Introduction to Statistical Learning Theory

Petra Philips

[1] Friedrich Miescher Laboratory, Tübingen

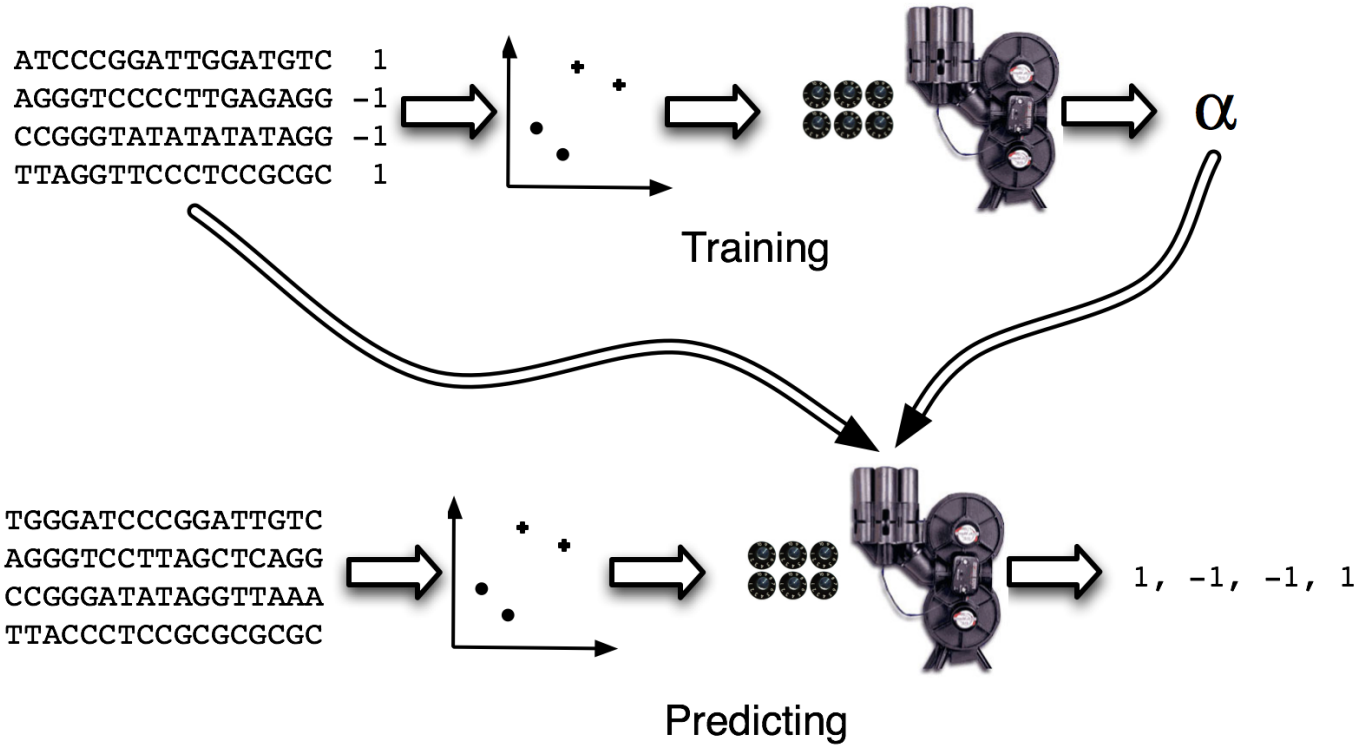
Vorlesung WS 2007/2008

Eberhard-Karls-Universität Tübingen

27 November 2007

http://www.fml.mpg.de/raetsch/lectures/amsa07

Retrospection



Supervised Learning in a Nutshell

MAX-PLANCK-GESELLSCHAFT

Given

Training data: A finite set of examples $\mathbf{x}_i \in \mathcal{X}$ and their associated labels $y_i \in \mathcal{Y}$.

