Machine Learning

MIRI Master

Lluís A. Belanche

belanche@cs.upc.edu

Soft Computing Research Group
Departament de Ciències de la Computació (Computer Science Department)
Universitat Politècnica de Catalunya - Barcelona Tech

Spring Semester 2017-2018

LECTURE 2: Theoretical issues (I): regression

Theoretical issues for regression

Outline

1. The regression framework

2. Bias-Variance analysis

3. Measuring complexity: the VC dimension

4. Empirical and Structural risk minimization

Theoretical issues for regression

The regression framework

Given data $\mathcal{D} = \{(\boldsymbol{x}_n, t_n)\}_{n=1,\dots,N}$, where $\boldsymbol{x}_n \in \mathbb{R}^d, t_n \in \mathbb{R}$,

Statistics: estimation of a continuous random variable (r.v.) T conditioned on a random vector $\boldsymbol{X}$

Mathematics: estimation of a real function f based on a finite number of "noisy" examples $(\boldsymbol{x}_n, f(\boldsymbol{x}_n))$

The departing **statistical setting** is $t_n = f(\boldsymbol{x}_n) + \varepsilon_n$; a **model** is any approximation of f

---

ε_n are i.i.d. continuous r.v. such that $\mathbb{E}[\varepsilon_n] = 0$ and $\mathrm{Var}[\varepsilon_n] = \sigma^2 < \infty$

Theoretical issues for regression

The regression framework

The **risk** of a model y is

$$R(y) := \int_{\mathbb{R}} \int_{\mathbb{R}^d} L\Big(t, y(\boldsymbol{x})\Big)\, p(t, \boldsymbol{x})\, d\boldsymbol{x}\, dt$$

where L is a suitable **loss** function:

- $L\Big(t, y(\boldsymbol{x})\Big) \geq 0$
- $L\Big(t, y(\boldsymbol{x})\Big) = 0$ if $t = y(\boldsymbol{x})$
- $L\Big(t, y(\boldsymbol{x})\Big)$ does not increase when $|t - y(\boldsymbol{x})|$ decreases

related to the distribution of the ε_n (the “noise model”)

Theoretical issues for regression

The regression framework

Since $\mathbb{E}[\varepsilon_n] = 0$, we can alternatively express the regression setting by stating that t is a continuous r.v. such that $f(\boldsymbol{x}) = \mathbb{E}[t|\boldsymbol{X} = \boldsymbol{x}]$:

$$\Longrightarrow f(\boldsymbol{x}) = \int_{\mathbb{R}} t\, p(t|\boldsymbol{x})\, dt$$

known as the **regression function**

Proof. (on the blackboard)

Theoretical issues for regression

The regression framework

Let us step firm ground and assume that $\varepsilon_n \sim N(0, \sigma^2)$ (implications?)

Using a **Maximum Likelihood** argument, it can be shown that the "right" loss is the **square error**:

$$L_{\mathsf{SE}}\Big(t, y(\boldsymbol{x})\Big) := \Big(t - y(\boldsymbol{x})\Big)^2$$

The **risk** is therefore

$$R(y) = \int_{\mathbb{R}} \int_{\mathbb{R}^d} \Big(t - y(\boldsymbol{x})\Big)^2 p(t|\boldsymbol{x})\, p(\boldsymbol{x})\, d\boldsymbol{x}\, dt$$

Theoretical issues for regression

The regression framework

If we enjoy complete freedom to choose y we should solve for:

$$y^* := \arg\min_y R(y)$$

The solution of which is:

$$y^*(\boldsymbol{x}) = \int_{\mathbb{R}} t\, p(t|\boldsymbol{x})\, dt = f(\boldsymbol{x})$$

(note it agrees with our previous result)

