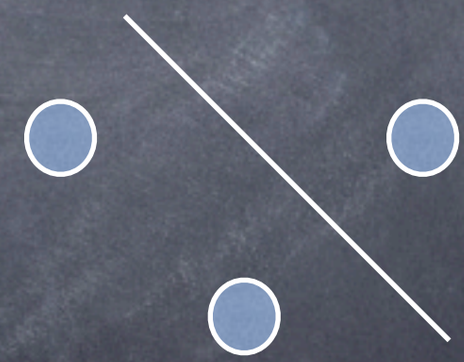
VC dimension

- The largest number h of points that can be separated **in all possible ways**, using functions of the given class.
- Allows one to find a bound on the generalization error.
- Linear models: h = d+1
  where d = dimension



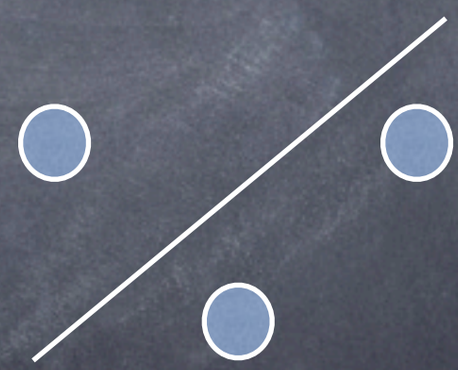
VC dimension

- The largest number h of points that can be separated **in all possible ways**, using functions of the given class.
- Allows one to find a bound on the generalization error.
- Linear models: h = d+1
  where d = dimension



VC dimension

- The largest number h of points that can be separated **in all possible ways**, using functions of the given class.
- Allows one to find a bound on the generalization error.
- Linear models: h = d+1
  where d = dimension

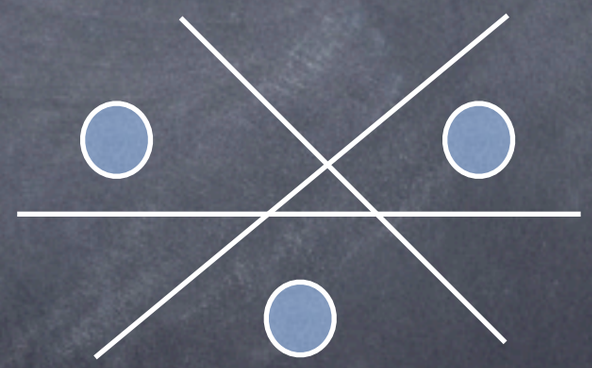


VC dimension

- The largest number h of points that can be separated **in all possible ways**, using functions of the given class.
- Allows one to find a bound on the generalization error.
- Linear models: h = d+1
  where d = dimension

VC dimension

- The largest number h of points that can be separated **in all possible ways**, using functions of the given class.
- Allows one to find a bound on the generalization error.
- Linear models: h = d+1
  where d = dimension



- Other alternative concepts for finding (tighter) error bounds: VC entropy, growth function, fat shattering dimension...

Generalization

- Generalization error ("risk"):

$$R[f] = \langle \frac{1}{2} |f(\vec{x}) - y| \rangle_{P(\vec{x},y)}$$

- Can only measure the training error ("empirical risk"):

$$R_{\mathrm{emp}}[f] = \frac{1}{2N} \sum_{i=1}^{N} |f(\vec{x}_i) - y_i|$$

- Those can differ. The generalization error can be larger than the training error:

$$R[f] = R_{\mathrm{emp}}[f] + \Phi$$

confidence

Error Bound

- With probability $1 - \eta$:

$$R[f] \leq R_{\text{emp}}[f] + \sqrt{\frac{h(\log \frac{2N}{h} + 1) - \log \frac{\eta}{4}}{N}}$$

can minimize this!

