# Aprentatge Automàtic 1

## GCED

Lluís A. Belanche

**belanche@cs.upc.edu**

Soft Computing Research Group
Dept. de Ciències de la Computació (Computer Science)

Universitat Politècnica de Catalunya

2018-2019

## LECTURE 3: Theory for regression and linear models (I)

# 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 $\mathcal{D} = \{(\boldsymbol{x}_n, t_n)\}_{n=1,\ldots,N}$, 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 $f$

---

$\varepsilon_n$ 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 $\varepsilon_n$ (the "noise model")

# 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 $y$, the solution is:

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

known as the **regression function**.

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

$$\Longrightarrow f = y^*$$

# Theoretical issues for regression

## The regression framework

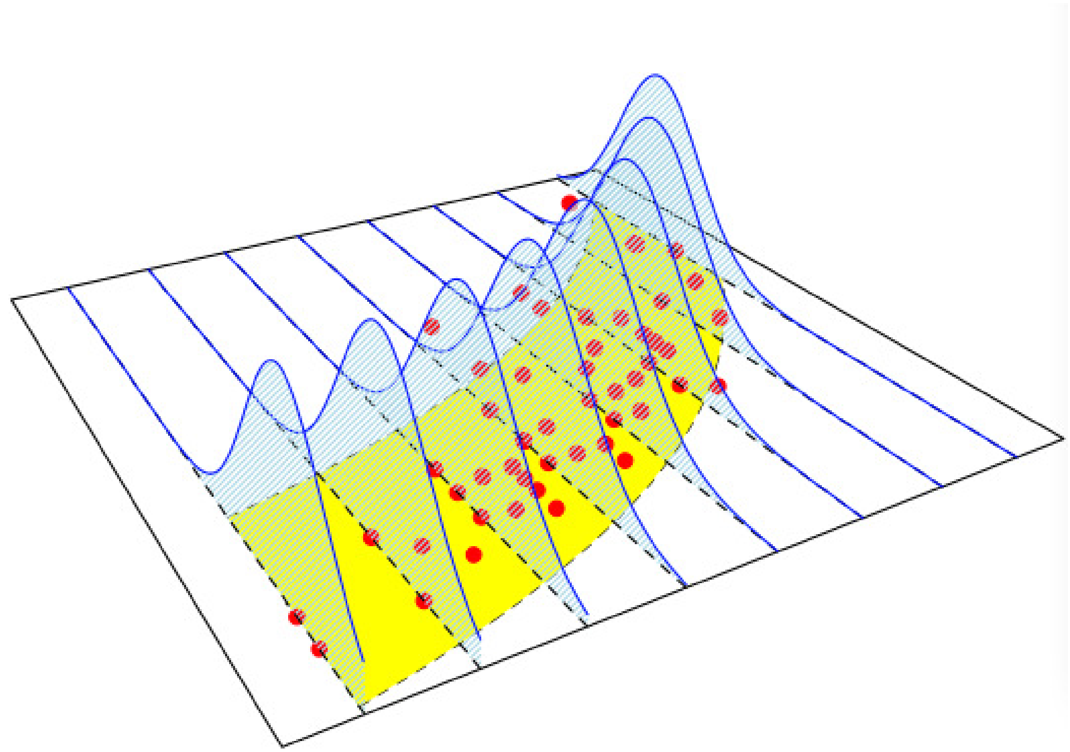

Illustration of the standard assumptions
(normality, homoscedasticity)

# 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 $\boldsymbol{x}_n \in \mathbb{R}^d, t_n \in \mathbb{R}$

- Intuition (?) is telling us to solve for $y$ in:

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

- 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 $R_{\text{emp}}(\widehat{y})$ 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 $|R_{\text{emp}}(\widehat{y}) - R(y)|$; at the very least, to bound $|\mathbb{E}[R_{\text{emp}}(\widehat{y})] - R(y)|$

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

# 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:

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

where $f$ is the **regression function**.

# Theoretical issues for regression

## Bias-Variance analysis

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

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

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

---

A **learning algorithm** for **regression** is a procedure that, given $\mathcal{D}$ and $\mathcal{Y}$, outputs a model $y_{\mathcal{D}} \in \mathcal{Y}$ that aims to minimize $\mathrm{MSE}(y)$.

# Theoretical issues for regression

## Bias-Variance analysis

- Consider now one particular $\boldsymbol{x}_0$: different $\mathcal{D}$ will produce different $y_{\mathcal{D}}$ and therefore different predictions $y_{\mathcal{D}}(\boldsymbol{x}_0)$ ...
- Let us concentrate on the quantity $\Big(f(\boldsymbol{x}_0) - y_{\mathcal{D}}(\boldsymbol{x}_0)\Big)^2$
- We wish to eliminate the dependence on $\mathcal{D}$; therefore we investigate its expected value:

$$\mathbb{E}_{\mathcal{D}}\Big[\Big(f(\boldsymbol{x}_0) - y_{\mathcal{D}}(\boldsymbol{x}_0)\Big)^2\Big], \qquad \text{taken over all possible } \mathcal{D} \text{ of size } N$$