Theoretical issues for regression

The regression framework

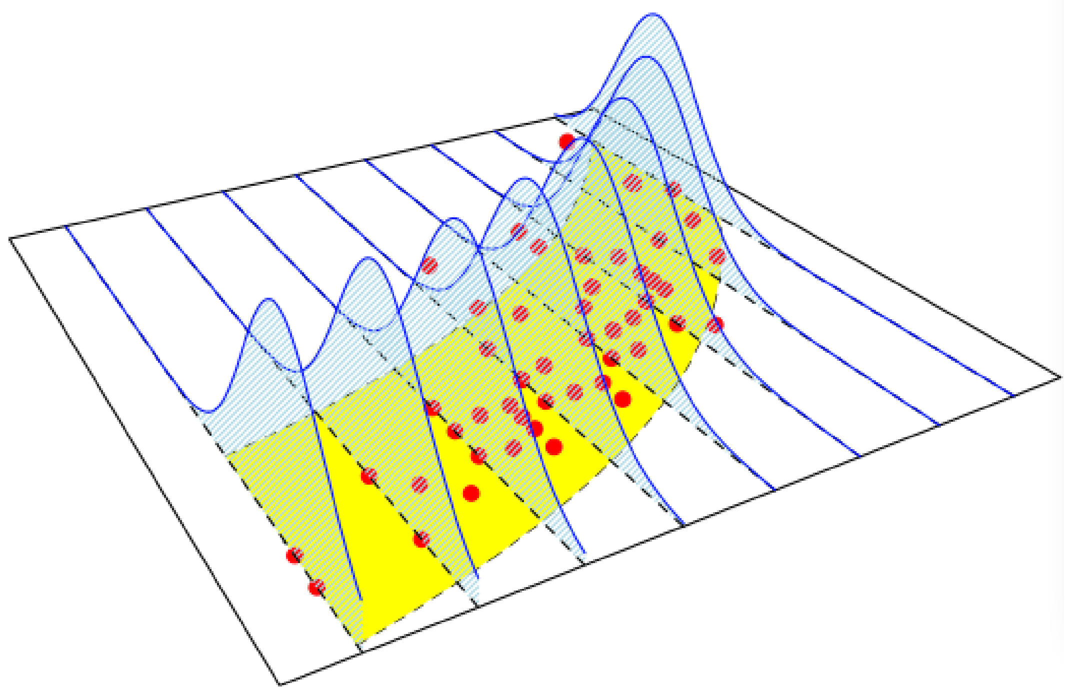


Illustration of the standard assumptions
(normality, homoscedasticity)

Theoretical issues for regression

The regression framework

In a practical setting, we do not know $p(t|\boldsymbol{x})$ ...

- Instead, we have a finite i.i.d. **data sample** of N labelled observations $\mathcal{D} = \{(\boldsymbol{x}_n, t_n)\}_{n=1,\ldots,N}$, where $\boldsymbol{x}_n \in \mathbb{R}^d, t_n \in \mathbb{R}$

- It seems natural to solve for y in (see below):

$$\int_{\mathbb{R}^d} \Big(f(\boldsymbol{x}) - y(\boldsymbol{x})\Big)^2 p(\boldsymbol{x})\, d\boldsymbol{x}$$

- We must impose restrictions on the possible solutions y (a specific **class of functions**)

Theoretical issues for regression

The regression framework

We can compute an approximation to the true risk, called the **empirical risk**, by averaging the loss function on the available data $\mathcal{D}$:

$$R_{\mathrm{emp}}(y) := \frac{1}{N} \sum_{n=1}^{N} (t_n - y(\boldsymbol{x_n}))^2$$

(this quantity is also known as the **training**, resubstitution or apparent **error**)

The **Empirical Risk Minimization** (ERM) principle states that a learning algorithm should choose a hypothesis (model) $\widehat{y}$ which minimizes the empirical risk among a predefined class of functions $\mathcal{Y}$:

$$\widehat{y} := \arg\min_{y \in \mathcal{Y}} R_{\mathrm{emp}}(y)$$

Theoretical issues for regression

The regression framework

The quantity $R_{\mathsf{emp}}(\widehat{y})$ is known as the **training error**

In theoretical ML, we are very much interested in:

1. how this error fluctuates as a function of $\mathcal{D}$

2. how far this error is from the true error, *i.e.*, to bound $|R_{\mathsf{emp}}(\widehat{y}) - R(y)|$; at the very least, to bound $|\mathbb{E}[R_{\mathsf{emp}}(\widehat{y})] - R(y)|$

3. how far this error is from the best possible error, *i.e.*, to bound $|R_{\mathsf{emp}}(\widehat{y}) - R(y^*)|$; at the very least, to bound $|\mathbb{E}[R_{\mathsf{emp}}(\widehat{y})] - R(y^*)|$

Theoretical issues for regression

Bias-Variance analysis

Recall the assumption that $\varepsilon_n \sim N(0, \sigma^2)$

In this case (using the square error), the risk can be decomposed as:

$$
\begin{aligned}
R(y) &= \int_{\mathbb{R}} \int_{\mathbb{R}^d} \Big(t - y(\boldsymbol{x})\Big)^2 p(t, \boldsymbol{x})\, d\boldsymbol{x}\, dt \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}^d} \Big(t - f(\boldsymbol{x})\Big)^2 p(t, \boldsymbol{x})\, d\boldsymbol{x}\, dt \\
&\quad + \int_{\mathbb{R}^d} \Big(f(\boldsymbol{x}) - y(\boldsymbol{x})\Big)^2 p(\boldsymbol{x})\, d\boldsymbol{x} \\
&= \sigma^2 + \int_{\mathbb{R}^d} \Big(f(\boldsymbol{x}) - y(\boldsymbol{x})\Big)^2 p(\boldsymbol{x})\, d\boldsymbol{x} =: \sigma^2 + \mathrm{MSE}(y)
\end{aligned}
$$

where f is the **regression function**. Hint: add and subtract $f(\boldsymbol{x})$