- Empirical risk depends on the function that the learning machine learns and can be minimized by choosing the right function out of a set of functions = Hypothesis class

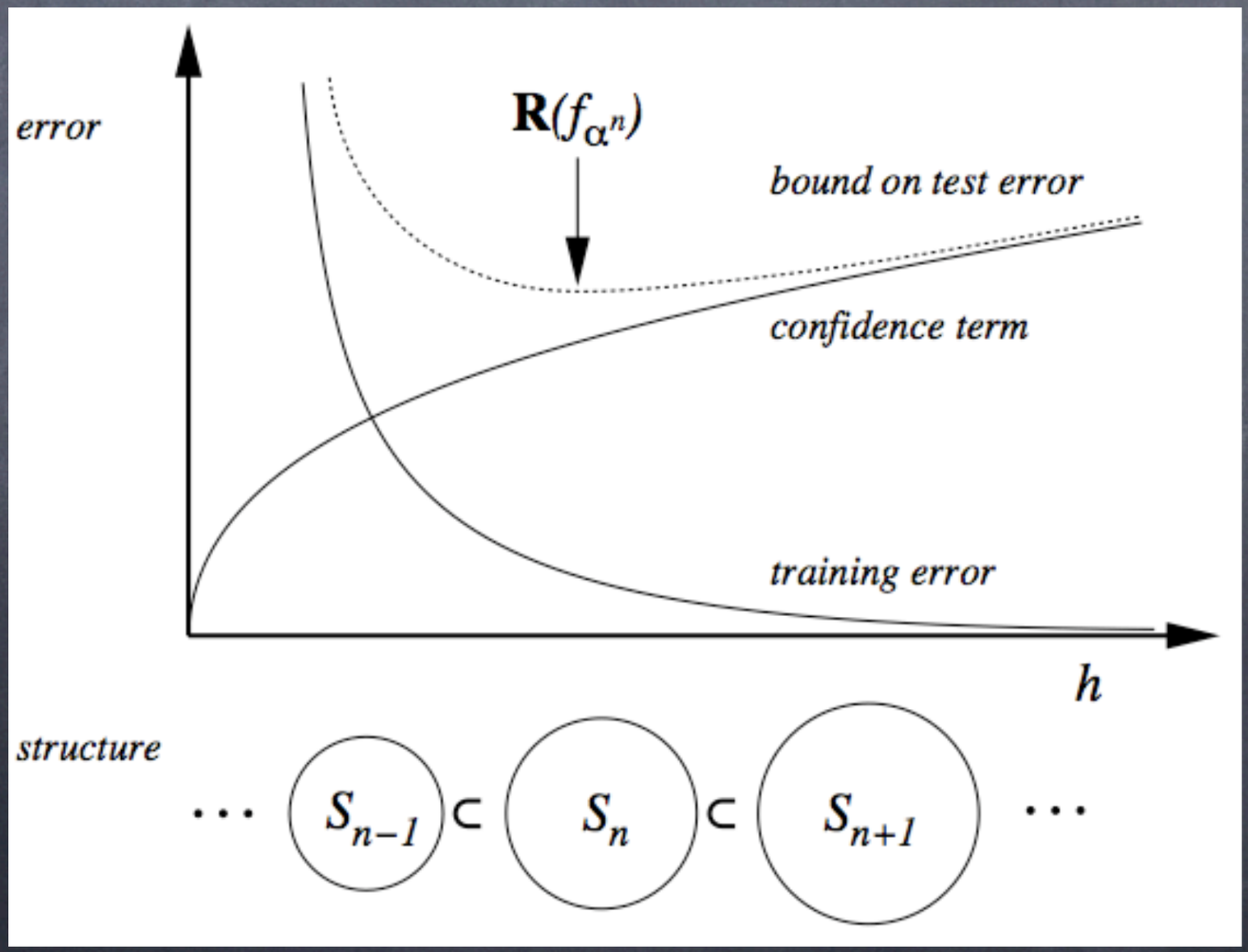
- Confidence term depends on the complexity of the Hypothesis class

$$R[f] \leq R_{\text{emp}}[f] + \frac{1}{\sqrt{N}} \sqrt{h\left(\log \frac{2N}{h} + 1\right) - \log \frac{\eta}{4}}$$

- consider N not large
- note: confidence term increases monotonically with h
  - imagine we have training data that is labeled at random (there is no structure)
  - learn function with low training error
  - must have large VC dimension to reproduce the random labels -> large confidence term
- thus we can not always expect small generalization error due to small training error