Wanted

The 'best' estimator modelling the relationship between the $\mathbf{x}_i$ and the associated labels y_i, i.e. the 'best' function

$$h : \mathcal{X} \to \mathcal{Y}.$$

Approach

- Restrict possible functions (e.g. hyperplanes).
- Quantify 'best' as the optimum of some computable objective function (usually error on training data).
- Evaluate prediction performance on new test data.

Challenge

Is there an a priori way to guarantee good performance?

Probabilistic Learning Model

Assumption

All data is generated by the same hidden probabilistic source!

Formally

- p is an unknown joint probability distribution over $\mathcal{X} \times \mathcal{Y}$
- Training data $((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n))$ is iid $\sim \mathrm{p}$
- Loss: the error for a particular example $\ell(h(\mathbf{x}_i), y_i)$.
- Aim: find best h^{**} that minimizes risk

$$\mathbf{L}(h) = \mathbf{E}_{\mathbb{X} \times \mathbb{Y}}\, l(\mathbb{Y}, h(\mathbb{X})) = \int \ell(h(\mathbf{x}), y) d\mathrm{p}.$$

Probabilistic Learning Model

Assumption

All data is generated by the same hidden probabilistic source!

Formally

- p is an unknown joint probability distribution over $\mathcal{X} \times \mathcal{Y}$
- Training data $((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n))$ is iid $\sim \mathrm{p}$
- Loss: the error for a particular example $\ell(h(\mathbf{x}_i), y_i)$.
- Aim: find best $h^* \in \mathcal{H}$ that minimizes risk

$$\mathbf{L}(h) = \mathbf{E}_{\mathbb{X} \times \mathbb{Y}}\, l(\mathbb{Y}, h(\mathbb{X})) = \int \ell(h(\mathbf{x}), y) d\mathrm{p}.$$

Probabilistic Learning Model

Assumption

All data is generated by the same hidden probabilistic source!

Formally

- p is an unknown joint probability distribution over $\mathcal{X} \times \mathcal{Y}$
- Training data $((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n))$ is iid $\sim \mathrm{p}$
- Loss: the error for a particular example $\ell(h(\mathbf{x}_i), y_i)$.
- Aim: find best $h^* \in \mathcal{H}$ that minimizes risk

$$\mathbf{L}(h) = \mathbf{E}_{\mathbb{X} \times \mathbb{Y}}\, l(\mathbb{Y}, h(\mathbb{X})) = \int \ell(h(\mathbf{x}), y) d\mathrm{p}.$$

- ERM: find best $h_n \in \mathcal{H}$ that minimizes empirical risk

$$\mathbf{L}_{emp}(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(h(\mathbf{x}_i), y_i).$$

Questions

- How can we know we are doing 'the right thing'?
- How to restrict the possible set of functions?

  Occam's Razor Of two equivalent models choose the simplest one. ?
- Can we quantify the 'complexity' of a learning problem?
- Is more data always better data?
- How much data do we need?

Statistical Learning Theory

- Provides a theoretical framework to study these questions.
- Started with ? which led to VC-Theory and SVM.
- Models the machine learning setting as a statistical phenomenon.
- Answers are probabilistic in nature.
- Tools: statistics, functional analysis, empirical processes, combinatorics, high-dimensional geometry, complexity theory.
- Newer view: ??.

Challenge Question

Is $\mathbf{L}(h_n)$ small, i.e. $\mathbf{L}(h_n) \approx \mathbf{L}(h^{**})$?

Magics?

Approximation & Estimation Error

$$\mathbf{L}(h_n) - \mathbf{L}(h^{**}) = \mathbf{L}(h_n) - \mathbf{L}(h^*) + \mathbf{L}(h^*) - \mathbf{L}(h^{**})$$

$\mathcal{H}$ large

- small approximation error
- overfitting

$\mathcal{H}$ small

- large approximation error
- better generalization but poor performance

Model selection

Choose $\mathcal{H}$ to get an optimal tradeoff between approximation and estimation error.

Estimation Error

$\mathbf{L}(h^*) - \mathbf{L}(h_n)$?