Theoretical issues for regression

Bias-Variance analysis

Therefore we arrive at $R(y) = \sigma^2 + \mathsf{MSE}(y)$

We can now "forget" about σ^2 and the risk and minimize instead the MSE "to the last bullet":

$$\mathsf{MSE}(y) = \int_{\mathbb{R}^d} \Big(f(\boldsymbol{x}) - y(\boldsymbol{x})\Big)^2 p(\boldsymbol{x})\, d\boldsymbol{x}$$

A **learning algorithm** is a procedure that, given $\mathcal{D}$ and $\mathcal{Y}$, outputs a model $y_{\mathcal{D}} \in \mathcal{Y}$

Theoretical issues for regression

Bias-Variance analysis

- Consider now one particular $\boldsymbol{x}_\circ$: different $\mathcal{D}$ will produce different $y_\mathcal{D}$ and therefore different predictions $y_\mathcal{D}(\boldsymbol{x}_\circ)$ ...

- Let us concentrate on the quantity $\Big(f(\boldsymbol{x}_\circ) - y_\mathcal{D}(\boldsymbol{x}_\circ)\Big)^2$

- We wish to eliminate the dependence on $\mathcal{D}$; therefore we investigate its expected value:

$$\mathbb{E}_\mathcal{D}\Big[\Big(f(\boldsymbol{x}_\circ) - y_\mathcal{D}(\boldsymbol{x}_\circ)\Big)^2\Big], \qquad \text{taken over all possible } \mathcal{D} \text{ of size } N$$

Theoretical issues for regression

Bias-Variance analysis

$$\mathbb{E}_{\mathcal{D}}\Big[\Big(f(\boldsymbol{x}_{\circ}) - y_{\mathcal{D}}(\boldsymbol{x}_{\circ})\Big)^2\Big] \;=\; \Big(f(\boldsymbol{x}_{\circ}) - \mathbb{E}_{\mathcal{D}}\big[y_{\mathcal{D}}(\boldsymbol{x}_{\circ})\big]\Big)^2 + \mathbb{E}_{\mathcal{D}}\Big[\Big(y_{\mathcal{D}}(\boldsymbol{x}_{\circ}) - \mathbb{E}_{\mathcal{D}}\big[y_{\mathcal{D}}(\boldsymbol{x}_{\circ})\big]\Big)^2\Big]$$

$$\Rightarrow \mathsf{MSE}(y_{\mathcal{D}}(\boldsymbol{x}_{\circ})) = \Big(Bias(y_{\mathcal{D}}(\boldsymbol{x}_{\circ}))\Big)^2 + \mathsf{Var}(y_{\mathcal{D}}(\boldsymbol{x}_{\circ}))$$

$$R(y_{\mathcal{D}}(\boldsymbol{x}_{\circ})) = \sigma^2 + \Big(Bias(y_{\mathcal{D}}(\boldsymbol{x}_{\circ}))\Big)^2 + \mathsf{Var}(y_{\mathcal{D}}(\boldsymbol{x}_{\circ}))$$

Theoretical issues for regression

Bias-Variance analysis

The prediction risk at any given point x_0 is the sum of three components:

The noise variance: variability of the target value around its conditional mean

The (squared) bias: average (square) deviation of our prediction at x_0 and the best possible prediction

The variance: variability of our prediction as a function of the used data sample (regardless of the underlying function!)