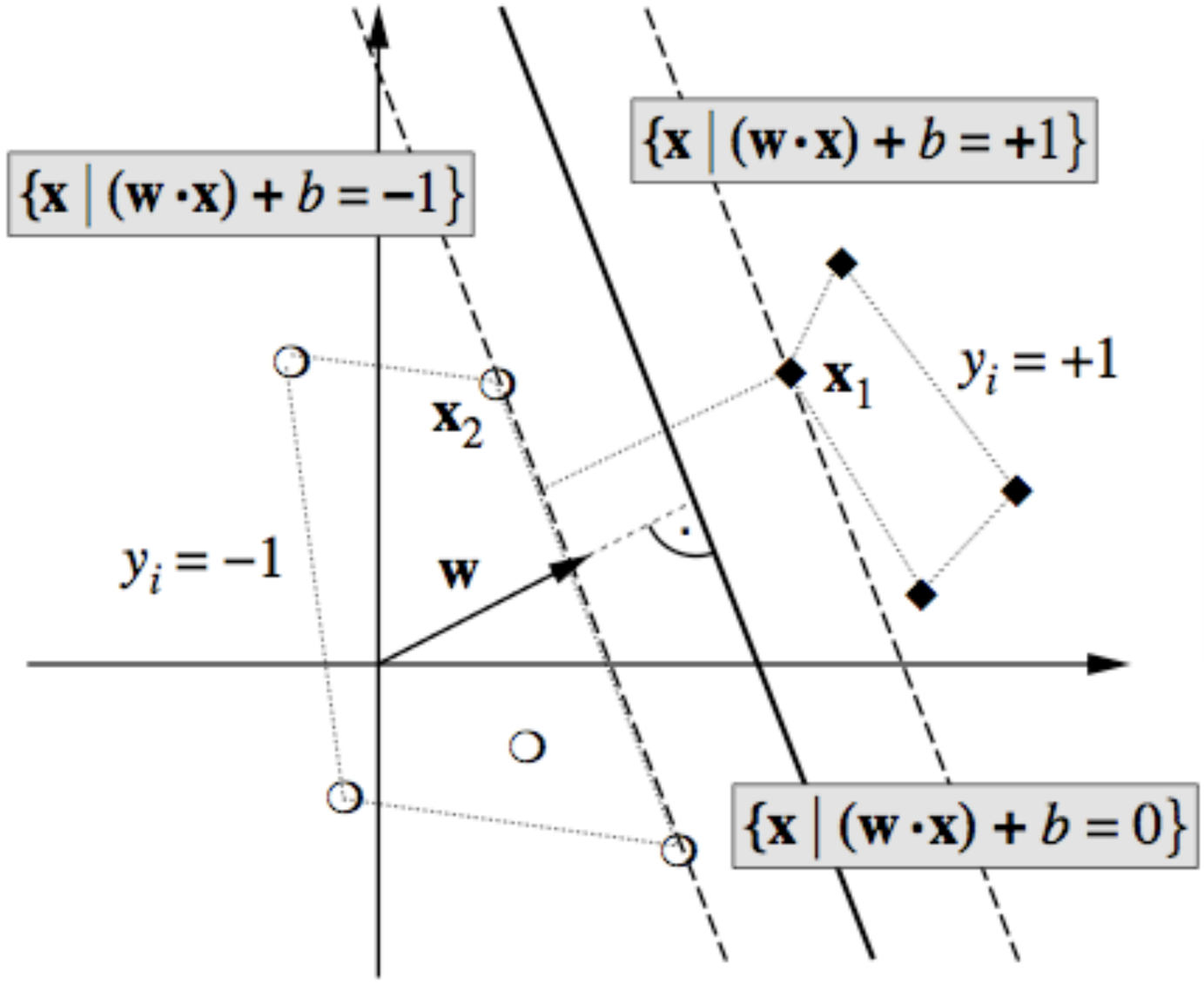
Controlling Complexity: Structural risk minimization

- Use a structure of nested subsets with increasing VC-dimension



Linear classifiers

- Vapnik and Chervonenkis (1964)
- separating hyperplane: $\vec{w} \cdot \vec{x} + b = 0$
- decision function: $f(\vec{x}) = sgn(\vec{w} \cdot \vec{x} + b)$
- unique separating hyperplane exists with maximum margin of separation (distance to nearest example point)
- the capacity decreases with increasing margin!



