Data Mining
(Mining Knowledge from Data)

Model evaluation

Marcel Jiřina, Pavel Kordík

ČESKÉ
VYSOKÉ
UČENÍ
TECHNICKÉ
V PRAZE

FIT

Classification

- Classification has 2 stages:
  - Learning (Training)
  - Recalling (Using, Applying the model)
- Our goal is to create a classifier with the best success in recalling stage
- The success of the recalling stage we can not measure, when we do not know all the data on which the classifier has been learned
- How to estimate the error of the classifier during the recall stage?

How to estimate error

- Calculating errors on the training data
  - We assume that the training data are a representative sample
  - But, learning of most of models is based on setting of parameters so that to minimize the error on the training data.
  - If we select a classifier that is the best for specific data, we can no longer use these data to evaluate the classifier -> risk of classifier overfitting.
  - This is not the appropriate method - do not use it!

Overfitting

- More complex models adapt more to the training data and thereby reduce the error on the training data
- The model is then trained accurately on specific data and thus loses its ability of generalization - for additional data from the same distribution the error raises steeply

Overfitting

- More complex models are more susceptible to overfitting
- Simpler models have a lower risk of overfitting, but are not able to cover more complex dependencies in the data
- The decision to select a simpler or more complex model is ambiguous and depends on specific data
- How to recognize that a model is overfitted?
  - If there is an error on the training data much smaller than on the test data, then the model is overfitted.

How to estimate error

- By splitting the training data into 2 parts:
  - Training data – The model is trained on the training data
  - Test data – the error of classification is calculated on the test data
- How to divide the training set?
  - 80 % training and 20 % test ...
  - By reducing the training set it is harder to learn a classifier and the classifier error increases
  - On the contrary, a small test set does not allow an accurate error detection
  - We balance between a more accurate classification and a better estimate of the error

Distribution

Sepal length	Sepal Width	Petal length	Petal Width	Species
5.1	3.5	1.4	0.2	Setosa
4.9	3.0	1.4	0.2	Setosa
4.7	3.2	1.3	0.2	setosa
7.0	3.2	4.7	1.4	Versicolor
6.4	3.2	4.5	1.5	versicolor
6.9	3.1	4.9	1.5	Versicolor