Theoretical issues for regression

Bias-Variance analysis

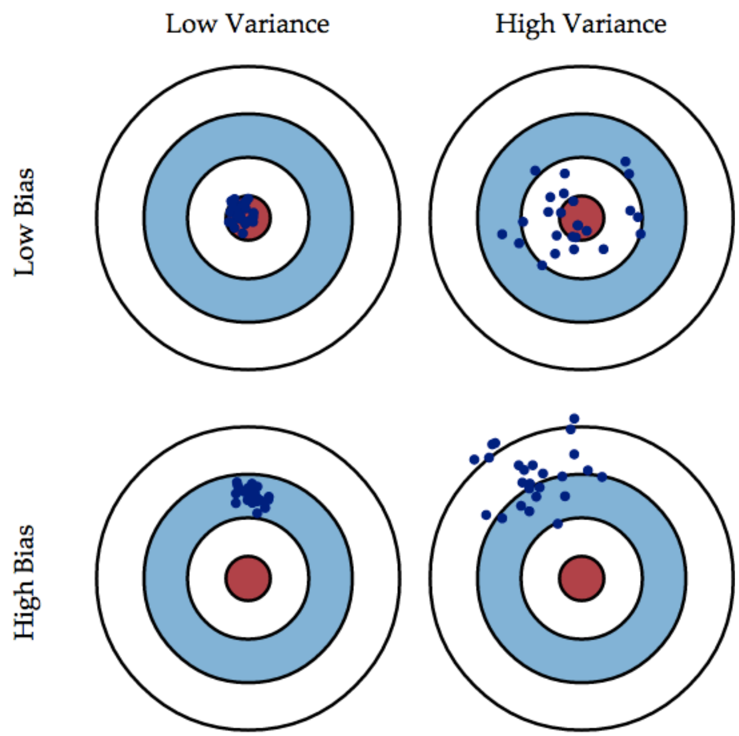


Illustration of the **Bias-Variance decomposition** using a dartboard

Theoretical issues for regression

Bias-Variance analysis

The derivation above depends on a particular point $\boldsymbol{x}_{\text{o}}$ ... let us put it back in place (*i.e.*, within their integrals):

$$\Big(Bias(y_{\mathcal{D}})\Big)^2 = \int_{\mathbb{R}^d} \Big(Bias(y_{\mathcal{D}}(\boldsymbol{x}))\Big)^2 p(\boldsymbol{x})\, d\boldsymbol{x}$$

$$\mathsf{Var}(y_{\mathcal{D}}) = \int_{\mathbb{R}^d} Var(y_{\mathcal{D}}(\boldsymbol{x}))\, p(\boldsymbol{x})\, d\boldsymbol{x}$$

$$R(y_{\mathcal{D}}) = \sigma^2 + \Big(Bias(y_{\mathcal{D}})\Big)^2 + \mathsf{Var}(y_{\mathcal{D}})$$

Theoretical issues for regression

Bias-Variance analysis

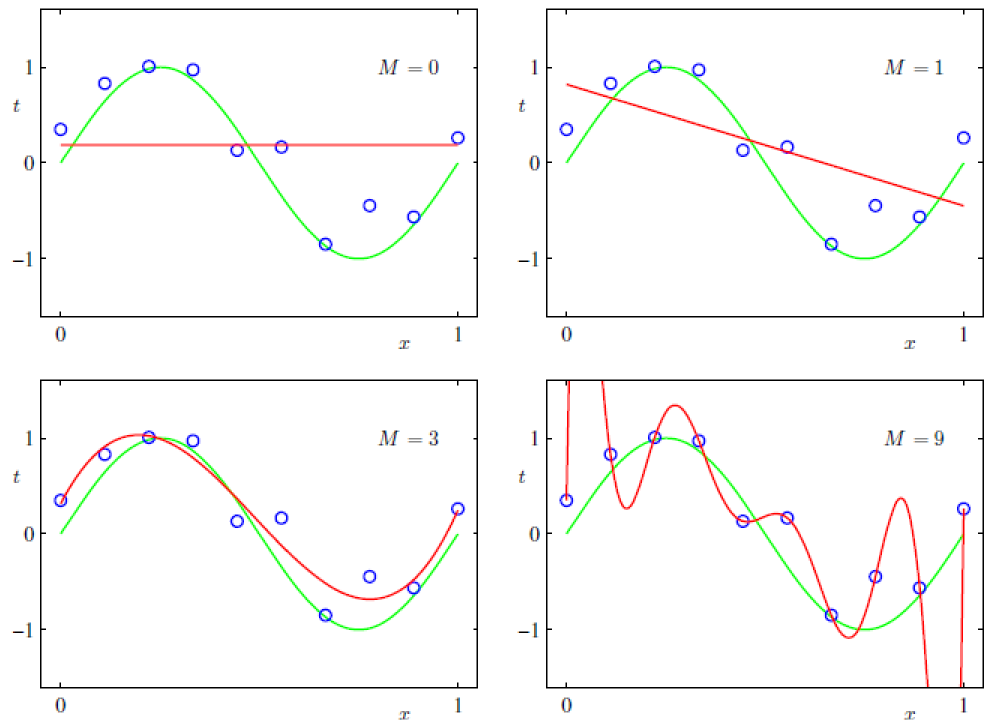


Illustration of the **Bias-Variance tradeoff** (a.k.a. **dilemma**)