Note:

$$(\mathbf{w}\cdot\mathbf{x}_1) + b = +1$$
$$(\mathbf{w}\cdot\mathbf{x}_2) + b = -1$$
$$\Rightarrow \quad (\mathbf{w}\cdot(\mathbf{x}_1 - \mathbf{x}_2)) = 2$$
$$\Rightarrow \left(\frac{\mathbf{w}}{\|\mathbf{w}\|}\cdot(\mathbf{x}_1 - \mathbf{x}_2)\right) = \frac{2}{\|\mathbf{w}\|}$$

From: Advances in Kernel Methods-Support Vector Learning (Chapter 1) B. Schoelkopf, C. Burges, A. Smolla (eds.), MIT Press, Cambridge, MA1999

- re-scale: |**wx** + b| = 1 for the points closest to the hyper-plane.
- Then, the margin = 2 / ||w||

Learning algorithm

- Make ||w|| small!

$$\begin{aligned} &\text{minimize} && \frac{1}{2}\|\vec{w}\|^2 \\ &\text{subject to} && y_i(\vec{w}\cdot\vec{x}_i + b) \geq 1 \quad \forall i \end{aligned}$$

- Use Lagrange multipliers:

$$L = \frac{1}{2}\|\vec{w}\|^2 - \sum_i \alpha_i \big(y_i(\vec{w}\cdot\vec{x}_i + b) - 1\big)$$

$$L = \frac{1}{2}\|\vec{w}\|^2 - \sum_i \alpha_i \big(y_i(\vec{w}\cdot\vec{x}_i + b) - 1\big)$$

- has to be minimized w.r.t. primary variables **w** and b
- has to be maximized w.r.t. dual variables α_i
- Find "saddle point".
- For all constraints which are not met precisely, the α_i must be zero (Karush-Kuhn-Tucker (KKT) complementary conditions), because that maximizes the dual.

- Saddle point conditions:

$$\frac{\partial}{\partial b} L(\vec{w}, b, \vec{\alpha}) = 0; \;\; \frac{\partial}{\partial \vec{w}} L(\vec{w}, b, \vec{\alpha}) = 0$$

- imply that at the optimum:

$$\sum_i \alpha_i y_i = 0$$

$$\vec{w} = \sum_i \alpha_i y_i \vec{x}_i$$

=> The optimal separating hyperplane can be written solely in terms of those points for which $\alpha_i \neq 0$, i.e for which

$$y_i(\vec{w} \cdot \vec{x}_i + b) = 1$$

- those are the points that lie on the margin! They are called "Support Vectors"

Dual representation

$$\begin{aligned} \text{maximize} \quad & \sum_i \alpha_i - \frac{1}{2} \sum_{ij} \alpha_i \alpha_j y_i y_j \vec{x}_i \cdot \vec{x}_j \\ \text{subject to} \quad & \alpha_i \geq 0,\ i = 1, \ldots, N,\ \text{and} \ \sum_i \alpha_i y_i = 0 \end{aligned}$$

Decision boundary

$$f(\vec{x}) = sgn\left(\sum_i \alpha_i y_i \vec{x} \cdot \vec{x}_i + b\right)$$

calculate b from: $y_i(\vec{w} \cdot \vec{x}_i + b) = 1$

- Never solve a problem that is harder than the one you need to solve (Vapnik) - Here: not modeling full distribution, just finding best separation line.