# Theoretical issues for regression

## Bias-Variance analysis

$$\mathbb{E}_{\mathcal{D}}\Big[\Big(f(\boldsymbol{x}_0) - y_{\mathcal{D}}(\boldsymbol{x}_0)\Big)^2\Big] \;=\; \Big(f(\boldsymbol{x}_0) - \mathbb{E}_{\mathcal{D}}\big[y_{\mathcal{D}}(\boldsymbol{x}_0)\big]\Big)^2 + \mathbb{E}_{\mathcal{D}}\Big[\Big(y_{\mathcal{D}}(\boldsymbol{x}_0) - \mathbb{E}_{\mathcal{D}}\big[y_{\mathcal{D}}(\boldsymbol{x}_0)\big]\Big)^2\Big]$$

$$\Rightarrow \mathsf{MSE}(y_{\mathcal{D}}(\boldsymbol{x}_0)) = \Big(Bias(y_{\mathcal{D}}(\boldsymbol{x}_0))\Big)^2 + \mathsf{Var}(y_{\mathcal{D}}(\boldsymbol{x}_0))$$

---

$$R(y_{\mathcal{D}}(\boldsymbol{x}_0)) = \sigma^2 + \Big(Bias(y_{\mathcal{D}}(\boldsymbol{x}_0))\Big)^2 + \mathsf{Var}(y_{\mathcal{D}}(\boldsymbol{x}_0))$$

# 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

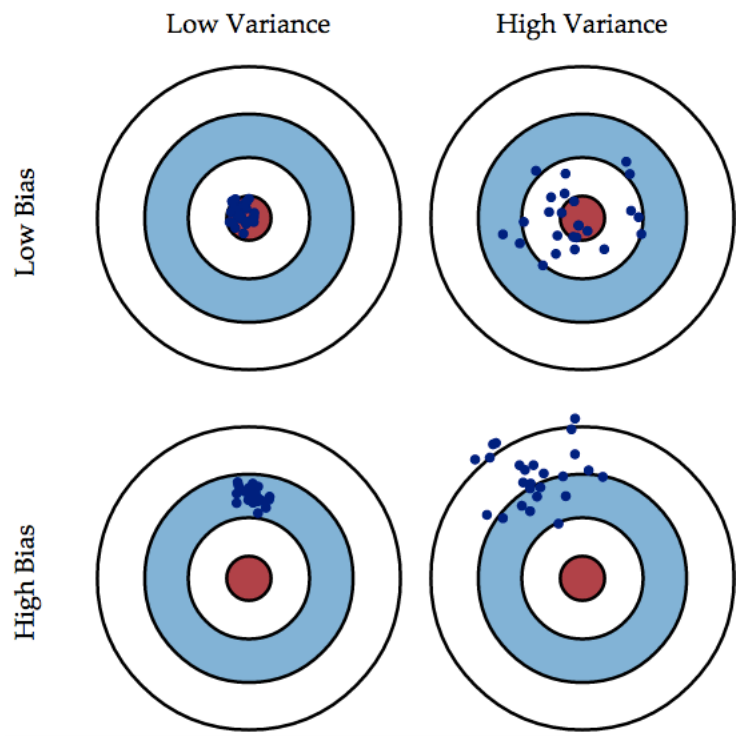


Illustration of the **Bias-Variance decomposition** using a dartboard

# Theoretical issues for regression

## Bias-Variance analysis

The derivation above depends on a particular point $\boldsymbol{x}_0$ ... 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

