Lecture 2

Overfitting, testing, validating and measuring

Machine Learning

Ivan Smetannikov

02.07.2019

Lecture plan

- Overfitting
- Model evaluation
- Validation
- Classification performance measures
- Regression performance measures

Lecture plan

- Overfitting
- Model evaluation
- Validation
- Classification performance measures
- Regression performance measures

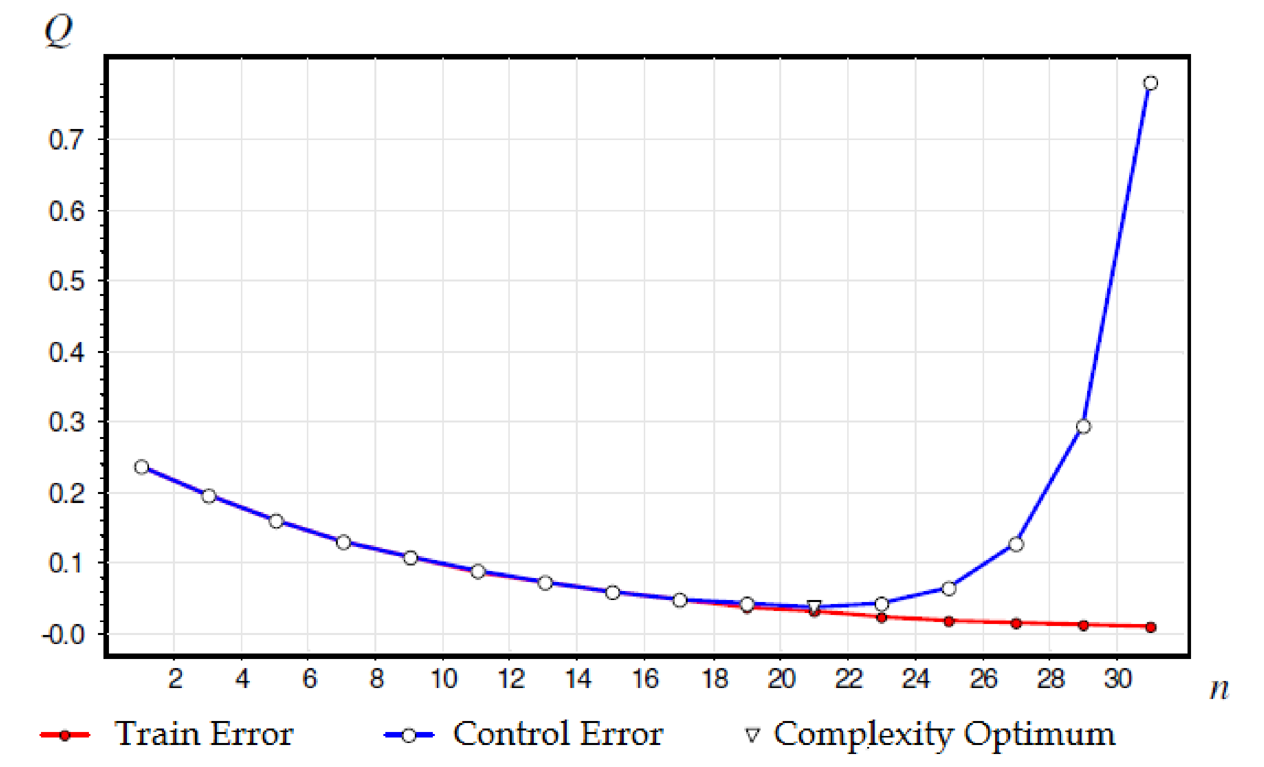
Overfitting problem

Overfitting problem: from a certain model complexity lever, the better an algorithm performs on train set X^{ℓ}, the worse it performs on real world objects.