Physical interpretation

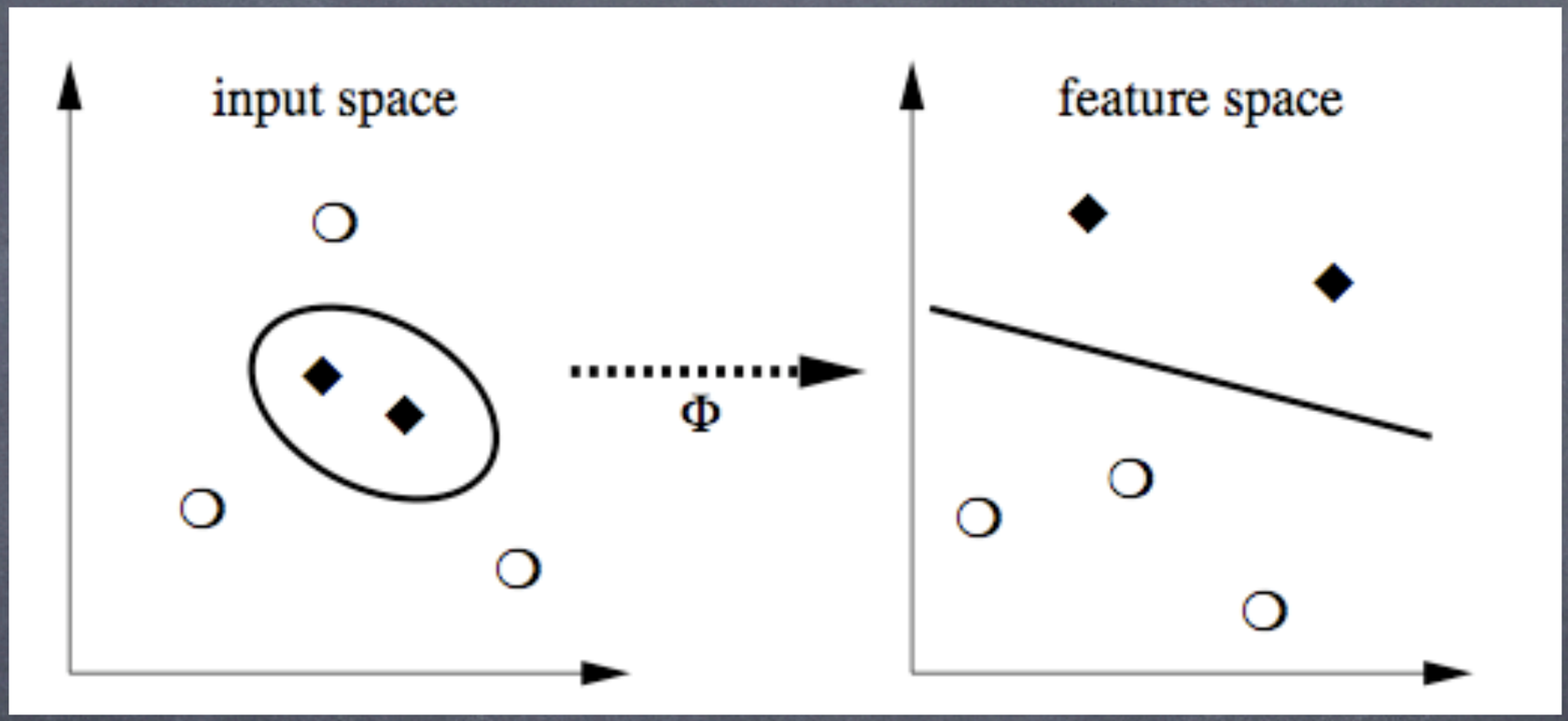
- Assume that each support vector exerts a force on the hyperplane (imagine a movable sheet there) in the direction of y_i, and with a magnitude α_i
- Solution <=> Mechanical Stability:
- Forces sum to zero: $\sum_i \alpha_i y_i = 0$
- Torques sum to zero:

$$\sum_i \alpha_i y_i \vec{x}_i \times \frac{\vec{w}}{\|\vec{w}\|} = \frac{\vec{w} \times \vec{w}}{\|\vec{w}\|} = 0$$

Feature Space

- Addresses the problem: Data is not linearly separable in input domain
- Map data into a feature space via a nonlinear feature map, in which the data is linearly separable.
- Find separating hyperplane in feature space
- Project onto input space -> decision boundary

Feature Space



Feature Space

- How to find the feature map Φ?
- This is as hard as solving the original problem of separating nonlinear data
- The cool thing:
- optimal hyperplane and decision boundary require only the evaluation of dot products of point
- That means, we never need to know $\Phi(\vec{x})$, only need $\Phi(\vec{x}) \cdot \Phi(\vec{y})$

