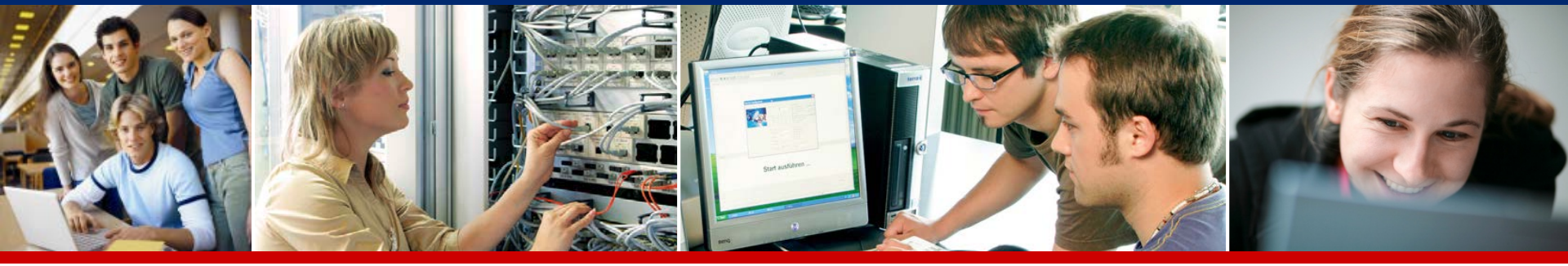


Model Evaluation

Prof. Dr. Stephan Trahasch

Offenburg University of Applied Sciences

1. Methods for Performance Evaluation
   How to obtain reliable estimates?

2. Metrics for Performance Evaluation
   How to evaluate the performance of a model?

3. Methods for Model Comparison
   How to compare the relative performance among competing models?

Evaluation: the key to success

To decide

- which model should be applied to which problem
- which trained classifier should be ultimately used

the models have to be be compared with each other.

How good are the forecasts of what has been learned?

- Error in the training data is not a good indicator of the quality of new data. Otherwise 1-NN would be the optimal classifier!
- Simple solution, split data into training & test set

But: mostly only limited amount of data is available

→ more sophisticated techniques necessary

Outline

- Methods for Performance Evaluation
- Metrics for Performance Evaluation
- Method for Model Comparison
  Receiver Operating Characteristic
- Summary

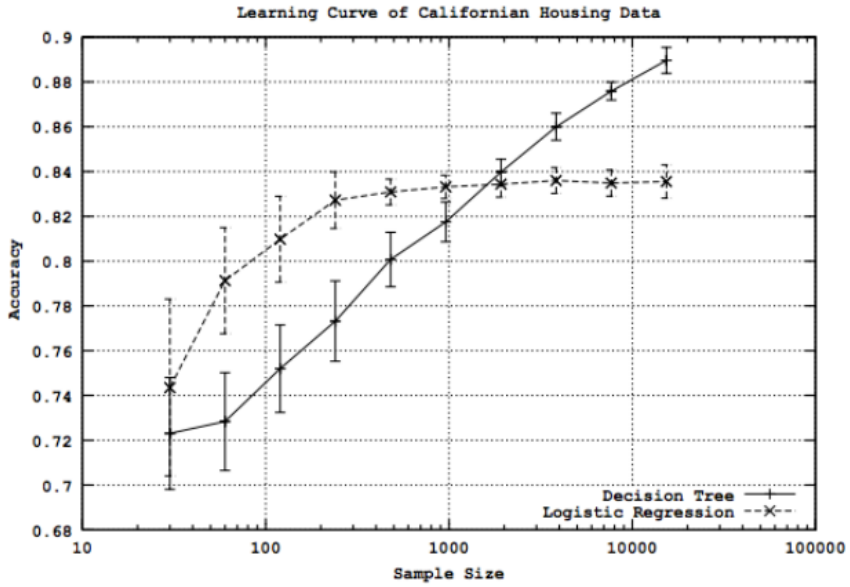
Methods for Performance Evaluation

- How to obtain a reliable estimate of performance?

- Performance of a model may depend on other factors besides the learning algorithm:
  - Class distribution
  - Cost of misclassification
  - Size of training and test sets

Learning Curves

given training/test set partition
    for each sample size s on learning curve
     (optionally repeat n times)
      – randomly select s instances from training set
      – learn model
      – evaluate model on test set to determine accuracy a
      – plot(s, a) or (s, avg.accuracy and error bars)



Learning Curves

How does the accuracy of a learning method change as a function of the training-set size?

This can be assessed by plotting learning curves

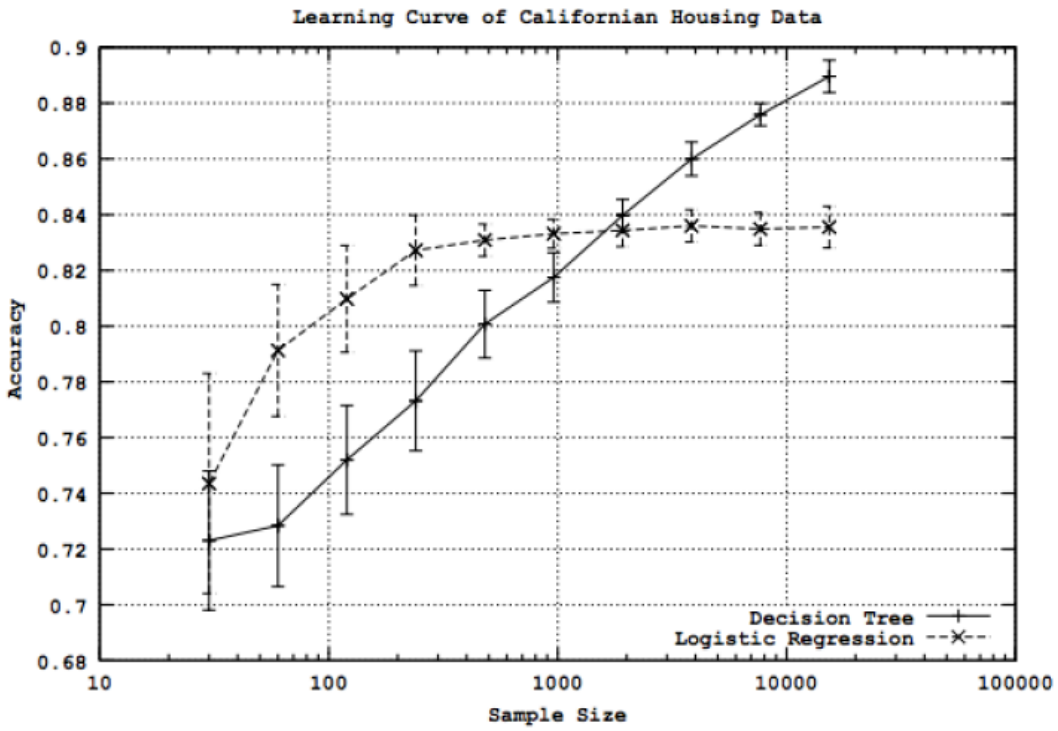
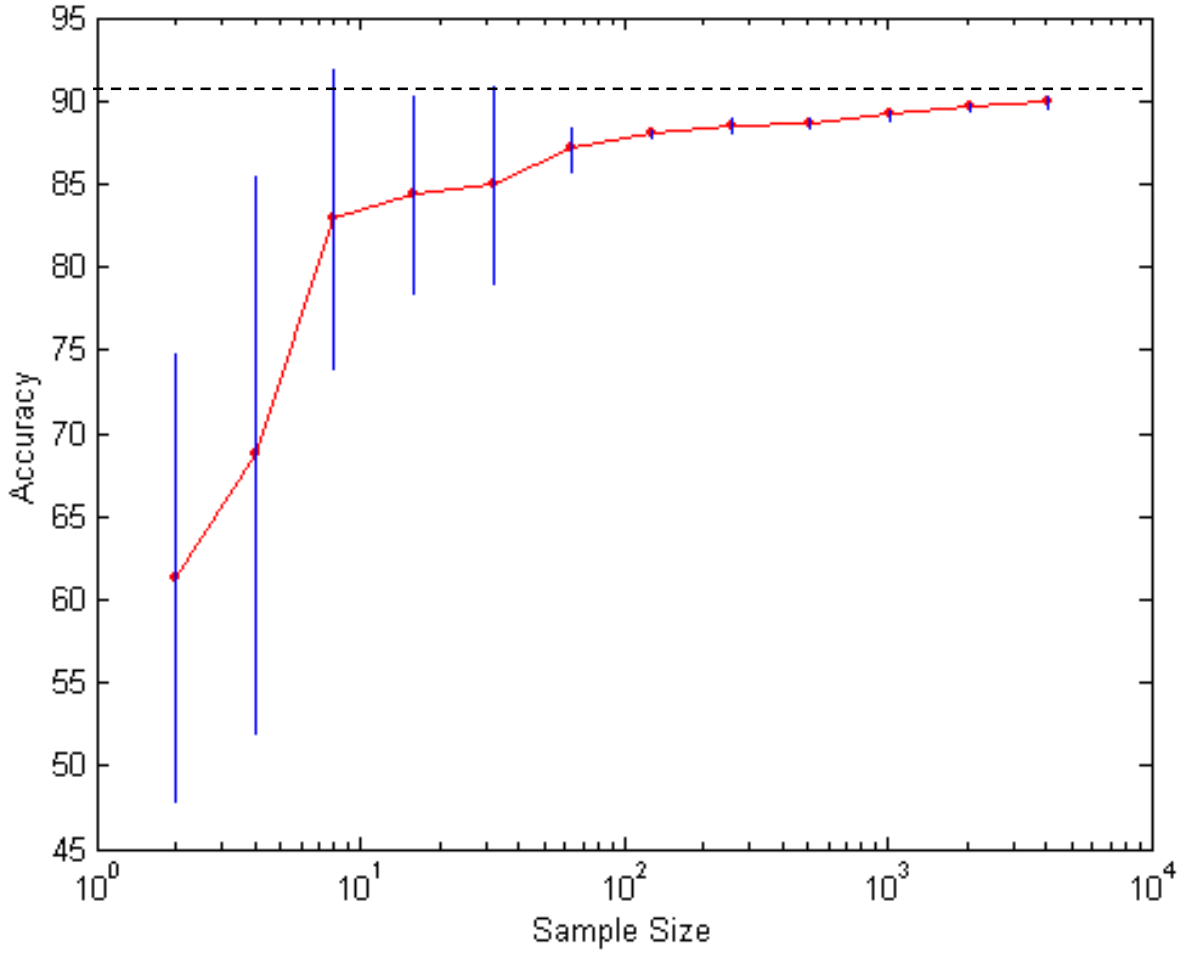