train

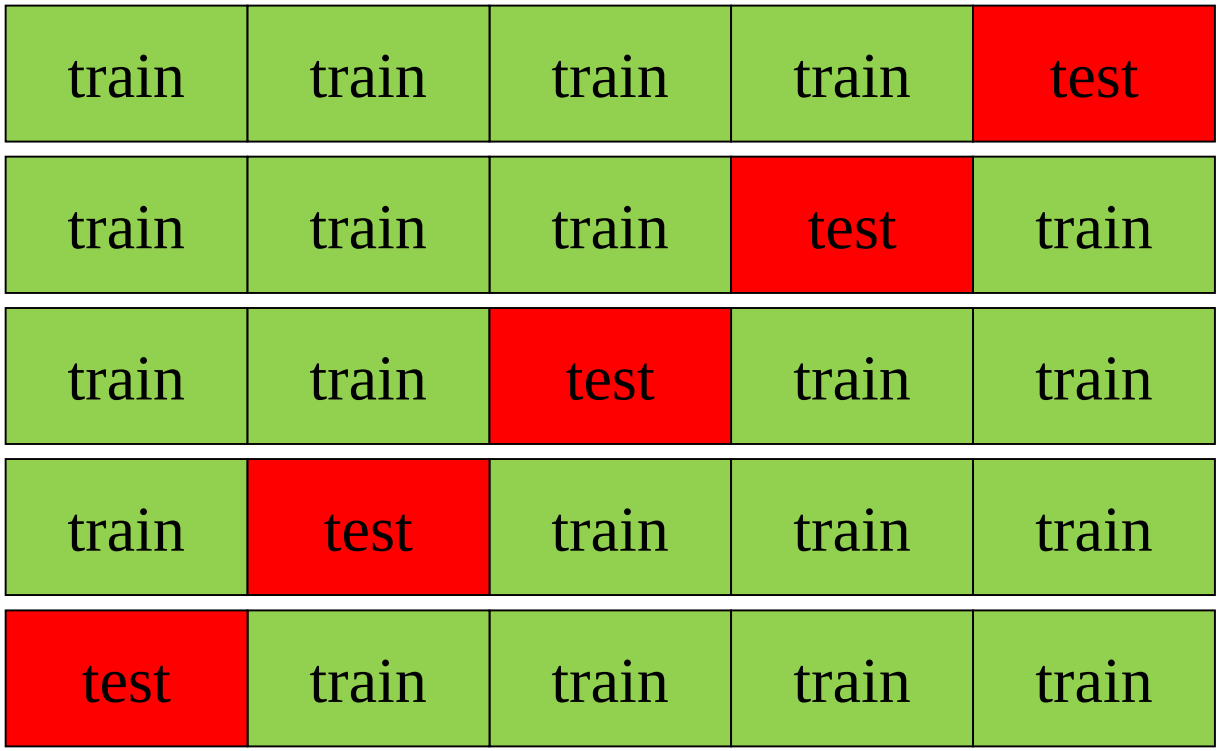
test

- Not this way!
- Data should be divided into training and test sets randomly.

Cross-validation (X-validation)

- By splitting data into training and test set, the ability to use a part of the data for learning is lost
- In cross-validation, the data are divided into N equal parts (folds)
- Repeat N-times:
  - Use N-1 parts for learning
  - Use 1 part for testing (error estimation)
- The resulting error is the average of N partial errors
- Each instance (pattern, case) is used N-1 times for learning, and once for the error calculation

Cross validation N=5



-> err_1

-> err_2

-> err_3

-> err_4

-> err_5

$$err = \frac{\sum_{i=1}^{N} err_i}{N}$$

Cross validation

- How the previous method (50%) test (50%) differs from the 2x cross validation?
- Can the cross validation be used to test the error of a single model?
- The cross-validation is dedicated to test the quality of the algorithm, which generates the models.
- What if, despite the use of the cross-validation, we can not statistically significantly distinguish the algorithms?
  - Increase the number of parts to N = number of instances (leave-one-out CV)
  - Use the Bootstrap validation

What is Bootstrap?

X=(3.12, 0, 1.57, 19.67, 0.22, 2.20)

Mean=4.46

X1=(1.57,0.22,19.67, 0,0,2.2,3.12)

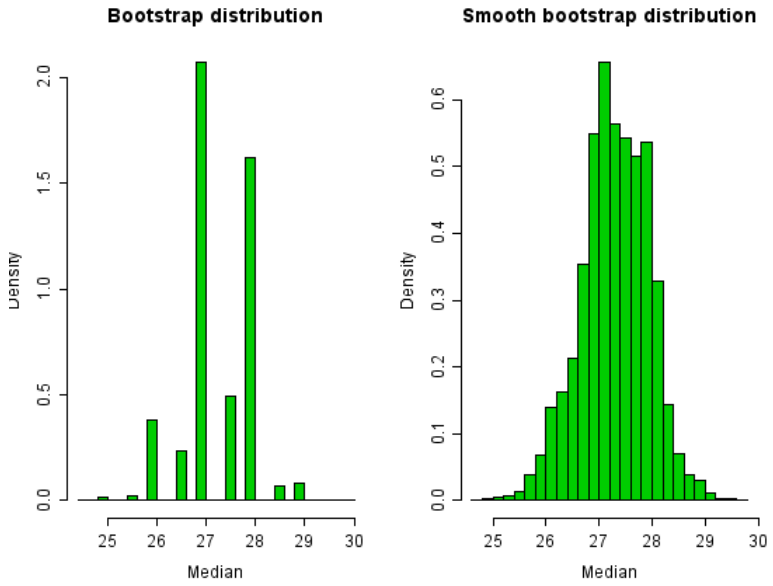
Mean=4.13

X2=(0, 2.20, 2.20, 2.20, 19.67, 1.57)

Mean=4.64

X3=(0.22, 3.12,1.57, 3.12, 2.20, 0.22)

Mean=1.74



statistics:

– confidence interval estimation

Examples of bootstrap (Opitz, 1999)

Training samples	1	2	3	4	5	6	7	8
Sample 1	2	7	8	3	7	6	3	1
Sample 2	7	8	5	6	4	2	7	1
…	…	…	…	…	…	…	…	…
Sample M	4	5	1	4	6	4	3	8

Bootstrap validation – it is necessary to ensure that the test data would not be used for the model learning