Example of overfitting

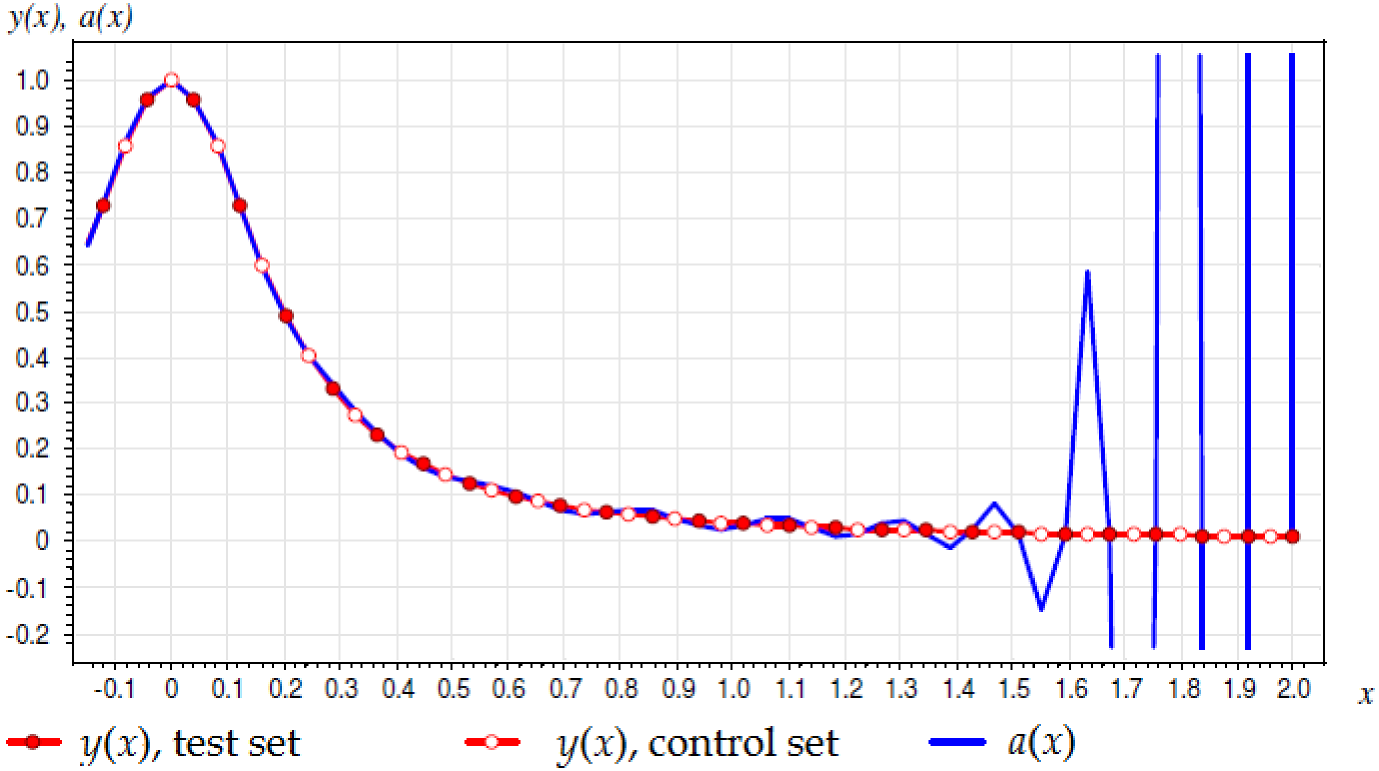
Dependency $y(x) = \frac{1}{1+25x^2}$ defined on $x \in [-2, 2]$.

Let search a function among polynomials with degree n.