Figure from Perlich et al. Journal of Machine Learning Research, 2003

Example: Learning Curve



Learning curve shows how accuracy changes with varying sample size

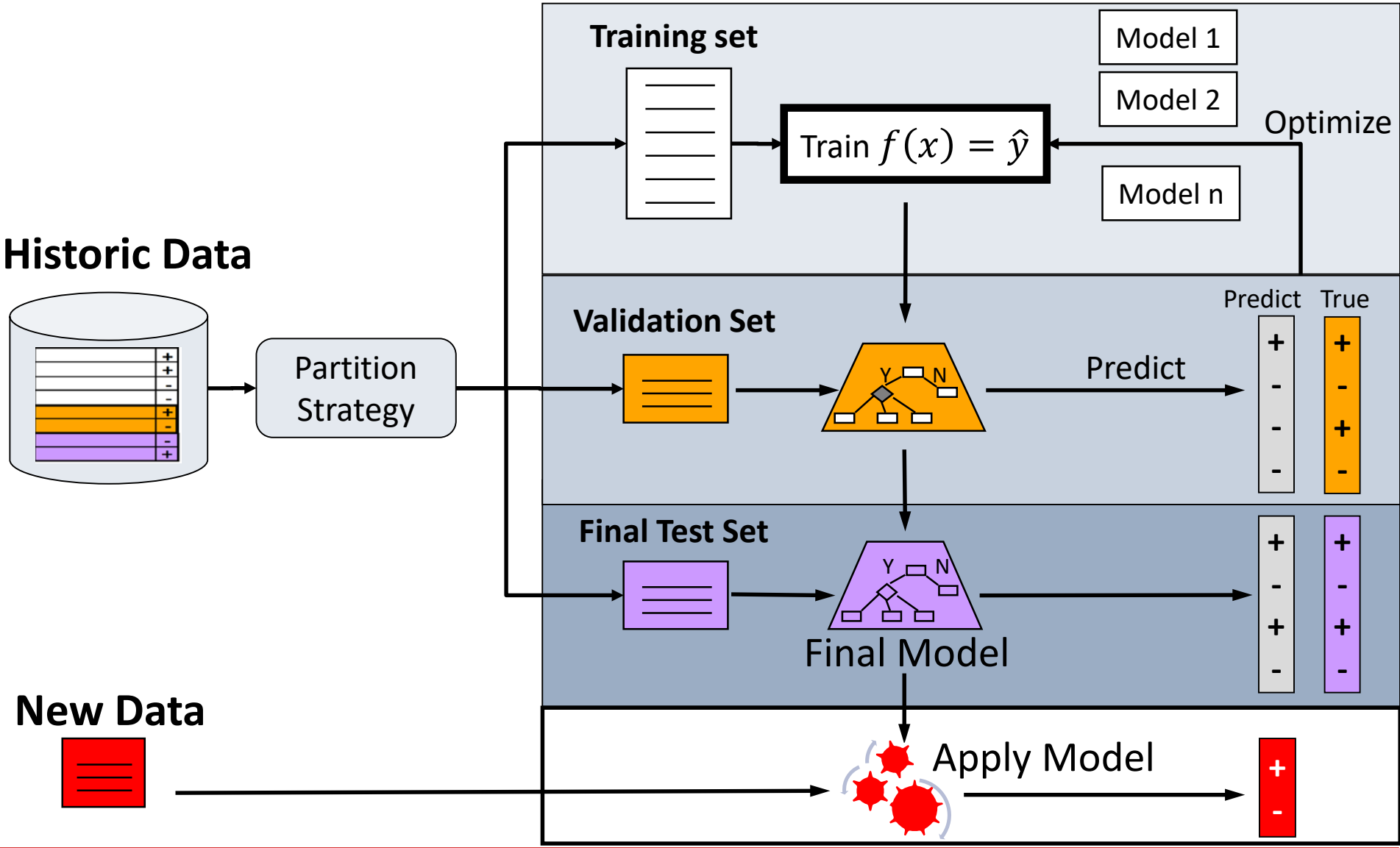
Effect of small sample size:

- Bias in the estimate
- Variance of estimate

Methods of Estimation

- Holdout
  - Reserve 2/3 for training and 1/3 for testing
- Stratification
  - oversampling vs undersampling
- Random subsampling
  - Repeated holdout
- Cross validation
  - Partition data into k disjoint subsets
  - k-fold: train on k-1 partitions, test on the remaining one
  - Leave-one-out: k=n
- Bootstrap
  - Sampling with replacement

Training, validation and test set



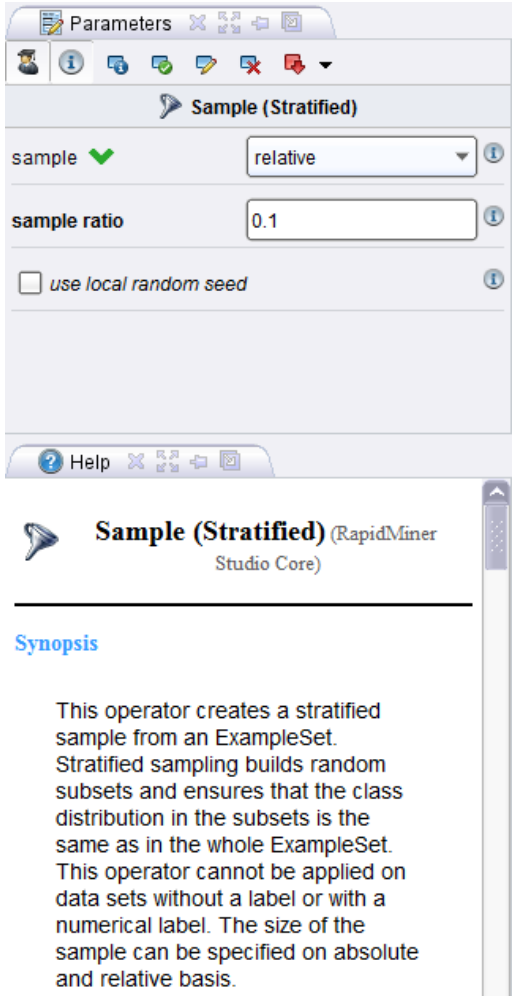
Stratification

Problem: The data selected for the sample may not be representative.