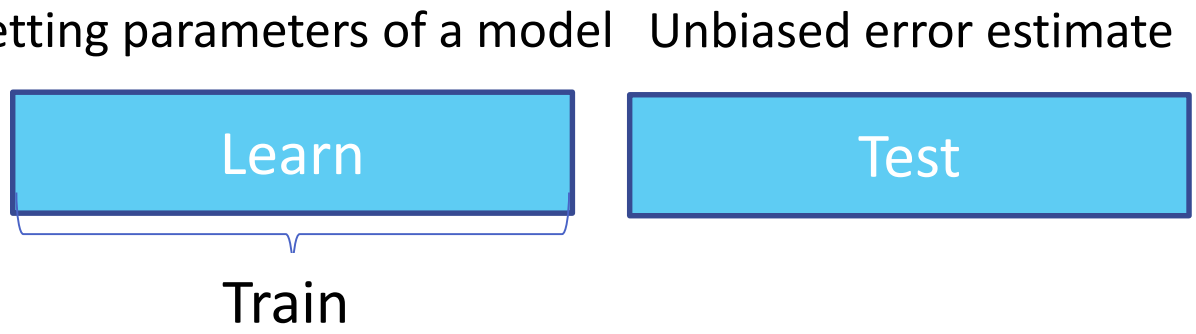
Validation data - problem

- Using the test data we test several models and choose the one with the smallest error.
- As we used the test data for model selection, we actually used them during the learning stage.
- Tis way, the test data become the validation data.
- The estimation of errors of the best model on these test data is no longer unbiased!
- If we want to compare this model with another model, new data has to be used.

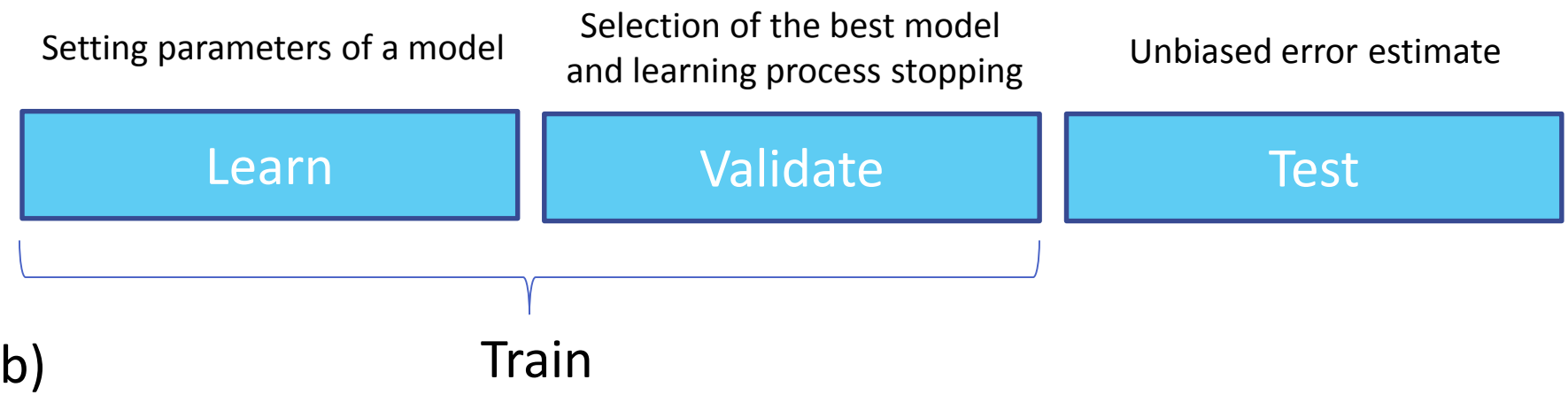
Learn – validate – test

- Attention: In various books, parts of data are referred different ways...
- Variants a) without and b) with using validation data

a)



b)



How to measure success of classification

- Direct solution:
  - $accuracy = \frac{number\ of\ correctly\ clasified\ patterns}{total\ number\ of\ patterns}$

- Analogy:
  - $error\ rate = \frac{number\ of\ incorrectly\ clasified\ patterns}{total\ number\ of\ patterns}$

- Frequently used but inadequate if the classes are unbalanced (unequal representation of classes).
- Why? What does it mean?