Theoretical issues for regression

Bias-Variance analysis

In general,

- an **underfit** model will have a high bias
- an **overfit** model will have a high variance

The "ability to fit" has a name: **complexity** of the function class

- Models that are "more complex than needed" will tend to have a large prediction error, which will be dominated by the **variance** term
- Models that are "less complex than needed" will tend to have a large prediction error, which will be dominated by the (square) **bias** term

Theoretical issues for regression

Bias-Variance analysis

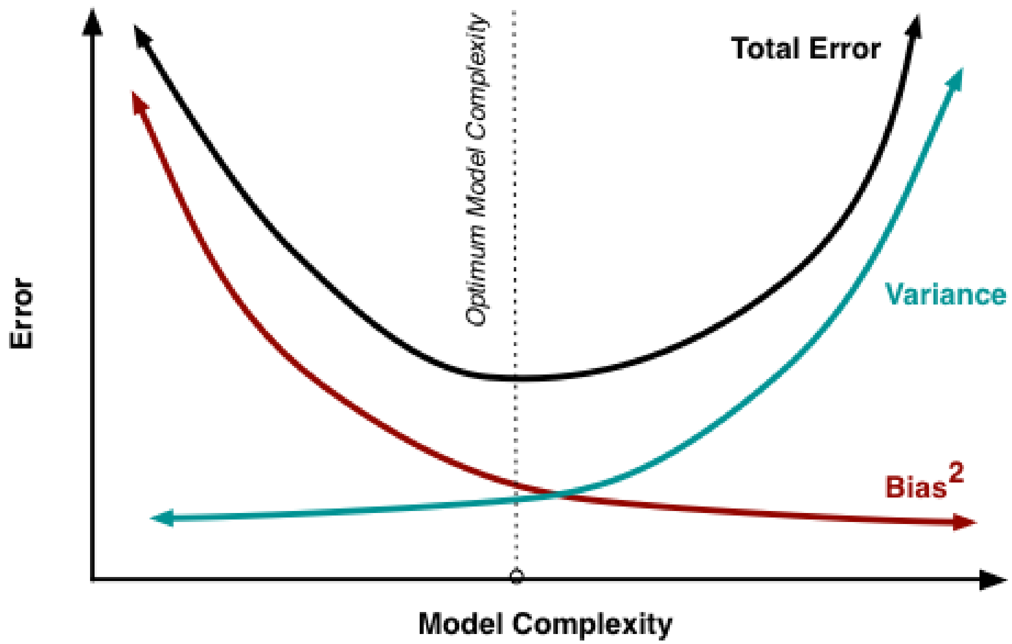


Illustration of **Bias^2**, **Var**, **MSE (Total Error)** and **Model Complexity**

Theoretical issues for regression

Measuring complexity: the VC dimension

How do we measure "**complexity** of the function class"?

Let

$$\mathcal{Y} = \Big\{ y(\boldsymbol{x}; \boldsymbol{\alpha}),\ \ \boldsymbol{\alpha} \in A \Big\}$$

be a class of parametric binary classifiers $y : \mathbb{R}^d \to \{-1, +1\}$

1. How "complex" is $\mathcal{Y}$?

2. How is complexity related to the number of parameters?

Theoretical issues for regression

Measuring complexity: the VC dimension

Example 1

Let $\mathcal{Y}_d = \left\{ y(\boldsymbol{x};\boldsymbol{\alpha}),\ \ \boldsymbol{\alpha} \in \mathbb{R}^{d+1} \right\}$

where $y : \mathbb{R}^d \to \{-1, +1\}$ is a class of **linear** classifiers in $\mathbb{R}^d$:

$$y(\boldsymbol{x};\boldsymbol{\alpha}) = \mathrm{sgn}\left(\alpha_0 + \sum_{i=1}^{d} \alpha_i x_i \right)$$