- depends on training data
- depends on $\mathcal{H}$
- depends on how algorithm chooses h_n
- depends on unknown p through h^* and risk

For ERM use uniform differences trick!

Uniform differences

$$|\mathbf{L}(h^*) - \mathbf{L}(h_n)| \leq 2 \sup_{h \in \mathcal{H}} |\mathbf{L}_{emp}(h) - \mathbf{L}(h)|$$

Empirical and Actual Risk

$\mathbf{L}_{emp}(h) \approx \mathbf{L}(h)$?

Asymptotics: Law of Large Numbers

For any fixed h, $|\mathbf{L}_{emp}(h) - \mathbf{L}(h)| \longrightarrow 0$ as $n \longrightarrow \infty$.

Finite Sample Result [Chernoff-Hoeffding]

For any fixed h, with high probability

$$|\mathbf{L}_{emp}(h) - \mathbf{L}(h)| \approx \frac{1}{\sqrt{n}}.$$

Does this mean that ERM finds optimal estimator h^* when training sample is getting large?

Empirical and Actual Risk

$\mathbf{L}_{emp}(h) \approx \mathbf{L}(h)$?

Asymptotics: Law of Large Numbers

For any fixed h, $|\mathbf{L}_{emp}(h) - \mathbf{L}(h)| \longrightarrow 0$ as $n \longrightarrow \infty$.

Finite Sample Result [Chernoff-Hoeffding]

For any fixed h, with high probability

$$|\mathbf{L}_{emp}(h) - \mathbf{L}(h)| \approx \frac{1}{\sqrt{n}}.$$

Does this mean that ERM finds optimal estimator h^* when training sample is getting large?

NO! h_n is a random variable and not fixed. A uniform LLN is needed, which holds simultaneously for all $h \in \mathcal{H}$. This is true only for classes $\mathcal{H}$ which are 'not too complex'.

Estimation Error

$\mathbf{L}(h_n) - \mathbf{L}(h^*)$?

Uniform differences

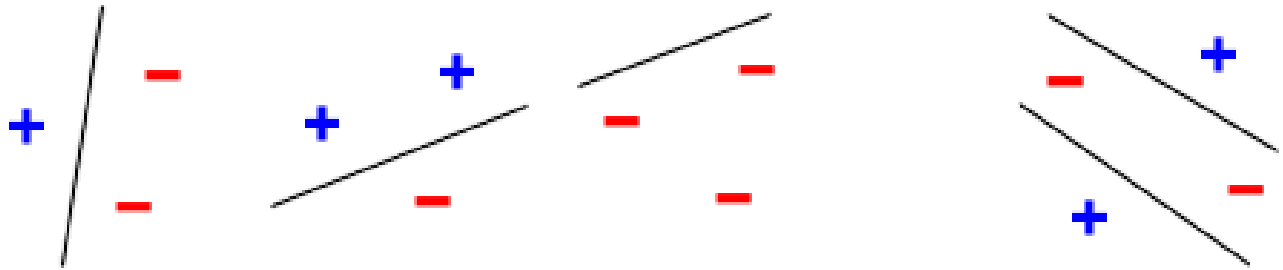
$$|\mathbf{L}(h_n) - \mathbf{L}(h^*)| \leq 2 \sup_{h \in \mathcal{H}} |\mathbf{L}_{emp}(h) - \mathbf{L}(h)|$$

Finite Sample Results

- One fixed function: $|\mathbf{L}_{emp}(h) - \mathbf{L}(h)| \approx 1/\sqrt{n}$
- $\mathcal{H}$ finite: $\sup_{h \in \mathcal{H}} |\mathbf{L}_{emp}(h) - \mathbf{L}(h)| \approx \sqrt{\log(|\mathcal{H}|)}/\sqrt{n}$
- $\mathcal{H}$ infinite: ?

VC Dimension

A model class shatters a set of data points if it can correctly classify any possible labeling.



Lines shatter any 3 points in $\mathbb{R}^2$, but not 4 points.

VC dimension [?]