Problem with “accuracy”

- Example:
  - We want to create a classifier, which decides whether the credit card transaction is fraudulent (stolen credit card)
  - Fraudulent transactions are very small in comparison with other transactions - only 0.01 %
  - A trivial classifier, which marks all transactions as correct, will have a very high success rate - 99.99 %
  - Such a classifier is useless

confusion matrix (contingency table)

- Binary classification is supposed

		Reality (observed values)	
		P	N
Classification (predicted values)	P	TP	FP
	N	FN	TN

- TP – true positives, correct classification to class P
- TN – true negatives, correct classification to class N
- FP – false positives, incorrect classification to class P (it should be N)
- FN – false negatives, incorrect classification to class N (it should be P)

confusion matrix, derived characteristics

		Reality (observed values)	
		P	N
Classification (predicted values)	P	TP	FP
	N	FN	TN

- true positive rate (sensitivity, recall) $TPr = \frac{TP}{TP+FN}$
- true negative rate $TNr = \frac{TN}{TN+FN}$
- false positive rate $FPr = \frac{FP}{FP+TN}$
- false negative rate $FNr = \frac{FN}{TN+FP}$
- specificity $spec = \frac{TN}{TN+FP}$
- precision $pre = \frac{TP}{TP+FP}$
- accuracy $accuracy = \frac{TP+TN}{TP+TN+FP+FN}$

F-measure

- F-measure is a compromise between precision and recall

$$\frac{2*Pre*TPr}{Pre+TPr}$$

- Calculate the F-measure for cases
  - Pre = TPr = 1 and
  - Pre = 0.1, TPr = 0.9

Example of a confusion matrix

- Iris data, 3 classes

accuracy: 91.33% +/- 9.45% (mikro: 91.33%)

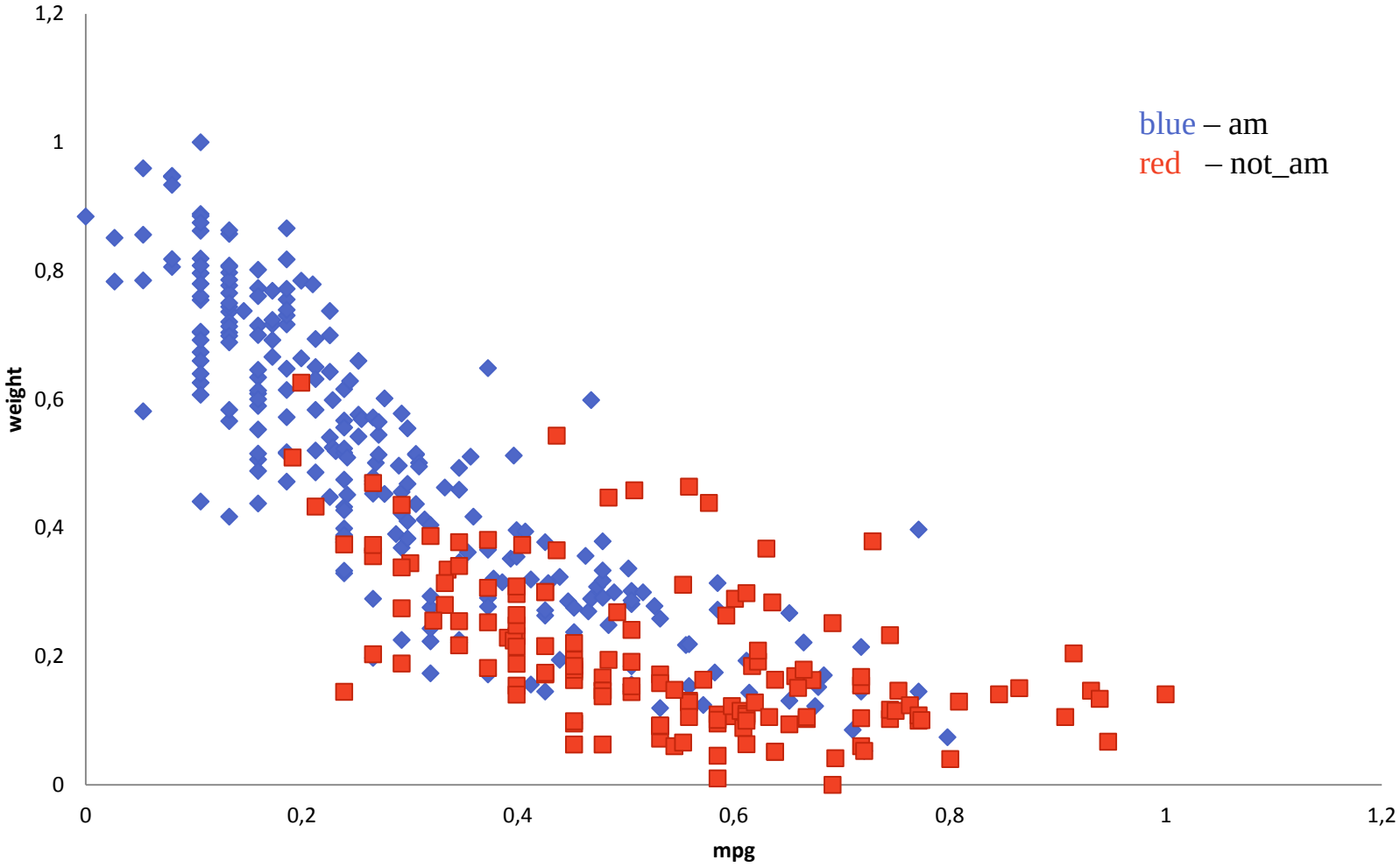
	true Iris-setosa	true Iris-versicolor	true Iris-virginica	class precision
pred. Iris-setosa	50	0	0	100.00%
pred. Iris-versicolor	0	47	10	82.46%
pred. Iris-virginica	0	3	40	93.02%
class recall	100.00%	94.00%	80.00%	