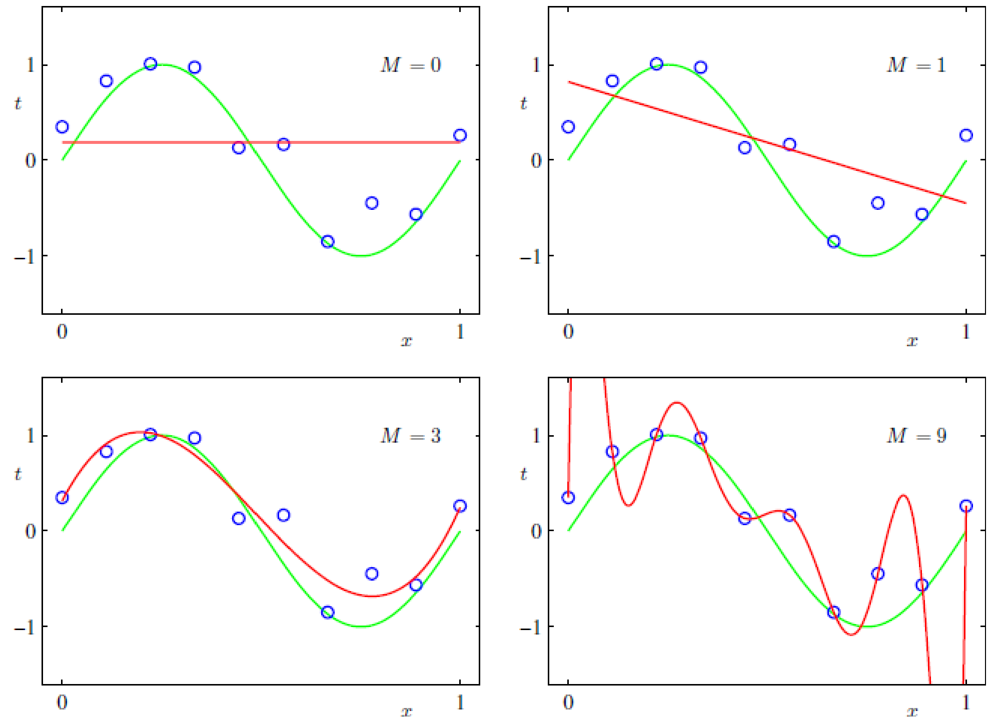


Illustration of the **Bias-Variance tradeoff** (a.k.a. **dilemma**)

# 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

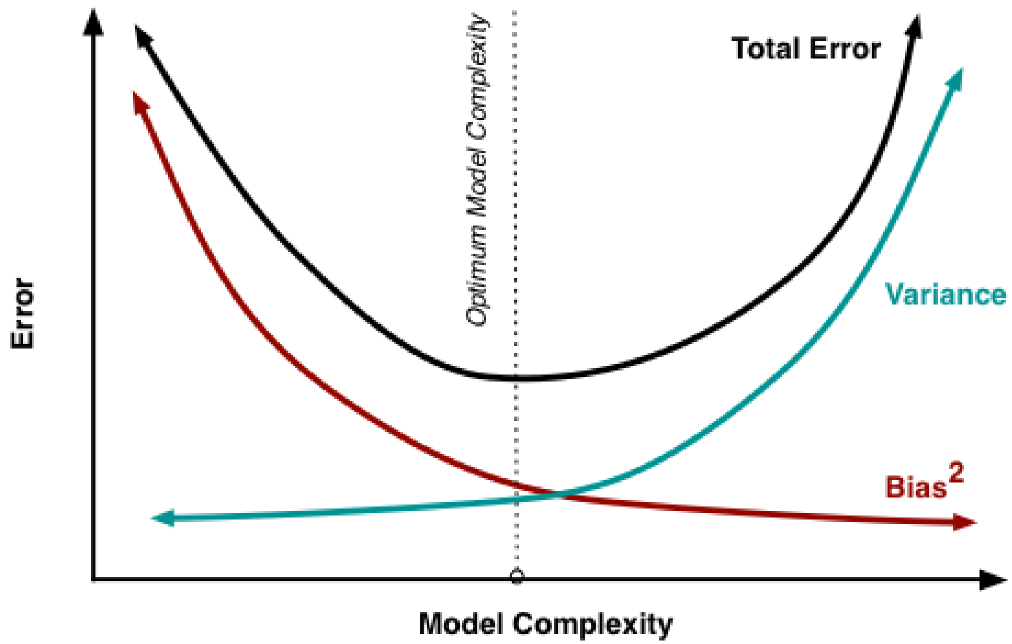


Illustration of **Bias**$^2$, **Var**, **MSE (Total Error)** and **Model Complexity**

# Theoretical issues for regression

## Bias-Variance analysis

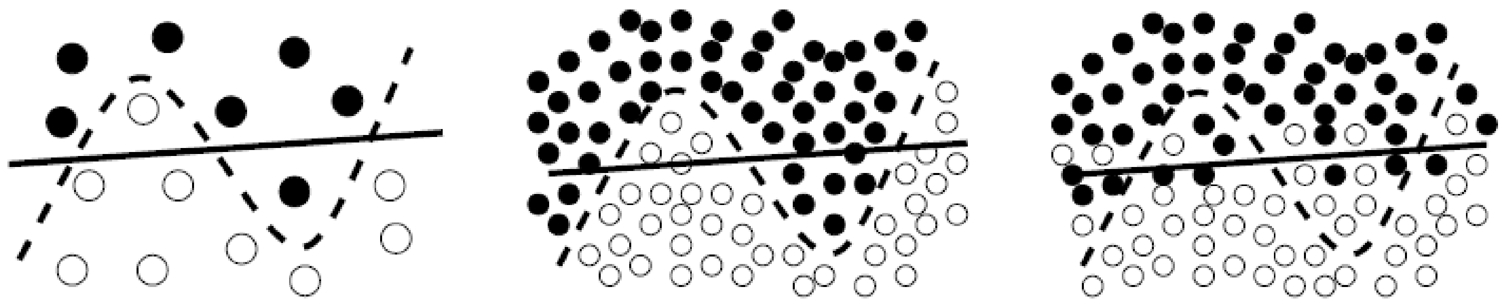

**Figure 2.1:** Illustration of the over–fitting dilemma: Given only a small sample (left) either, the solid or the dashed hypothesis might be true, the dashed one being more complex, but also having a smaller training error. Only with a large sample we are able to see which decision reflects the true 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)