The VC dimension of a model class is the maximum h such that some data point set of size h can be shattered by the model. (e.g. VC dimension of $\mathbb{R}^2$ is 3.)

A small VC dimension implies small complexity.

Why?

Shattering Coefficient: The number of distinct patterns a function class $\mathcal{H}$ can produce on a sample $\mathbf{x}$.

$$S_n(\mathcal{H}, \mathbf{x}) = |\mathcal{H}/\mathbf{x}|$$

Sauer-Shelah Lemma:

$$S_n(\mathcal{H}, \mathbf{x}) \leq \sum_{i=1}^{\mathsf{VC}(\mathcal{H})} \binom{n}{i} \sim n^{\mathsf{VC}(\mathcal{H})}$$

A function class 'behaves' on sample like a class with cardinality $n^{\mathsf{VC}(\mathcal{H})}$.

Optimal and Empirical Estimator

$\mathbf{L}(h^*) \approx \mathbf{L}(h_n)$?

Uniform differences

$$|\mathbf{L}(h^*) - \mathbf{L}(h_n)| \leq 2 \sup_{h \in \mathcal{H}} |\mathbf{L}_{emp}(h) - \mathbf{L}(h)|$$

Finite Sample Results

- One fixed function: $|\mathbf{L}_{emp}(h) - \mathbf{L}(h)| \approx 1/\sqrt{n}$
- $\mathcal{H}$ finite: $\sup_{h \in \mathcal{H}} |\mathbf{L}_{emp}(h) - \mathbf{L}(h)| \approx \sqrt{\log(|\mathcal{H}|)}/\sqrt{n}$
- $\mathcal{H}$ infinite: $\sup_{h \in \mathcal{H}} |\mathbf{L}_{emp}(h) - \mathbf{L}(h)| \approx \sqrt{\mathsf{VC}(\mathcal{H})\log(n)}/\sqrt{n}$

All results hold with high probability over the random draw of training samples!

Beyond VC Dimension

Rethinking: A function class 'behaves' in the worst case on sample like a class with cardinality VC$(\mathcal{H})$.

Rademacher Averages: Average of highest correlation of functions with n random patterns $\epsilon_i = +/-1$.

$$R_n(\mathcal{H}) = \mathbf{E}_{\mathbb{X},\epsilon}\left(\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^{n}\epsilon_i h(X_i)\right|\right)$$

- Extends theory to general loss functions.
- Finite VC dimension leads to upper estimate.
- More sophisticated mathematical machinery which avoids union bound.
- Better understanding of learning phenomenon.

??

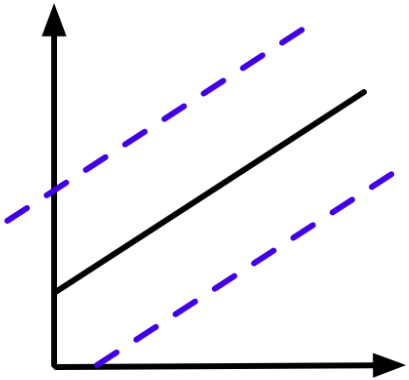
Implications

- VC dimensions are meaningful complexity measures.
- Do model selection by minimizing VC dimension.
- More data gives more likely a good predictor.

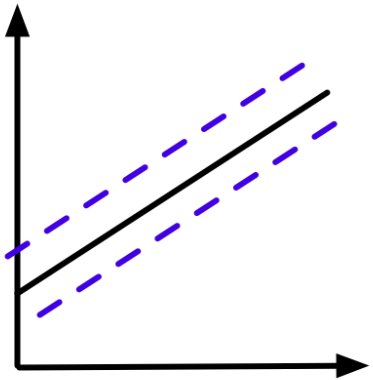
Larger Margin Classifiers

Large Margin $\Rightarrow$ Small VC dimension

Hyperplane classifiers with large margin have small VC dimension [?].