Example: Class does not appear in test data

- Divide the amount of data so that the relative class frequencies in test and training data are the same.
- Ensures that each class occurs in both subsets with approximately the same relative frequency

→ stratified Holdout



Repeated Holdout Method

Holdout estimation can be made more reliable by repeating the process with different samples.

In each iteration a certain amount of data is randomly selected for training (possibly with stratification)

The error rates of the different iterations are averaged to calculate a total error rate.

→ repeated Holdout

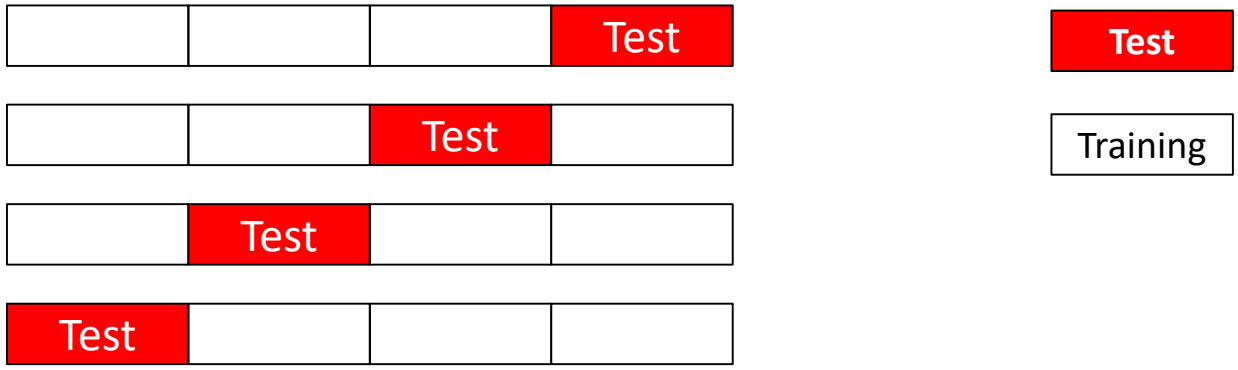
Still not optimal, because the different test quantities overlap.

Can overlaps be completely avoided?

Cross Validation

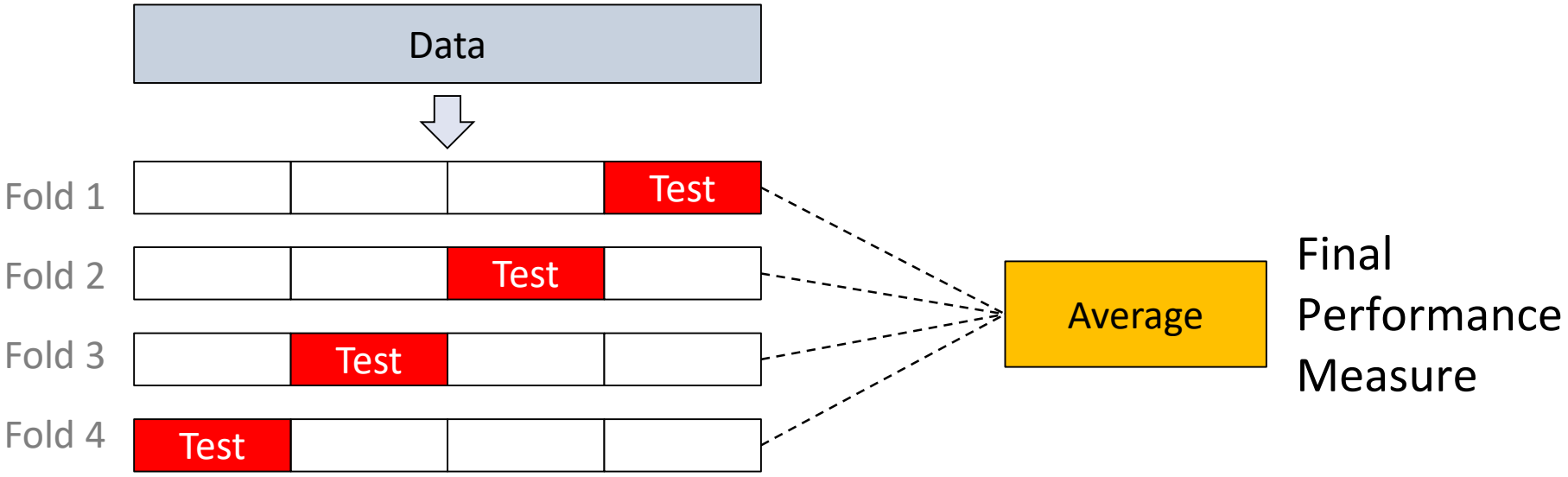
Cross validation avoids overlapping

- Divide data into k subsets of the same size
- Use each subset in turn for testing, rest for training
- Example: k = 4



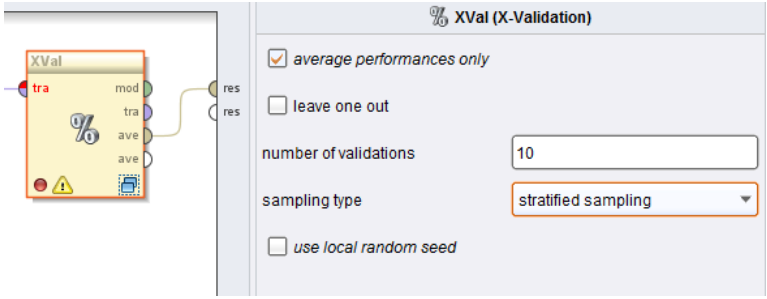
- Often, subsets are stratified before cross-validation is performed.
- Error rates are averaged to calculate the total error rate.

Cross Validation



Stratified 10 Cross Validation

Stratified 10-fold cross-validation
is the standard evaluation method!



Why 10?

- Extensive experiments have shown that this is the best choice to get reliable estimates.
- There are also theoretical reasons for this.
- Stratification reduces the variance of estimates
- Even better: Repeated stratified cross-validation
- E. g.: 10-fold cross-validation is repeated 10 times and the results are averaged (reduces the variance)

Leave-One-Out Cross Validation (LOOC)

- Leave-One-Out is a special form of cross-validation
- Number of executions = number of training instances
- This means that the classifier is learned n times for n training instances.
- Uses the data optimally
- No random sample selection!
- Drawback: large computing effort

Leave-One-Out-CV and Stratification

- Disadvantage of Leave-One-Out-CV:
  Stratification is not possible
- Procedure guarantees a non-stratified sample because the test set contains only one instance!

Extreme example: Data quantity in which two classes occur equally frequently

- Simple learner predicts the majority class
- 50% accuracy on fresh data
- Leave-One-Out-KV would deliver 100% error rate

Bootstrap – also called 0.632 Bootstrap

Given a record with n instances (observations).
Bootstrap sample: draw a sample of the same length from the original dataset with replacement.

Original Data	1	2	3	4	5	6	7	8	9	10
Bootstrap 1	7	8	10	8	2	5	10	10	5	9
Bootstrap 2	1	4	9	1	2	3	2	7	3	2
Bootstrap 3	1	8	5	10	5	5	9	6	3	7
…										

For each set of size n, the probability that a given example appears in it is $\Pr(x \in B_i) = 1 - \left(1 - \frac{1}{n}\right)^n \xrightarrow[n\to\infty]{} 0.6322$

On average, less than 2/3 of the examples appear in any single bootstrap sample.

Outline

- Methods for Performance Evaluation
- Metrics for Performance Evaluation
  - Categorical Targets
  - Continuous Targets
- Method for Model Comparison
  Receiver Operating Characteristic
- Summary

Occam's Razor - Ockhams Rasiermesser

There are two models with the same error rate.

The simplest model (hypothesis) that can explain the data should be preferred.

Wilhelm von Ockham
1288–1347

Complex models are more likely to have adapted to errors in the data.

Model complexity should be taken into account in the evaluation.

Evaluation Aspects

Selection of the quality measure:

- Number of correct classifications
- Errors in numerical predictions
- Accuracy of probability estimates

Statistical reliability of observed differences in quality?

→ signifiance tests

Costs for different types of errors

→ Costs are relevant for many practical applications

Quality measure for classification problems: error rate

Focus on the predictive capability of a model.

Rather than how fast it takes to classify or build models, scalability, etc.

Correct: The class of an instance is correctly predicted.