Mercer Kernel

- Can introduce a kernel function: $k(\vec{x}, \vec{y})$
- kernels of positive integral operators give rise to maps, such that

$$k(\vec{x}, \vec{y}) = \Phi(\vec{x}) \cdot \Phi(\vec{y})$$

- (Theorem, Mercer 1909)

Kernels used (examples)

- polynomial $k(\vec{x}, \vec{y}) = (\vec{x} \cdot \vec{y})^d$
- sigmoidal $k(\vec{x}, \vec{y}) = \tanh(\beta \vec{x} \cdot \vec{y} + \gamma)$
- radial-basis (gaussian) $k(\vec{x}, \vec{y}) = e^{(-\frac{\|\vec{x}-\vec{y}\|^2}{2\sigma^2})}$
- string kernel (for text and genes)

Support Vector Machines

- Compute optimal hyperplane in feature space substitute $\Phi(\vec{x}_i)$ for $\vec{x}_i$, and mercer kernel for the dot product: $k(\vec{x}_i, \vec{x}_j) = \Phi(\vec{x}_i) \cdot \Phi(\vec{x}_j)$
- => decision function becomes

$$f(\vec{x}) = sgn\Big(\sum_i \alpha_i y_i k(\vec{x}_i, \vec{x}_j) + b\Big)$$

- quadratic optimization program

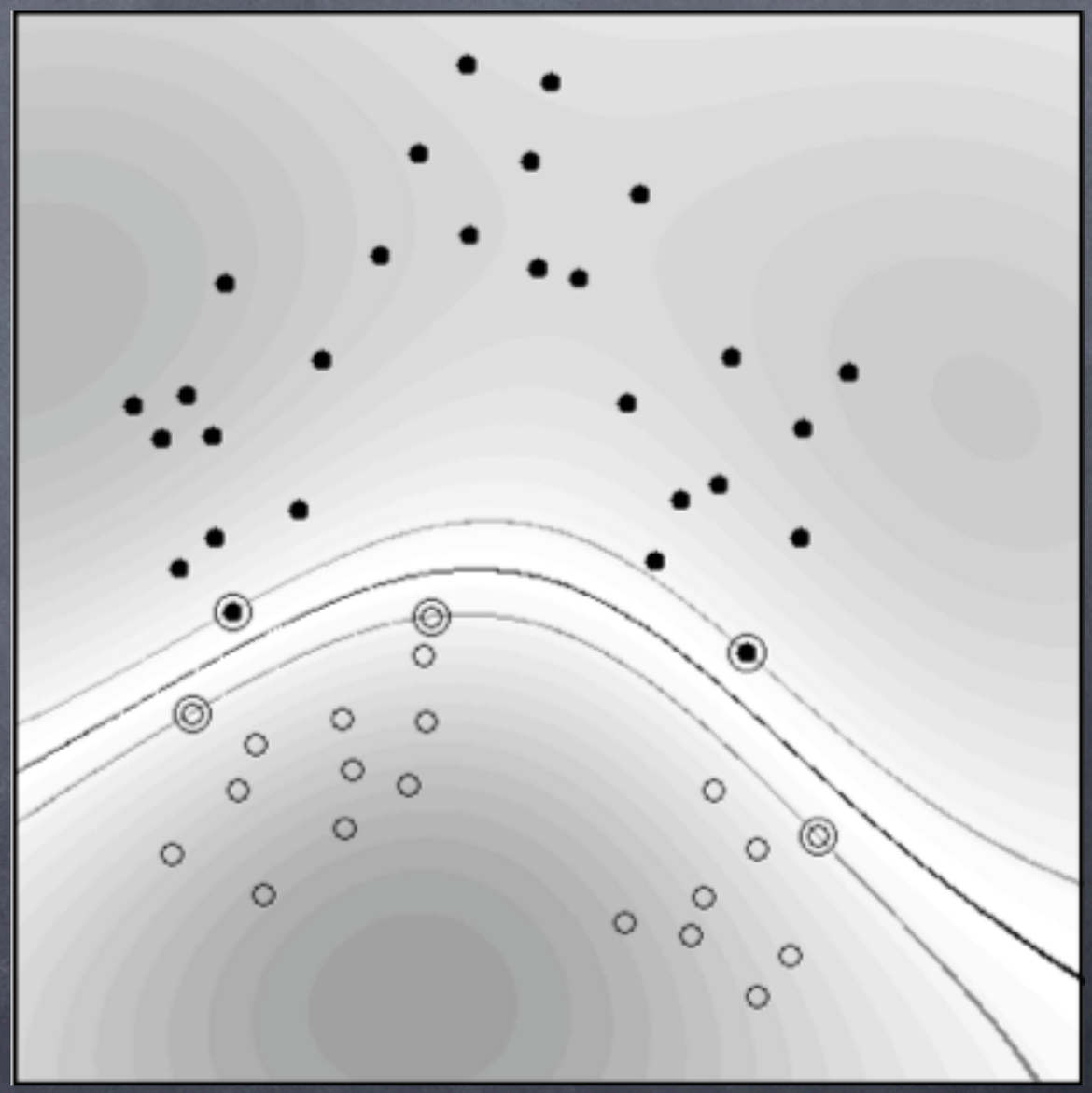
$$\begin{aligned} \text{maximize} \quad & \textstyle\sum_i \alpha_i - \frac{1}{2} \sum_{ij} \alpha_i \alpha_j y_i y_j k(\vec{x}_i, \vec{x}_j) \\ \text{subject to} \quad & \alpha_i \geq 0,\ i = 1, \ldots, N, \text{ and } \textstyle\sum_i \alpha_i y_i = 0 \end{aligned}$$