VC dim. small



VC dim. large

Maximum Margin $\Rightarrow$ Minimum Complexity

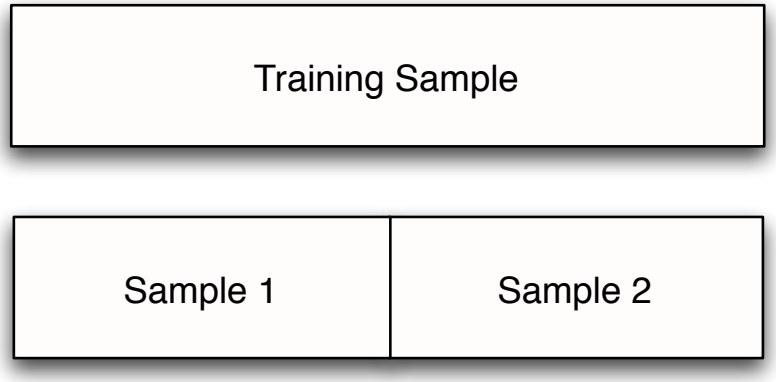
Minimize complexity by maximizing margin (irrespective of the dimension of the space).

Model Selection In Practice

VC-type complexities are hard to compute, but

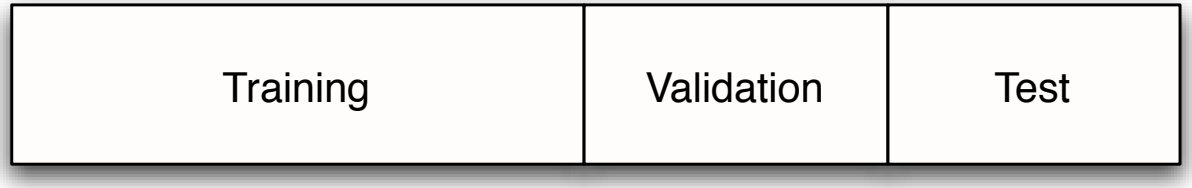
$$\mathbf{L}_{emp}(h) \sim \mathbf{L}(h) \Rightarrow \mathbf{L}_{emp2}(h) \sim \mathbf{L}(h)$$

Strategy: Choose among good empirical hypotheses the ones which are similar on independent samples.



Validation

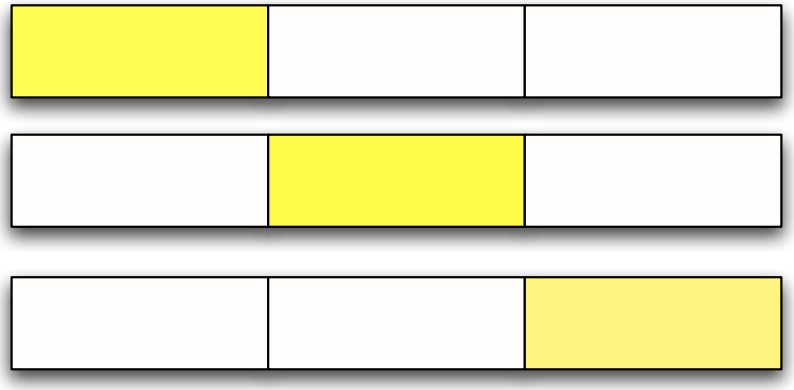
- Model selection and error estimation.
- Randomly chosen subsets of disjoint training, validation, test data.
- Works well for large data sets.



Crossvalidation

K-fold Crossvalidation:

- Splitting into K sample sets.
- K times training on K-1-sets.
- Error estimation through averaging on each of the K 'left-out' test sets.
- K trades bias vs. variance (in practice K=5,10).



Summary - SLT

- Provides a statistical framework to study learning algorithms.
- Quantifies the generalization ability in terms of
  - complexity of estimator functions
  - number of training examples.
- Results are probabilistic in nature (confidences).
- Results teach us
  - When and why our intuitive solutions were right (SVM, boosting, some forms of crossvalidation).
  - Why and how to restrict class of estimators and to regularize.
  - That more data is best because it increases confidence in result.
- But: Limited model, many questions not yet understood!