Error: The class is incorrectly predicted.

Error Rate: Percentage of errors in decisions for a set of instances.

Is accuracy an adequate measure of predictive performance?

Accuracy may not be useful measure in cases where

- there is a large class skew
- Is 98% accuracy good if 97% of the instances are negative?

There are differential misclassification costs.

Getting a positive wrong costs more than getting a negative wrong

- Consider a medical domain in which a false positive results in an extraneous test but a false negative results in a failure to treat a disease

Confusion Matrix for a binary classifier

Let model M be a classifier for 2 classes: {Yes, No}

Predicted Class \ True Class	Yes	No
Yes	TP = True Positive	FP = False Positive
No	FN = False Negative	TN = True Negative

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

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

$$\text{Error rate} = \frac{FP + FN}{P + N}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

Confusion matrix for mulitple classes C_1,.., C_n

Comparison of the frequencies of the determined classes and the actual classes of a test case.

Predicted Class \ True Class	C1	C2	C3	C4	Σ
C1	12	7	3	4	**26**
C2	1	9	8	2	**20**
C3	6	4	9	13	**32**
C4	94	12	3	5	**114**
Σ	**113**	**32**	**23**	**24**	**192/192**

Accuracy:

Error rate:

Limitation of Accuracy

- Consider a 2-class problem
- Number of Class 0 examples = 9990
- Number of Class 1 examples = 10

If model predicts everything to be class 0,
accuracy is 9990/10000 = 99.9 %

Accuracy is misleading because model does not detect any class 1 example

Consideration of costs

In practical applications, different types of errors often lead to different costs.

Examples:

Cost Matrix

	True Class		
Predicted Class	**C(i|j)**	**Class=Yes**	**Class=No**
	Class=Yes	C(Yes|Yes)	C(No|Yes)
	Class=No	C(Yes|No)	C(No|No)

C(i|j): Cost of misclassifying class j example as class i

Computing Cost of Classification

Cost Matrix	True		
	C(i|j)	+	-
Predicted Class	+	-1	1
	-	100	0

Model M_1	True Class		
		+	-
Predicted Class	+	150	60
	-	40	250

Accuracy = 80%

Cost = 3910

Model M_2	True Class		
		+	-
Predicted Class	+	250	5
	-	45	200

Accuracy = 90%

Cost = 4255

Increase diagram

- In practice, the costs are often unknown
- Decisions are made by comparing different possible scenarios

Example: direct mail to 1,000,000 households

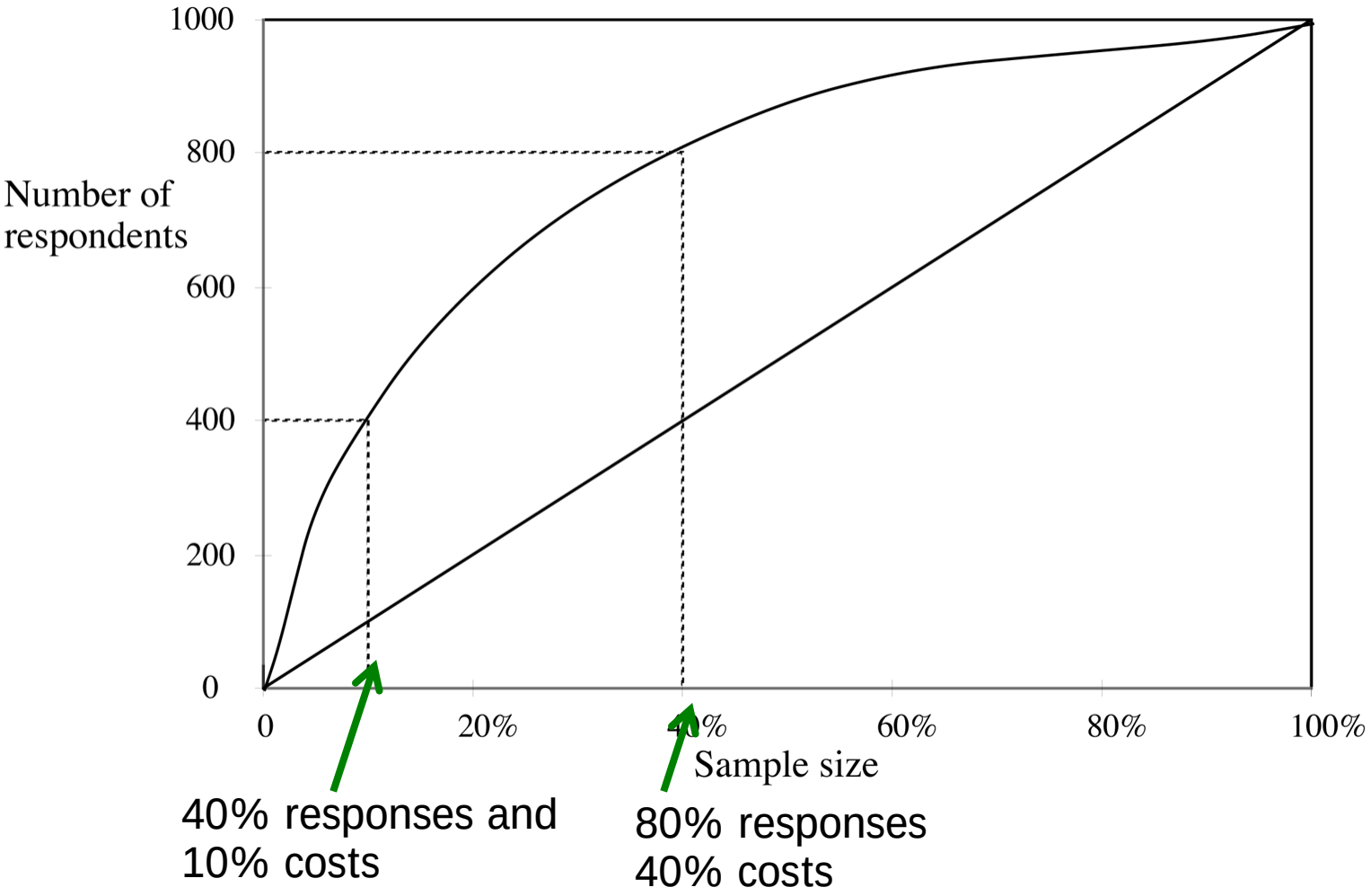
- Send to all: 0.1% reply (1000)
- Model identifies subset of 100,000 prospective sites, 0.4% of which respond (400) 40% of responses for 10% of the cost can be worthwhile
- Identify subset of 400,000 prospects, 0.2% of which answer (800)
- An increase diagram allows visual comparison

How to generate a increase diagram

Sort instances according to the estimated probability of success:

	Predicted probability	Actual class
1	0.95	Yes
2	0.93	Yes
3	0.93	No
4	0.88	Yes
...	...	...

Example


1000
800
600
400
200
0
Number of respondents
0
20%
60%
80%
100%
Sample size
40% responses and
10% costs
80% responses
40% costs

Outline

- Methods for Performance Evaluation
- Metrics for Performance Evaluation
  - Categorical Targets
  - Continuous Targets
- Method for Model Comparison
  Receiver Operating Characteristic
- Summary

Linear Regression

Prediction of a variable y by a variable x (vector).

Prediction is only possible if x and y are related, i. e. correlate with each other.