Theoretical issues for regression

Measuring complexity: the VC dimension

1. Take a number N of data vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ in $\mathbb{R}^d$

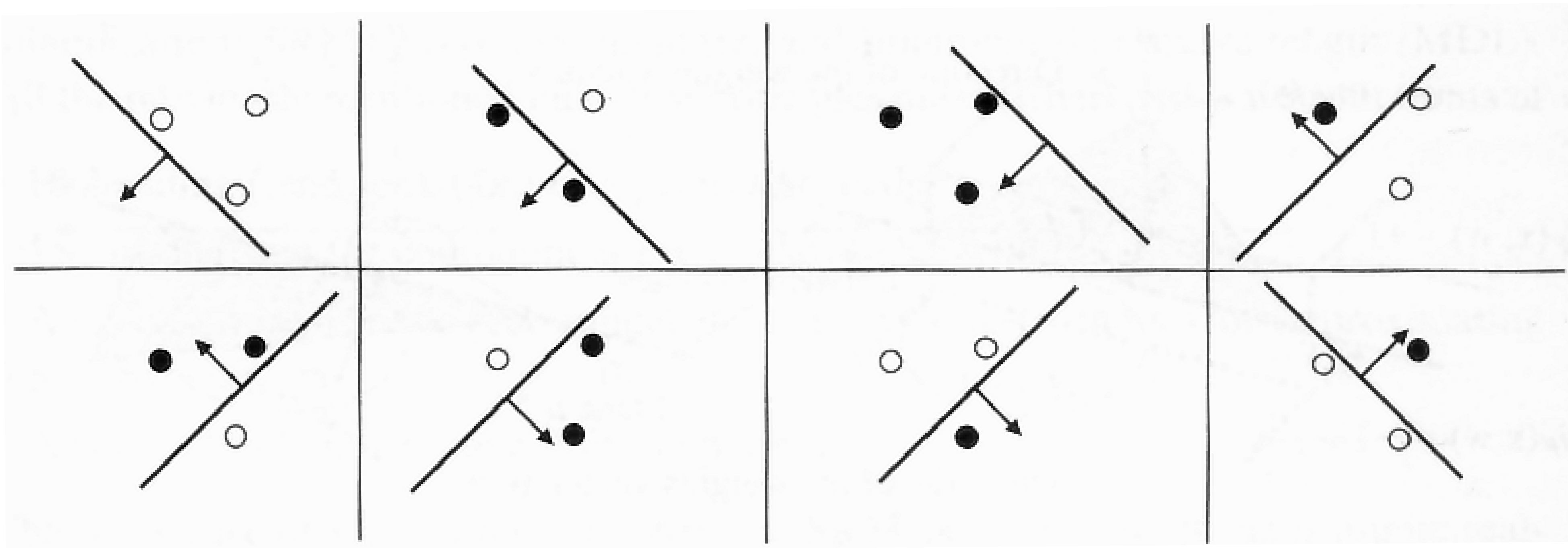
2. Consider all 2^N possible $\{-1, +1\}$-labellings of these vectors

3. We say that a function class $\mathcal{Y}$ **shatters** the vectors if, for all possible labellings, there exists a function in $\mathcal{Y}$ (a classifier) that perfectly separates the vectors

4. The **VC dimension** of a function class $\mathcal{Y}$ is the **maximum** $N \in \mathbb{N}$ for which N data vectors can be found that can be shattered by $\mathcal{Y}$

Theoretical issues for regression

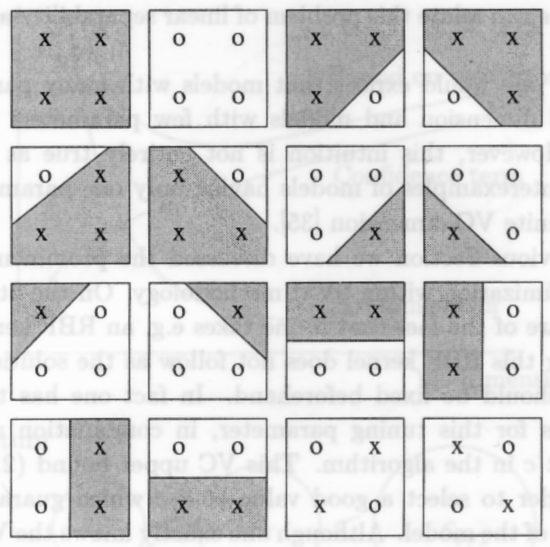
Measuring complexity: the VC dimension



VC-dim($\mathcal{Y}_2$) ≥ 3

Theoretical issues for regression

Measuring complexity: the VC dimension



- VC-dim($\mathcal{Y}_2$) < 4 and therefore VC-dim($\mathcal{Y}_2$) $= 3$
- It can be shown that VC-dim($\mathcal{Y}_d$) $= d + 1$ (i.e., the number of parameters)

Theoretical issues for regression

Measuring complexity: the VC dimension

In order to prove that VC-dim($\mathcal{Y}$) $= N$ for some N we have to:

1. find a set of N data vectors that can be shattered by $\mathcal{Y}$

2. prove that no set of $N + 1$ data vectors can be shattered by $\mathcal{Y}$

If, for all $N \in \mathbb{N}$, we can *always* find a set of N data vectors that can be shattered by $\mathcal{Y}$, we say that VC-dim($\mathcal{Y}$) $= \infty$