Example – classification of cars

- We have data with various pieces of information about 392 cars
- The goal is to determine whether the car has been made in USA

mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin
0	1	0,609819	0,798913	0,884321	0,625	0	am
0,026596	1	0,754522	0,918478	0,851148	0,357143	0	am
0,026596	1	0,617571	0,836957	0,783385	0,416667	0	am
0,053191	1	0,645995	0,891304	0,785086	0,327381	0	am
0,053191	1	0,932817	0,880435	0,856252	0,178571	0,166667	am
0,332447	0,2	0,136951	0,347826	0,279841	0,285714	0,583333	not_am
0,335106	0,2	0,136951	0,375	0,335129	0,458333	0,666667	not_am
0,398936	0,2	0,170543	0,271739	0,308761	0,327381	0,416667	not_am
0,398936	0,2	0,131783	0,277174	0,264247	0,535714	0,416667	not_am
0,404255	0,6	0,20155	0,402174	0,373405	0,345238	0,916667	not_am
...							

- Scatter plot for atributues
  - mpg – miles per gallon
  - weight



- Comparison of results of two trivial classifiers:
  1. mpg < 0,35 => the origin of the car is USA
  2. All cars has been made in USA

1. classifier – based on the mpg attribute

- mpg < 0,35 => the origin of the car is USA
- Confusion matrix:

TP – American cars marked as American cars

		Reality	
		American	Non-American
Classification	American	169	24
	Non-American	75	124