The variable y to be predicted is called criterion variable

The variable x used for prediction is called predictor variable

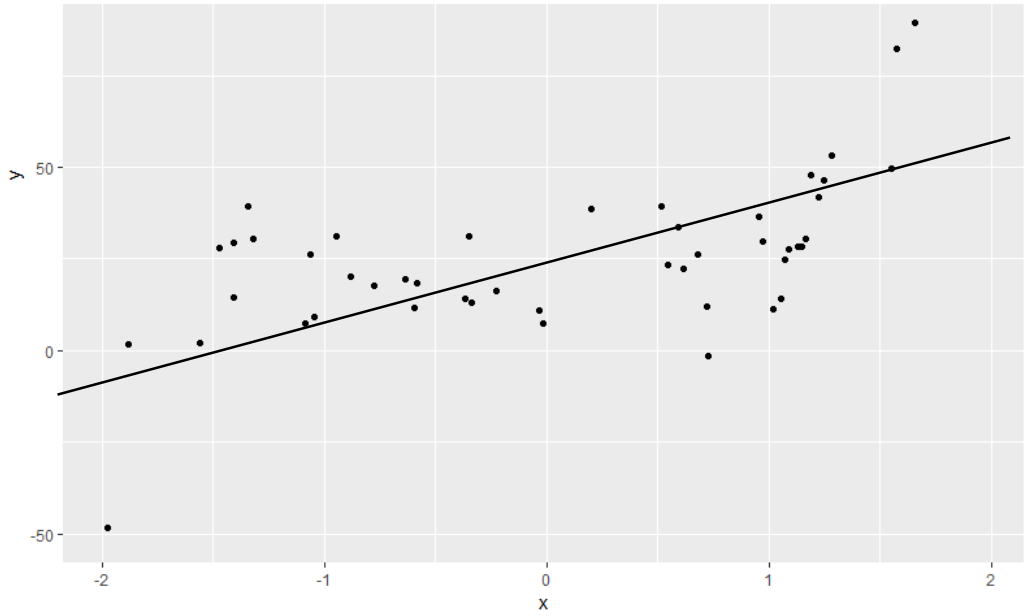
Application examples:

- Values of X have already been collected, values of Y are not known
- X can be recorded at the present time, Y only much later
- X is easy to acquire (easy, inexpensive, fast) and Y can only be collected by expensive, time-consuming examination

Linear Regression

Intuition:
A straight line is determined which describes the relationship between x and y.

With such a straight line, a value of y can be predicted for each value of x.



Linear Regression

General function of a line: $y = b \cdot x + a$

where b stands for the slope and a for the y-axis section.

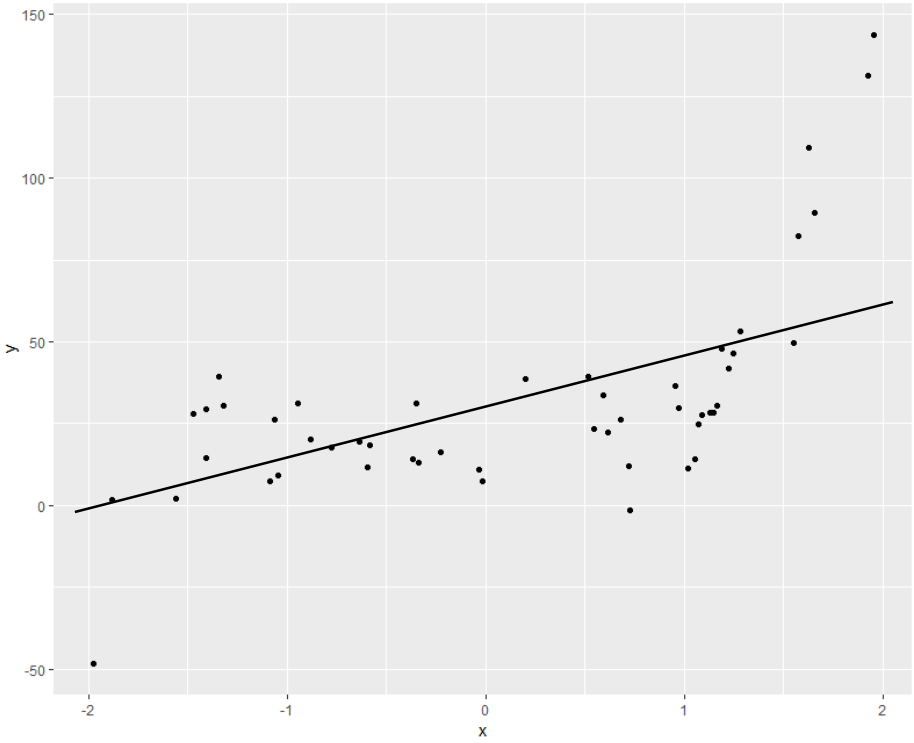
$$\hat{y}_i = \beta_1 \cdot x_i + \beta_0 + \varepsilon$$

$\hat{y}_i$: predicted value

β_1 : Regression coefficient

β_0 : additive constant

ε : noise $\quad \varepsilon \sim N(0, \sigma)$



Residual Sum of Squares - Least Squares Method

β_0 and β_1 are calculated in such a way that the squared prediction error is minimal across all predictions:

Residual Sum of Squares $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

Square deviation has two advantages...

1. Deviation values are always positive.
2. Large deviations are taken more into account than small deviations.

→ details later on the topic Linear Regression

Basic measures of error

Mean squared error $= \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}$

Root mean squared error $= \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$

Mean absolute error $= \frac{\sum_{i=1}^{n} abs((y_i - \hat{y}_i))}{n}$

Mean absolute percentage error $= \frac{100}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{|y_i|}$

Outline

- Methods for Performance Evaluation
- Metrics for Performance Evaluation
- Method for Model Comparison
  Receiver Operating Characteristic
- Summary

Comparing Learning Models

How can we determine if one learning model provides better performance than another …

- for a particular task?
- across a set of tasks / data sets?

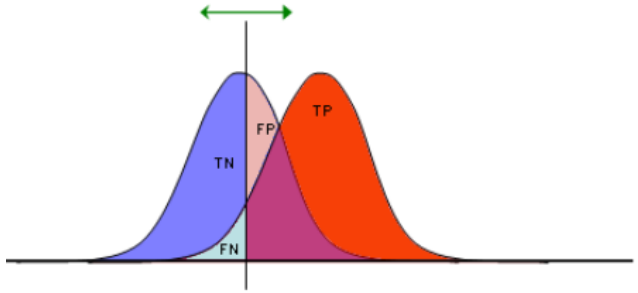
Example: **Accuracies on test sets**

Model 1:	80%	50	75	…	99
Model 2:	79	49	74	…	98
δ :	+1	+1	+1	…	+1

Mean accuracy for Model 1 is better, but the standard deviations for the two clearly overlap.

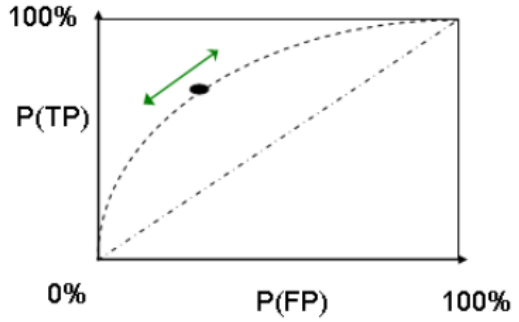
ROC Analysis – Receiver Operating Characteristic

Origin Signal theory: 2 signal sources.
To which source does
a received signal belong?



TP	FP
FN	TN
1	1

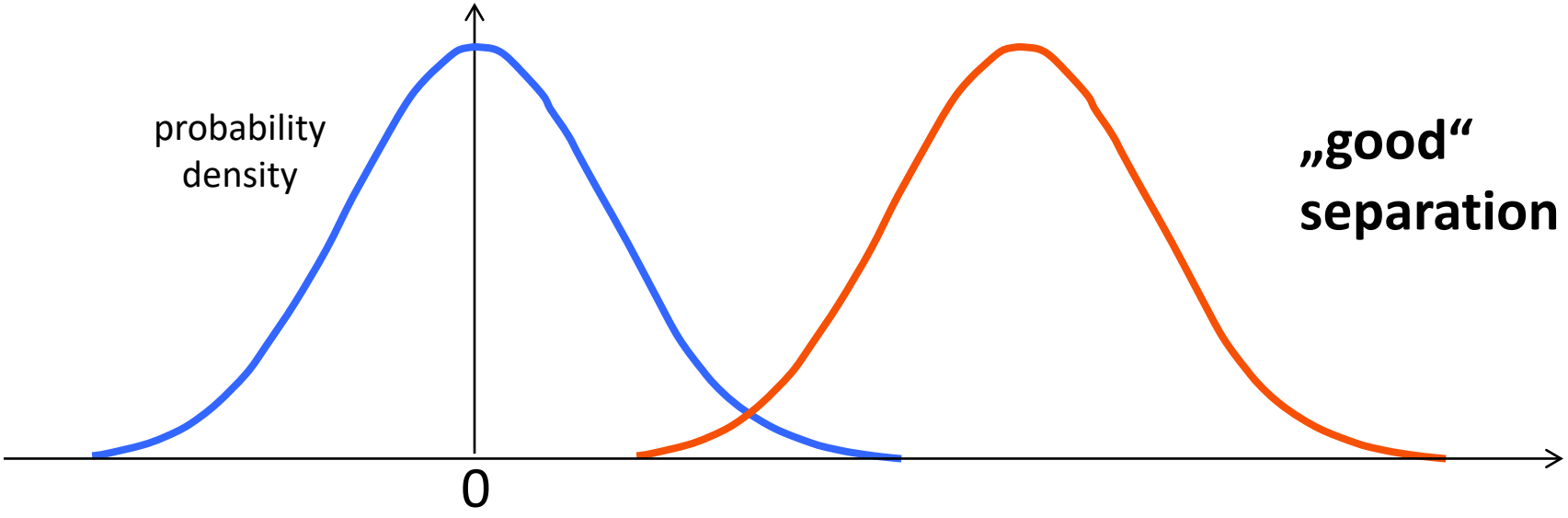
Objective:
What is the best method for
varying parameter values?