SVM Example

- Radial-basis function kernel

$$k(\vec{x}, \vec{y}) = e^{(-\|\vec{x}-\vec{y}\|^2)}$$

- Gray scale reflects argument of the decision function



From: B. Schoelkopf, A. Smola, K.-R. Mueller (1996) Technical Report 44, Max Planck Institue for Biological Cybernetics, Tuebingen, Germany

Noisy data

- In practice, data sets are usually corrupted by noise.
- This may cause the classes to overlap
- => no separating hyperplane exists
- How do we deal with noise?

Soft Margin

- allow for the possibility that examples violate the constraints by introducing "slack variables" $\xi_i \geq 0$
- relax the constraints $y_i(\vec{w} \cdot \vec{x}_i + b) \geq 1 - \xi_i$
- => new objective function: control the capacity by maximizing the margin (as before, via minimization of ||w||), and minimize the margin error. C controls the trade-off.

$$\frac{1}{2}\|\vec{w}\|^2 + C\sum_i \xi_i$$

Support Vector Regression

- Idea: construct the analog to the margin
- ϵinsensitive loss: $|y - f(\vec{x})|_\epsilon = \max\{0; |y - f(\vec{x})| - \epsilon\}$
- To estimate a linear regression with precision ϵ, one minimizes:

$$\frac{1}{2}\|\vec{w}\|^2 + C\sum_i |y_i - f(\vec{x}_i)|_\epsilon$$

Support Vector Regression

- Fit a tube with radius ϵ to the data.
- Introduce slack variables as before, solve the program:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|\vec{w}\|^2 + C\sum_i(\xi_i + \xi_i^*) \\ \text{subject to} \quad & (\vec{w}\cdot\vec{x}_i + b) - y_i \leq \epsilon + \xi_i \\ & y_i - (\vec{w}\cdot\vec{x}_i + b) \leq \epsilon + \xi_i^* \\ & \xi_i,\ \xi_i^* \geq 0 \end{aligned}$$