Theoretical issues for regression

Measuring complexity: the VC dimension

Example 2

Let $\mathcal{Y} = \Big\{ y(x;\alpha),\ \ \alpha \in \mathbb{R} \Big\}$

where $y : \mathbb{R}^d \to \{-1, +1\}$ is the class of **sine** classifiers in $\mathbb{R}$:

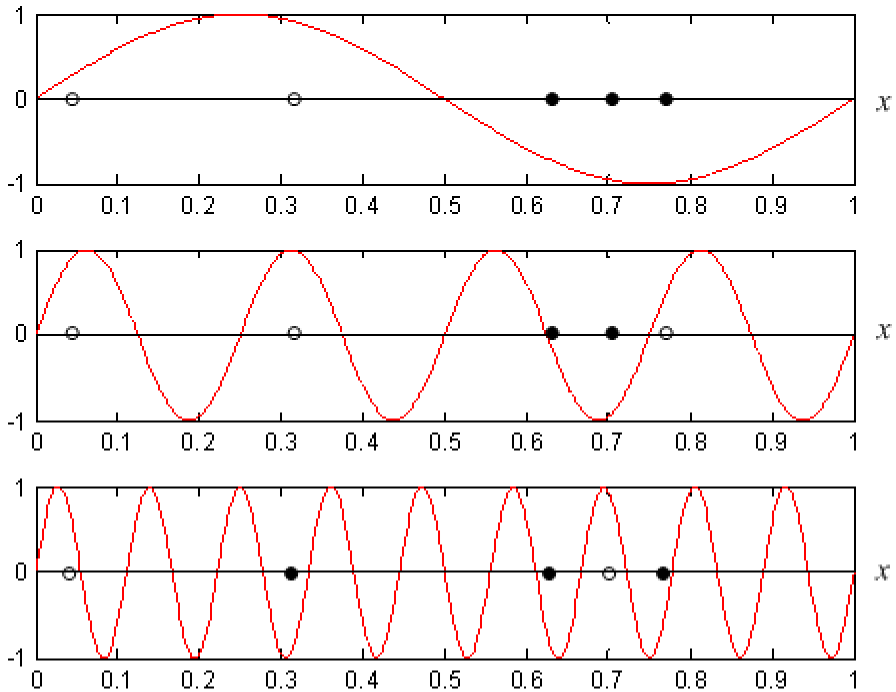
$$y(x;\alpha) = \mathrm{sgn}\Big(\sin(\alpha x) \Big)$$

It can be shown that VC-dim$(\mathcal{Y}) = \infty$, with the choice $x_n = 10^{-n}$ and

$$\alpha = \pi \left(1 + \frac{1}{2} \sum_{n=1}^{N} (1 - t_n) 10^n \right)$$

Theoretical issues for regression

Measuring complexity: the VC dimension



Plot of the function $\sin(\alpha x)$, for different α and arbitrary $\{-1, +1\}$-labellings of $N = 5$ points (in black and white)

Theoretical issues for regression

Using the VC dimension for two-class classification

Theorem (Vapnik and Chervonenkis, 1974). Let $\mathcal{D}$ be an i.i.d. data sample of size N and $\mathcal{Y}$ a class of parametric binary classifiers. Let ϑ denote the VC dimension of $\mathcal{Y}$. Take $y \in \mathcal{Y}$ with empirical error $R_{\text{emp}}(y)$ on $\mathcal{D}$. For all $\eta \in (0,1)$ it holds true that, with probability at least $1-\eta$, the true error of y is bounded by:

$$R(y) \leq R_{\text{emp}}(y) + H(N, \vartheta, \eta)$$

where

$$H(N, \vartheta, \eta) := \sqrt{\frac{\vartheta(\ln(2N/\vartheta) + 1) - \ln(\eta/4)}{N}}$$

Theoretical issues for regression

Structural risk minimization

Consider a nested sequence of function classes:

$\mathcal{Y}_1 \subset \mathcal{Y}_2 \subset \ldots \mathcal{Y}_k \subset \ldots$ with respective VC-dimensions $\vartheta_1 < \vartheta_2 \ldots < \vartheta_k \ldots$

- The **Structural Risk Minimization** (SRM) principle states that a learning algorithm should choose a hypothesis (model) which minimizes the previous bound on the true error
- The SRM principle can also be applied to the regression case, by extending the definition of VC-dimension
- Other definitions of complexity (to measure the "richness" of classes of real functions) have been proposed (Pseudo-Dimension, Fat-Shattering Dimension, Rademacher complexity)