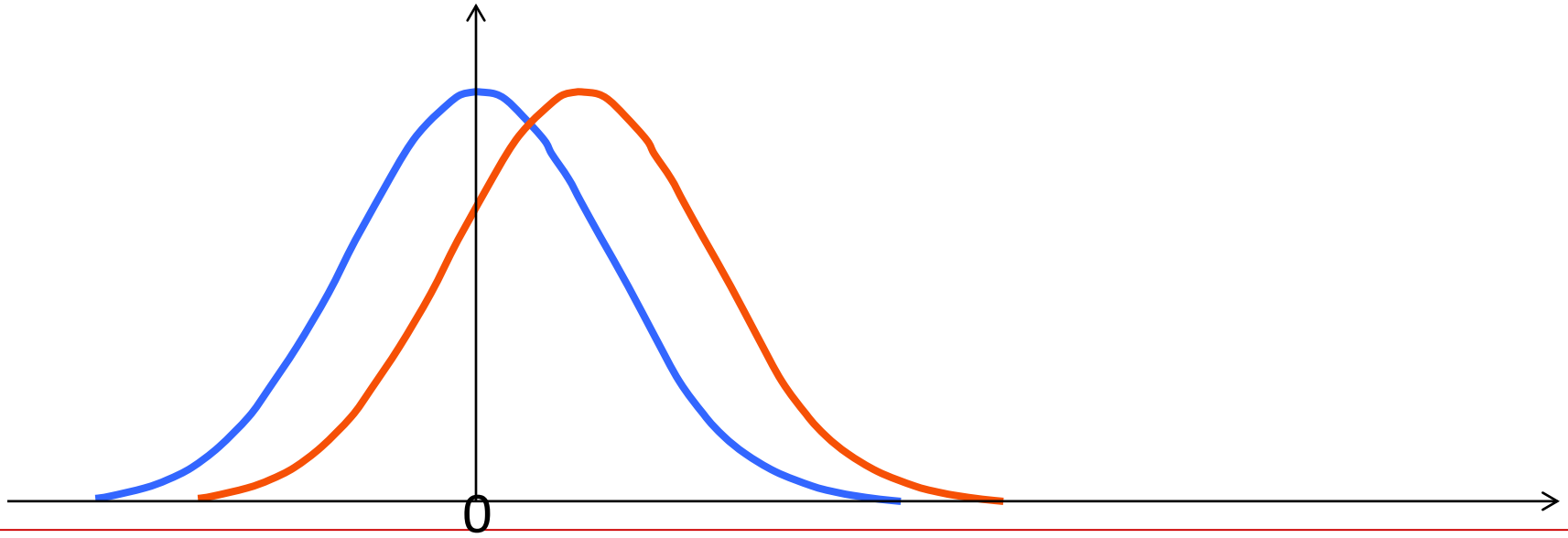
Method:

- Visualization as ROC curve
- fpr and tpr for each binary classifier
- x-axis: false positive rate fpr
- y-axis: true positive rate tpr