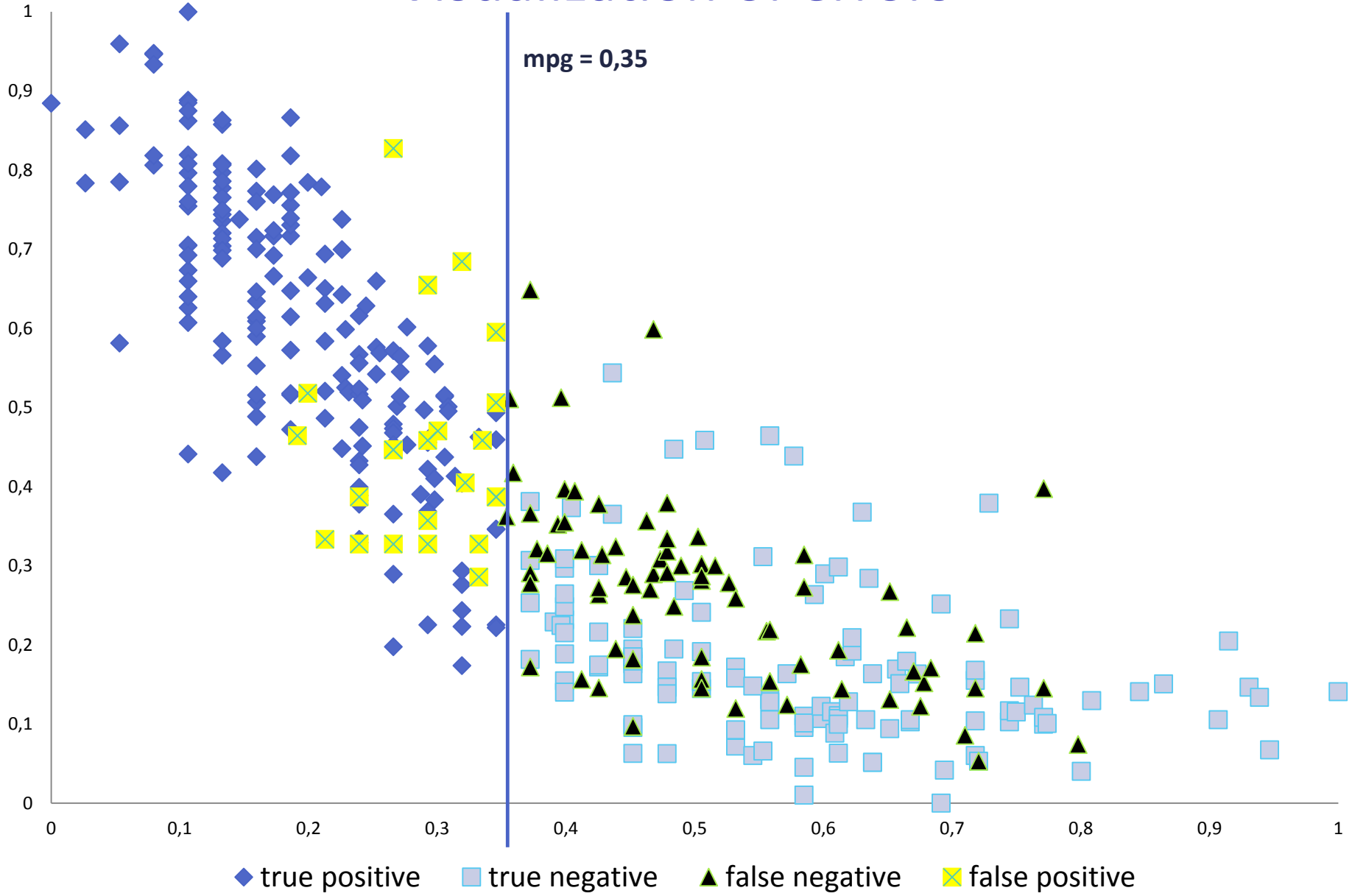
FP – Non-American cars incorrectly marked as American cars

FN – American cars incorrectly marked as non-American cars

TN – Non-American cars marked as Non-American cars

- Percentage of correctly classified cars (accuracy)
  - $\frac{169+124}{392} = 74{,}74\ \%$
- Values on the diagonal – correct classification
- Values aside from the diagonal – errors

Visualization of errors



		Reality	
		American	Non-American
Classification	American	169	24
	Non-American	75	124

- Precision
  - How many cars classified as American cars were really American cars?
  - $precision_{american\ cars} = \frac{169}{169+24} = 0,875$
  - How many cars classified as non-American cars were really non-American cars?
  - $precision_{non-american\ cars} = \frac{124}{124+75} = 0,623$
- Recall (completeness)
  - How many American cars has been marked as American cars?
  - $recall_{american\ cars} = \frac{169}{169+75} = 0,692$
  - How many non-American cars has been marked as non-American cars?
  - $recall_{non-american\ cars} = \frac{124}{124+24} = 0,838$

2. Classifier – All cars are American cars

- Confusion matrix

		Reality	
		American	Non-American
Classification	American	245	148
	Non-American	0	0

TP – American cars correctly marked as American cars

FP – Non-American cars incorrectly marked as American cars

- Percentage of correctly marked cars (Accuracy)
  - $\frac{245+0}{392} = 62{,}24\ \%$
  - $> 50\ \%$ because there are more American cars in the data set than non-American cars

		Reality	
		American	Non-American
Classification	American	245	148
	Non-American	0	0

- Precision
  - How many cars classified as American cars were really American cars?
  - $precision_{american\ cars} = \frac{245}{245+148} = 0{,}62$
- Recall (completeness)
  - How many American cars has been marked as American cars?
  - $recall_{american\ cars} = \frac{245}{245+0} = 1$
  - How many non-American cars has been marked as non-American cars?
  - $recall_{non-american\ cars} = \frac{0}{0+148} = 0$