- ϵ and C are chosen a priori

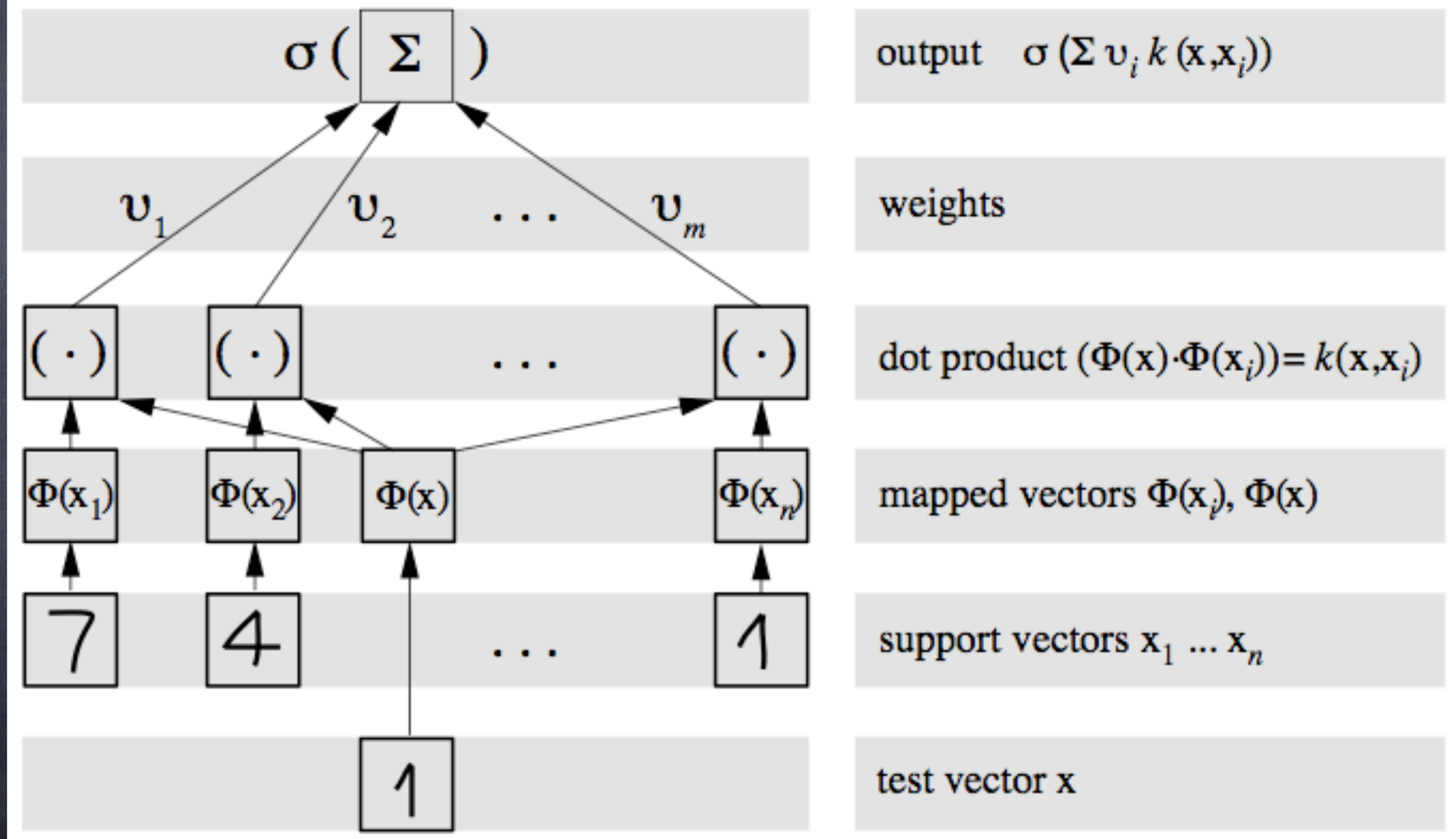
Support Vector Regression

- As before, use Lagrange multipliers, compute dual.
- Deal with nonlinear regression by introducing a kernel function (analogous to classification)
- Eventually, the regression function is given by:

$$f(\vec{x}) = \sum_i (\alpha_i^* - \alpha_i) k(\vec{x}_i, \vec{x}) + b$$

Sketch of SVM Architecture

Application: Character recognition



Some Applications

- OCR: Optical Character Recognition
- Object recognition
- Outlier detection (jet engines, email spam, ...)
- Data and text classification
- Bioinformatics

Other Kernel Machines

- Kernel PCA (Principle Component Analysis)
- Kernel CCA (Canonical Correlation Analysis)
- Kernel K-means
- Some methods are interpretable as kernel methods, such as ICA (Independent Component Analysis), Fisher discriminant, and several clustering algorithms, and more...