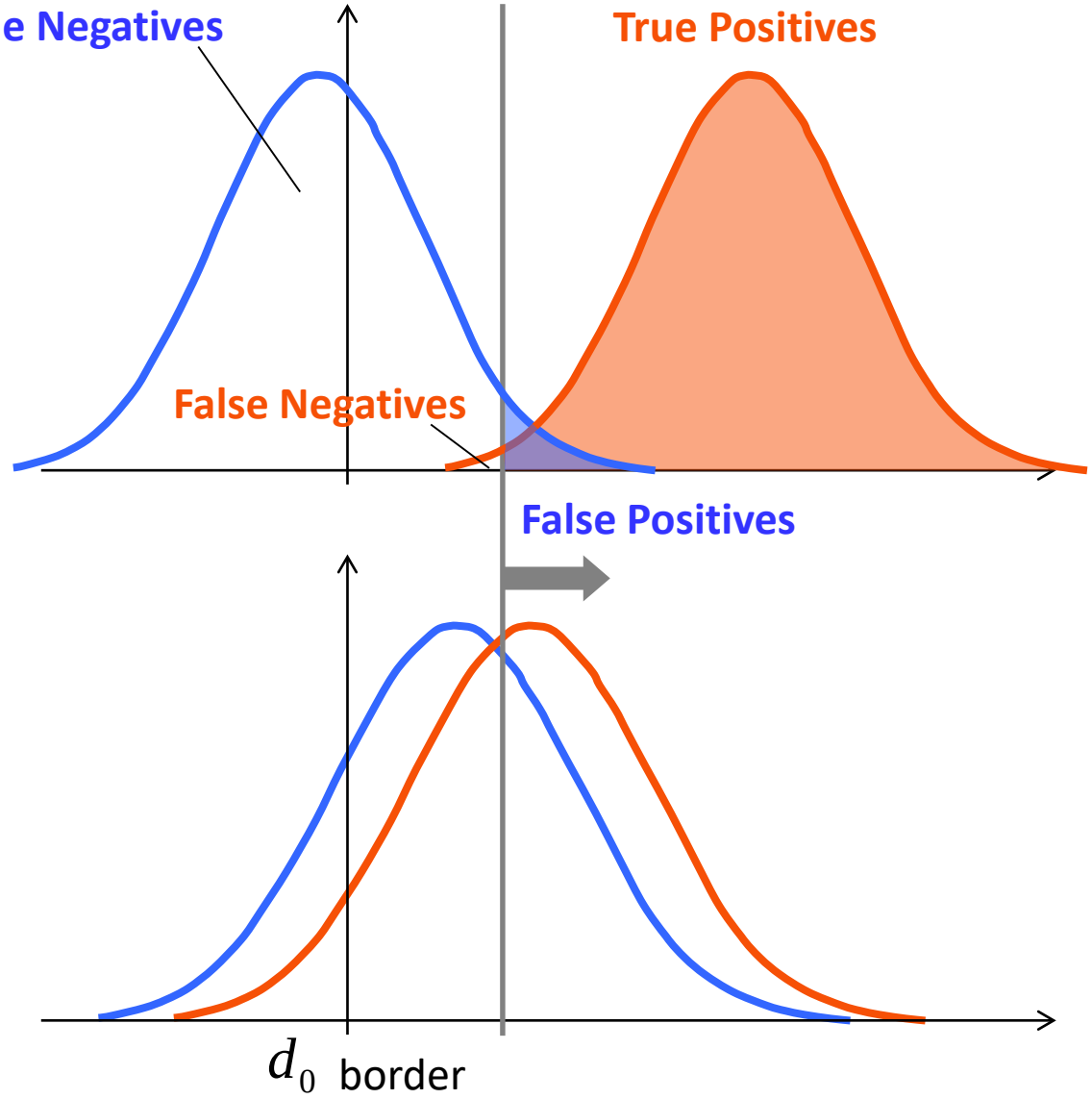
Separation of classes



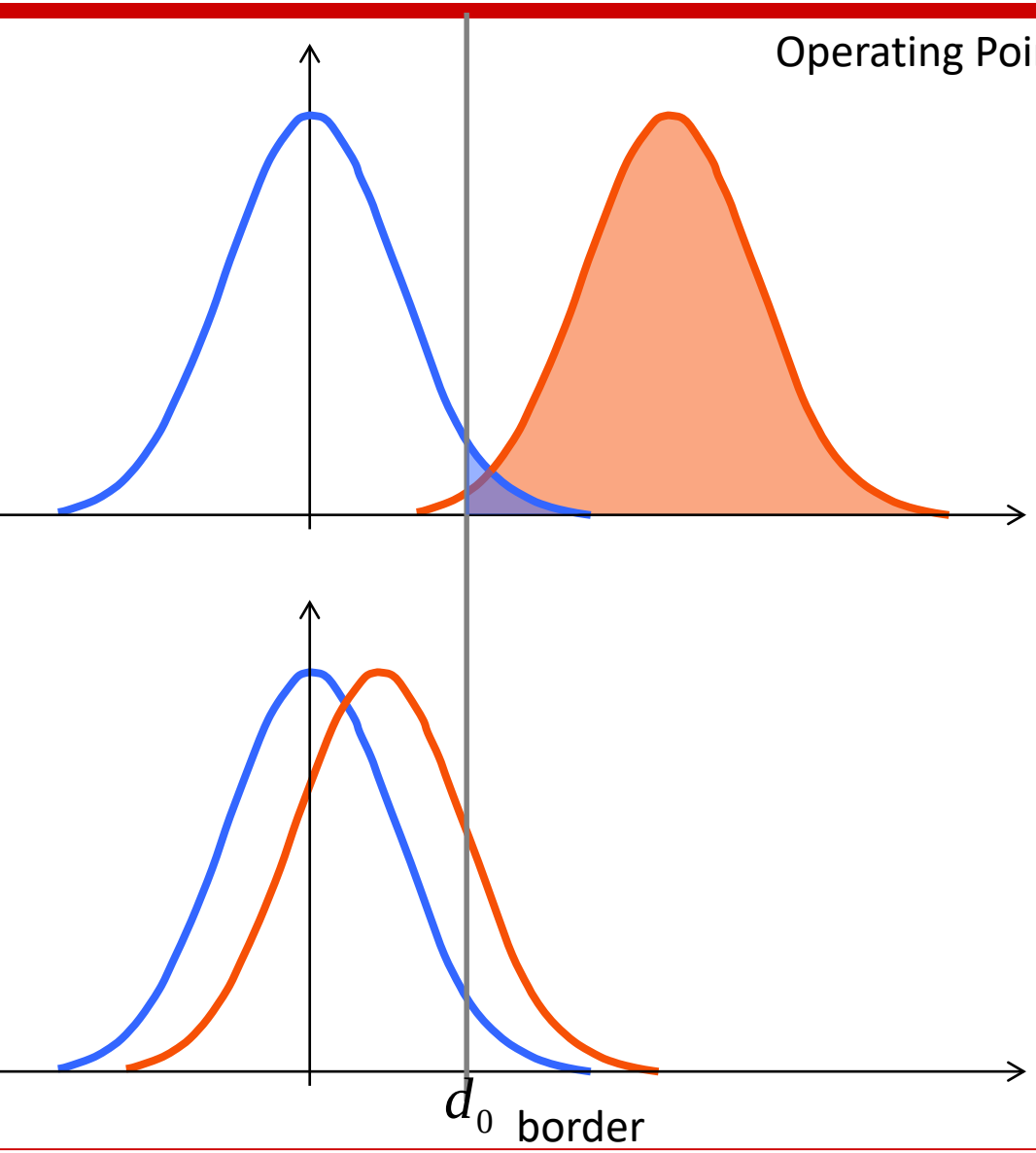
„good“ separation

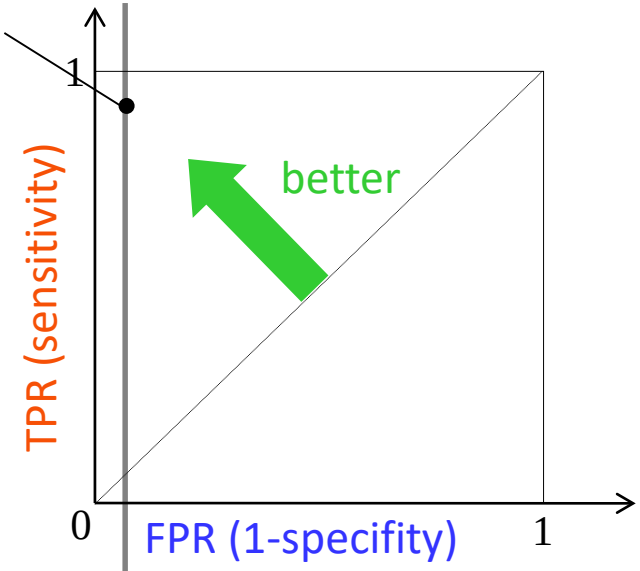


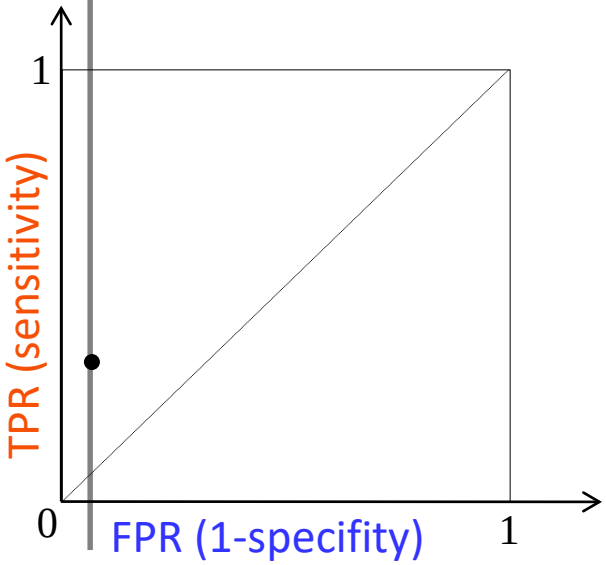
True positives and negatives



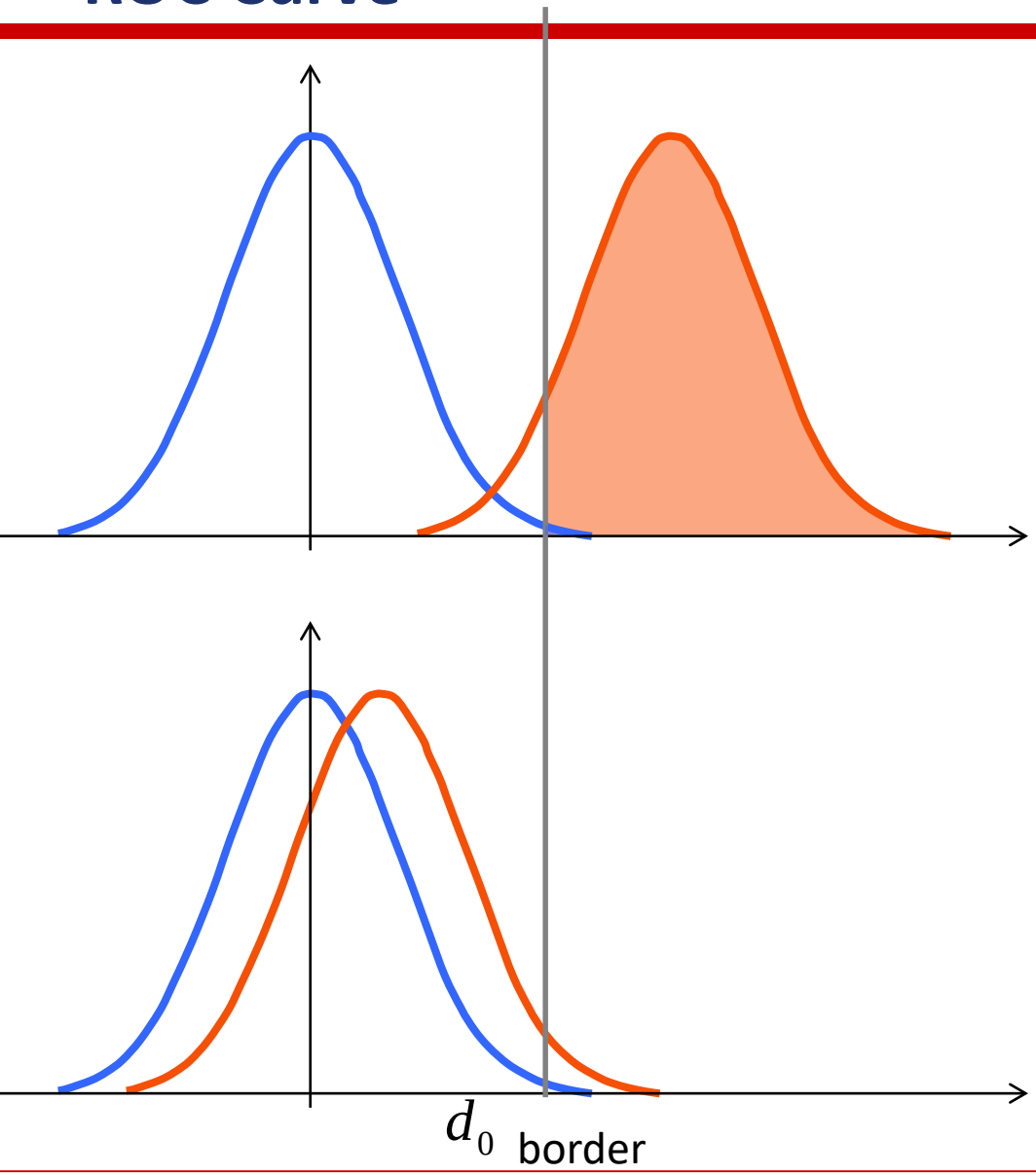
ROC Curve