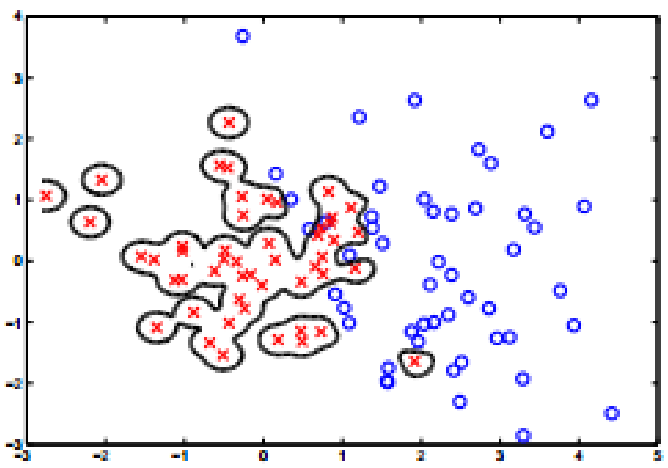


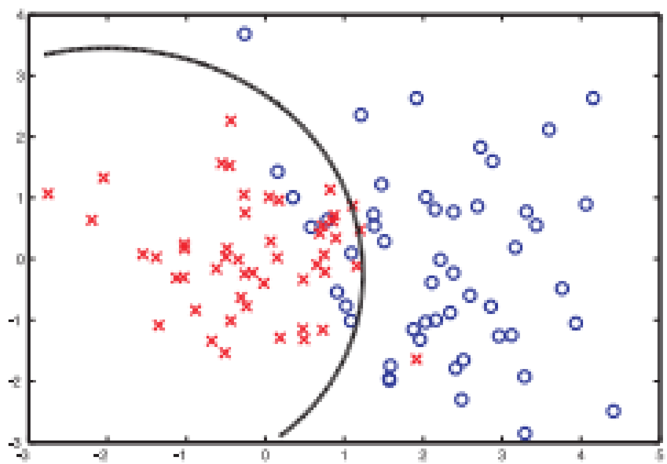
Overfitted algorithm

$$y(x) = \frac{1}{1 + 25x^2}; \quad a(x) \text{ — polynomial of degree } n = 38$$



Tuning model complexity





Lecture plan

- Overfitting
- Model evaluation
- Validation
- Classification performance measures
- Regression performance measures

Hold-out validation

Hold-out validation, HO

Split training sample into two parts:

$$T^{\ell} = T^{t} \cup T^{\ell-t}$$

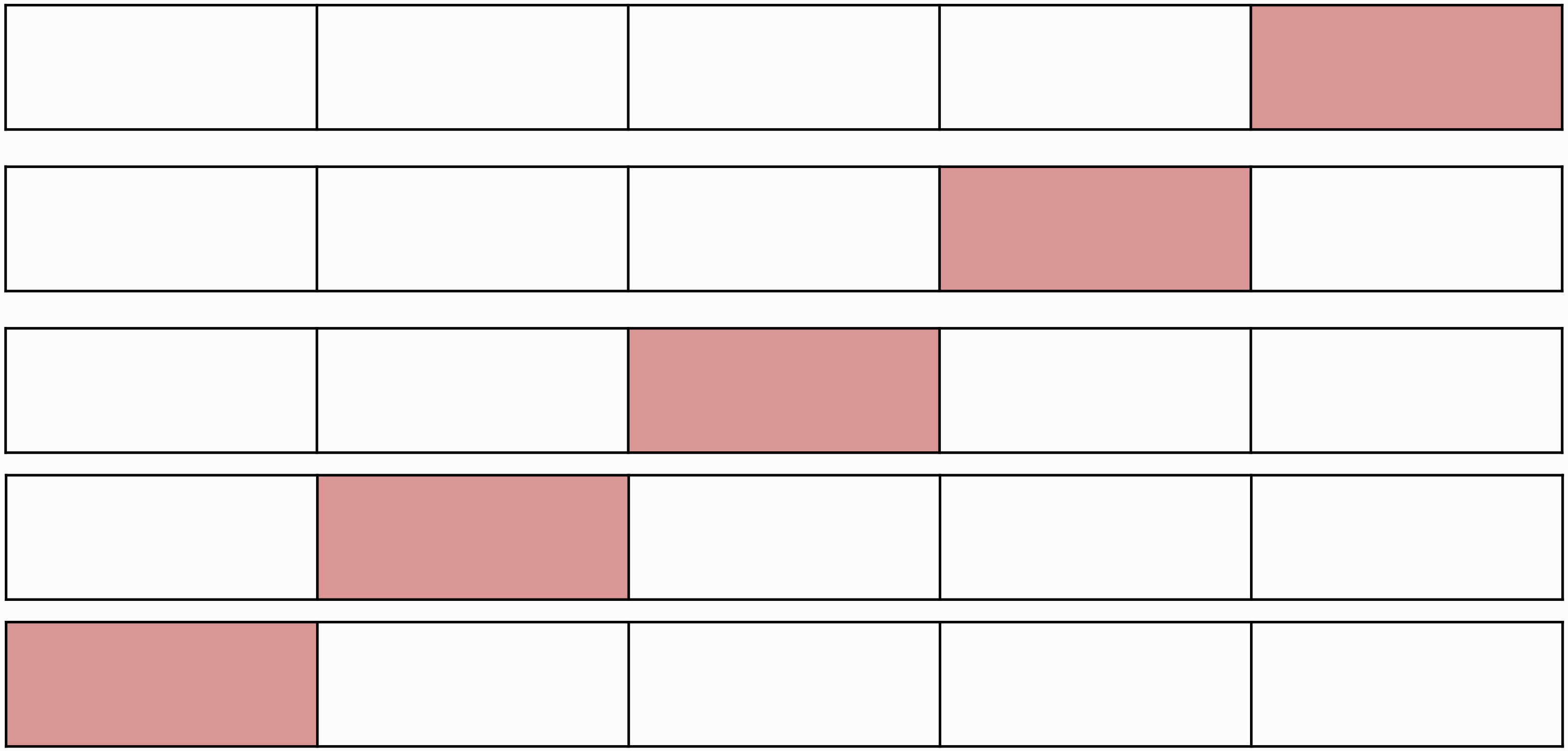
Train, T^t	Test, $T^{\ell-t}$

Solve the optimization problem:

$$\mathrm{HO}\left(\mu, T^{t}, T^{\ell-t}\right) = Q\left(\mu(T^{t}), T^{\ell-t}\right) \to \min$$

Cross-validation

Split sample to k parts k times



Complete cross-validation

Choose value of t.

Split the sample with all the possible ways on T^t and $T^{\ell-t}$.

Solve the optimization problem:

$$\mathrm{CVV}_t = \frac{1}{C_\ell^{\ell-t}} \sum_{T^\ell = T^{\ell-t} \cup T^t} Q\left(\mu(T^t), T^{\ell-t}\right) \to \min$$

k-fold cross-validation

k-fold cross-validation

Each of k blocks is a test sample once.

k is usually 10 (5 is small sample size).

Split $T^\ell = F_1 \cup \dots \cup F_k$, $|F_i| \approx \frac{\ell}{k}$.

Solve the optimization problem:

$$\mathrm{CV}_k = \frac{1}{k} \sum_{i=1}^{k} Q\left(\mu\left(T^\ell \setminus F_i\right), F_i\right) \to \min.$$

$t \times k$-fold cross-validation

Repeat t times: split sample on k blocks, each of k blocks is a test sample once.

k is usually 10 , t is usually 10 or less.

Split T^ℓ t times randomly:

$T^\ell = F_{(1,1)} \cup \dots \cup F_{(k,1)} = \dots = F_{(1,t)} \cup \dots \cup F_{(k,t)}$,

$\left|F_{(i,j)}\right| \approx \frac{\ell}{k}$.

Solve the optimization problem:

$$\mathrm{CV}_{t\times k} = \frac{1}{tk} \sum_{j=1}^{t} \sum_{i=1}^{k} Q\left(\mu\left(T^\ell \backslash F_{(i,j)}\right), F_{(i,j)}\right) \to \min.$$

Leave one out

Leave-one-out cross-validation, LOO

Split sample into $\ell - 1$ and 1 objects ℓ times.

Train, $T^{\ell-1}$	$\{x_i\}$

Solve the optimization problem:

$$\text{LOO} = \frac{1}{\ell}\sum_{i=1}^{\ell} Q\left(\mu\left(T^{\ell}\backslash p_i\right), p_i\right) \to \min.$$

where $p_i = (x_i, y_i)$.

Lecture plan

- Overfitting
- Model evaluation
- Validation
- Classification performance measures
- Regression performance measures

Matryoshic structure

For tuning hyperparameters, you need to treat your train set as a new dataset.

Lecture plan

- Overfitting
- Model evaluation
- Validation
- Classification performance measures
- Regression performance measures

Contingency table

	Positive	Negative
Classified as positive	**TP** = True Positive	**FP** = False Positive
Classified as negative	**FN** = False Negative	**TN** = True Negative

FN in math. stat. – I type error

FP in math. stat. – II type error

P = **TP** + **FN** – number of positive examples

N = **FP** + **TN** – number of negative examples

Some definitions

Sensitivity or **Recall**:

$$\text{Recall} = \text{TPR} = \frac{\text{TP}}{\text{P}}$$

Specificity:

$$\text{SPC} = \frac{\text{TN}}{\text{N}}$$

Precision:

$$\text{Precision} = \text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Accuracy:

$$\text{Accuracy} = \text{ACC} = \frac{\text{TP} + \text{TN}}{\text{P} + \text{N}}$$

F-measure

We will not lose much in accuracy performing badly on small classes.