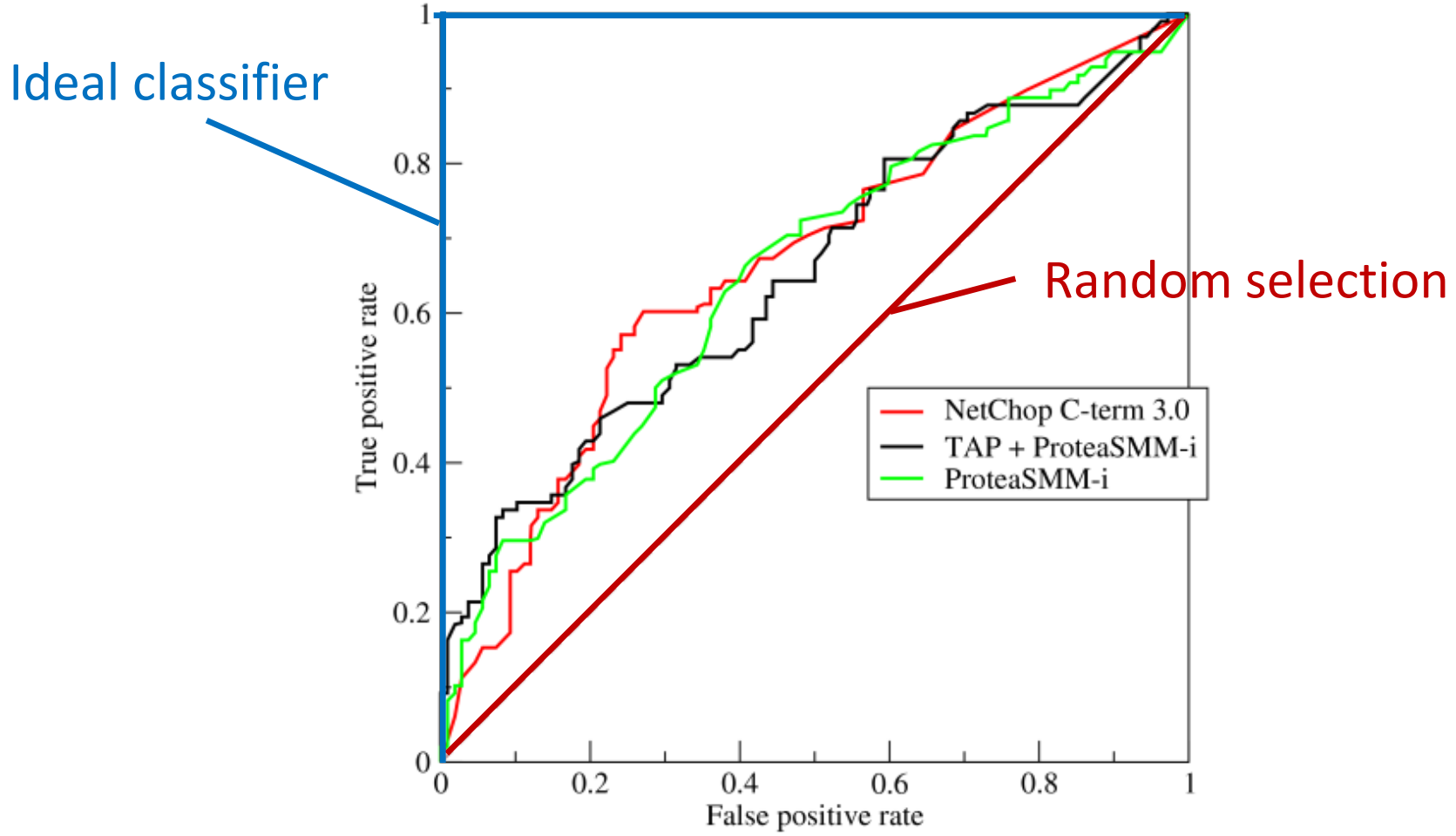
Threshold

- We can usually easily increase TPr at the expense of TNr and vice versa.
- How?
- In k-NN classifier we can do it using a threshold instead of the majority.
- Similarly for the other classifiers.