Theoretical issues for regression

Structural risk minimization

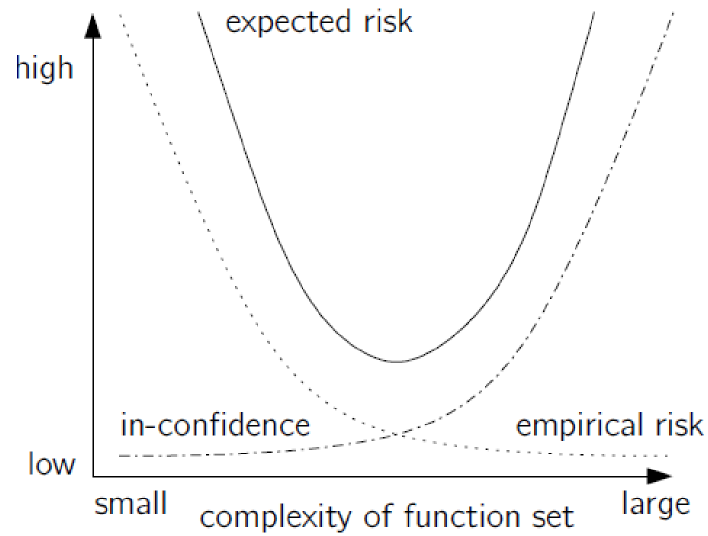


Figure 2.2: Schematic illustration of (2.8). The dotted line represents the training error (empirical risk), the dash-dotted line the upper bound on the complexity term (confidence). With higher complexity the empirical error decreases but the upper bound on the risk confidence becomes worse. For a certain complexity of the function class the best expected risk (solid line) is obtained. Thus, in practice the goal is to find the best trade-off between empirical error and complexity.

Illustration of the **Empirical error vs. complexity tradeoff**

Theoretical issues for regression

Structural risk minimization

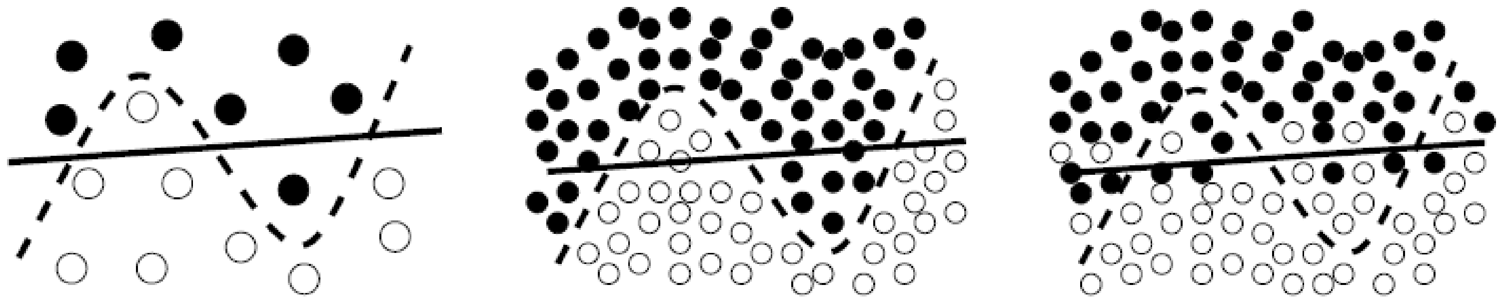


Figure 2.1: Illustration of the over–fitting dilemma: Given only a small sample (left) either, the solid or the dashed hypothesis might be true, the dashed one being more complex, but also having a smaller training error. Only with a large sample we are able to see which decision reflects the true distribution more closely. If the dashed hypothesis is correct the solid would under-fit (middle); if the solid were correct the dashed hypothesis would over-fit (right).

Interpretation of the **Overfitting vs. underfitting dilemma**

(last two figures from S. Mika's PhD dissertation, Technische Universität Berlin, 2002)

Machine Learning

Syllabus

1. Introduction to Machine Learning

2. Theoretical issues (I): regression

3. Linear regression and beyond

4. Theoretical issues (II): classification

5. Generative classifiers

6. Discriminative classifiers

7. Clustering: k-means and E-M
8. Learning with kernels (I): The SVM
9. Learning with kernels (II): Kernel functions
10. Artificial neural networks (I): Delta rule, MLP-1
11. Artificial neural networks (II): MLP-2, RBF
12. Artificial neural networks (III): DL and CNNs
13. Ensemble methods: Random Forests
14. Advanced topics and frontiers