F_β-measure:

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}}$$

F_1-measure:

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

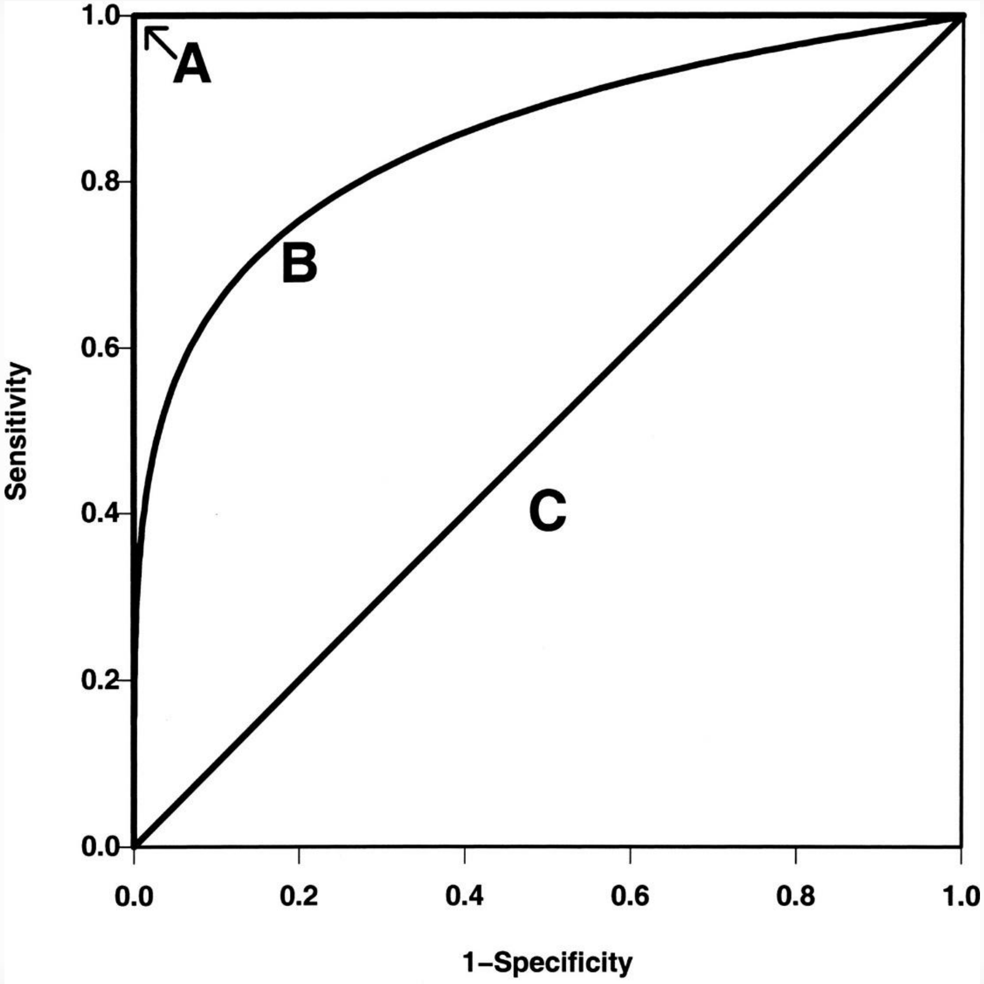
ROC-curve

A is the best algorithm

B is a typical algorithm

C is the worst algorithm

FPR vs TPR



AUC

Area under the curve (AUC) is area under the ROC-curve.

Connected with Mann-Whitney U.
Can be expressed with Gini-index.

Out of date measure.

Multiclass case

- One vs one classification
- One vs all (one vs rest) classification
- Hierarchical classification
- Confusion matrix

Lecture plan

- Overfitting
- Model evaluation
- Validation
- Classification performance measures
- Regression performance measures

Errors

- Root mean squared error (RMSE)
- Mean absolute error (MAE)
- Mean squared error (MSE)