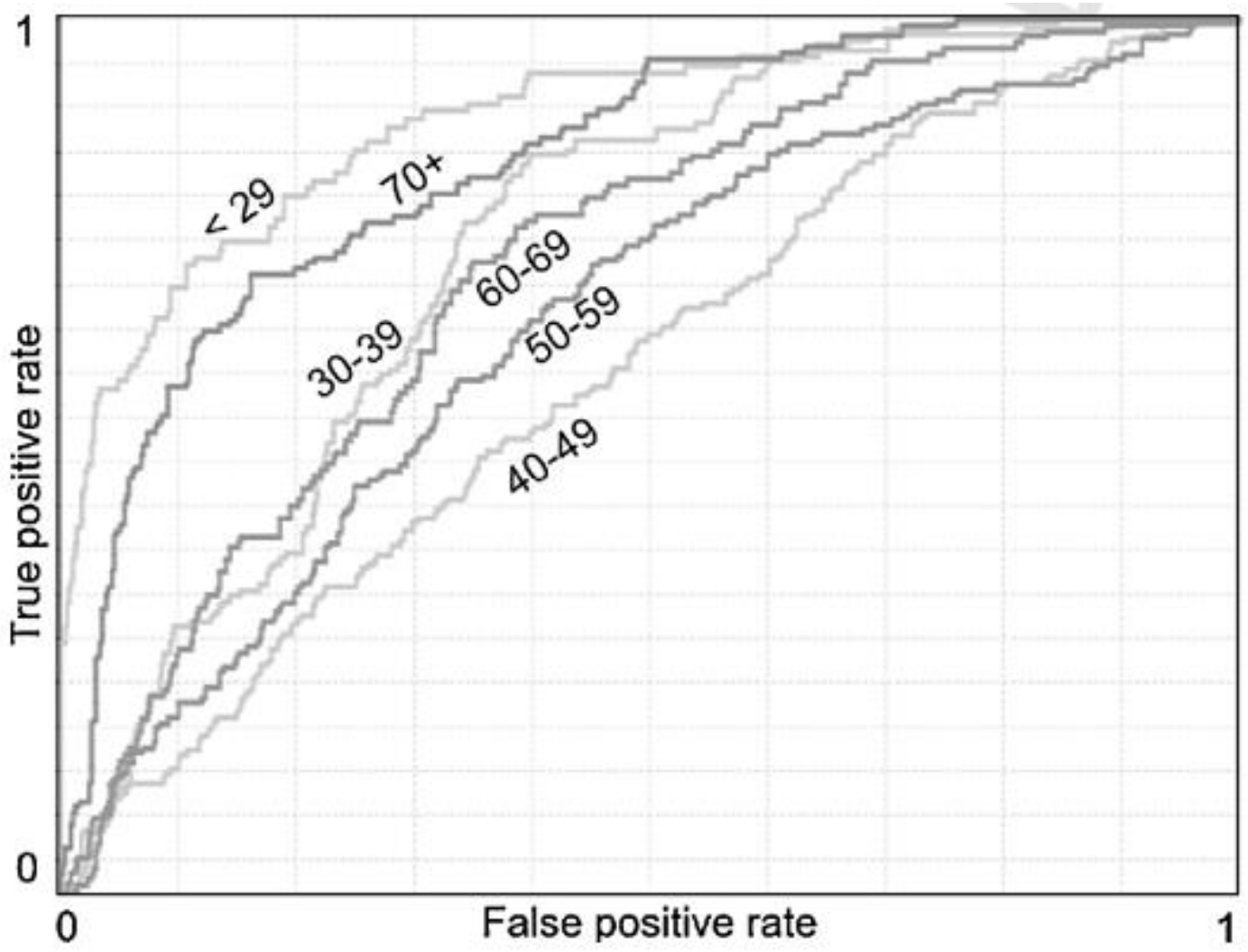
- Let's change a threshold from 0 to 1 and count TPr and TNr:

ROC

- Dependency of TPr on FPr is expressed by the ROC



What class is best separable?



AUC

- AUC = Area Under Curve
- 0,5 corresponds to random selection
- 1 corresponds to ideal (perfect) classifier

- http://www.anaesthetist.com/mnm/stats/roc/

Cost Sensitive Learning – CSL

- CSL – Learning that takes into account the fact that some errors are more serious than others.
- Is it better not to let enter a staff into a building or to let enter a terrorist?

- Cost matrix (function)
- Default form of the matrix is that there are *zeros* on the diagonal and *ones* off the diagonal

		Reality	
		Authorized	Denied
Classification	Authorized	0	10
	Denied	1	0

- When calculating errors, values in the respective fields of the cost matrix and confusion matrix are multiplied.

Recapitulation

- Train, test (Split validation)
- Train, validation, test
- X-validation
- Leave-one-out CV
- Bootstrap validation
- FP, FN, TP, TN, precision, recall, specificity, sensitivity
- F-measure
- Confusion matrix
- ROC
- AUC
- Cost-sensitive learning, cost matrix