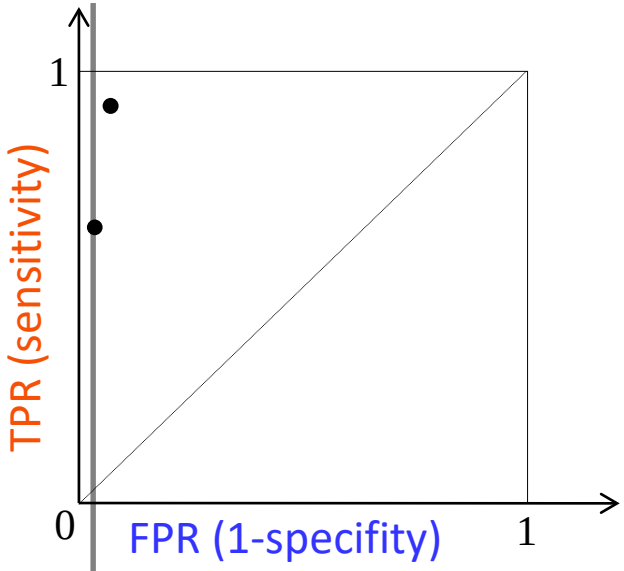

Operating Point
d_0 border

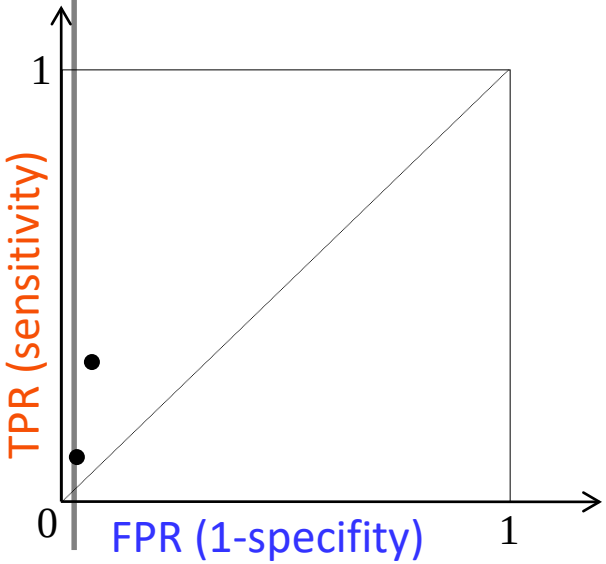

1
better
TPR (sensitivity)
0
FPR (1-specifity)
1


1
TPR (sensitivity)
0
FPR (1-specifity)
1

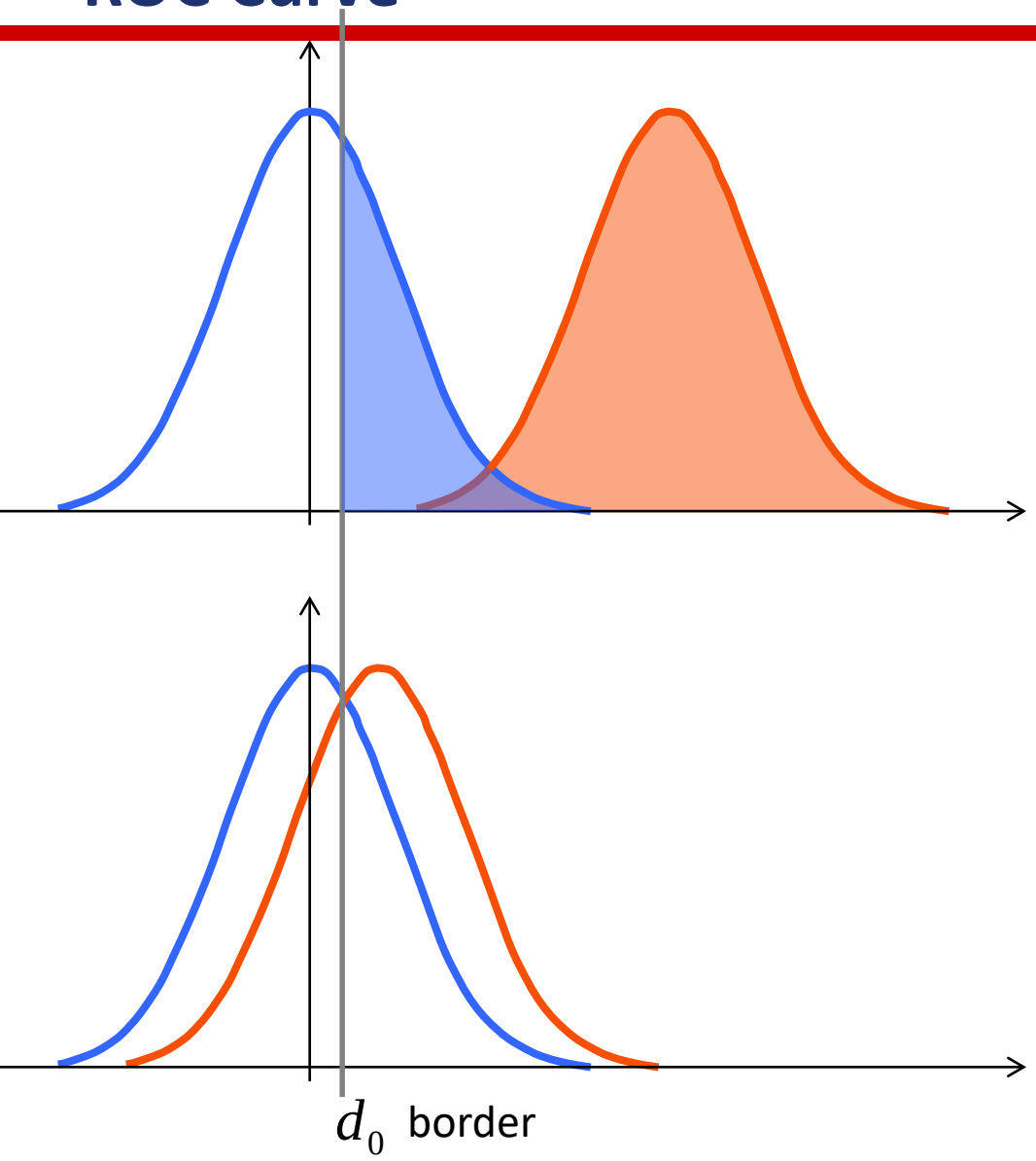
ROC Curve

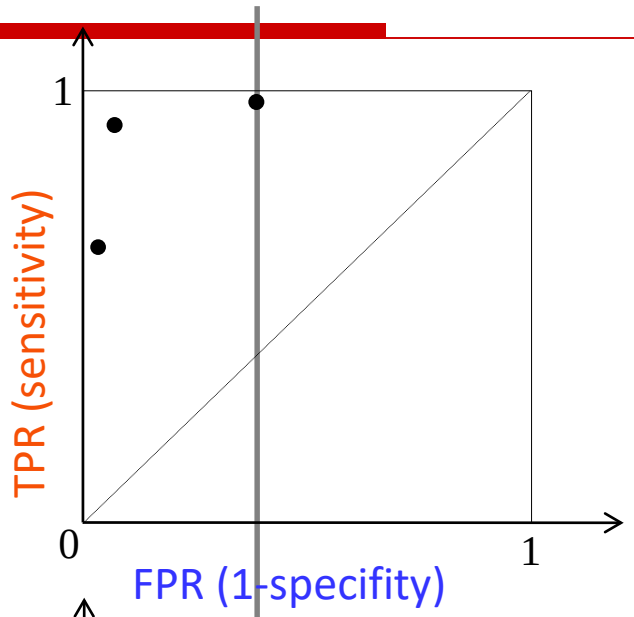

d_0 border

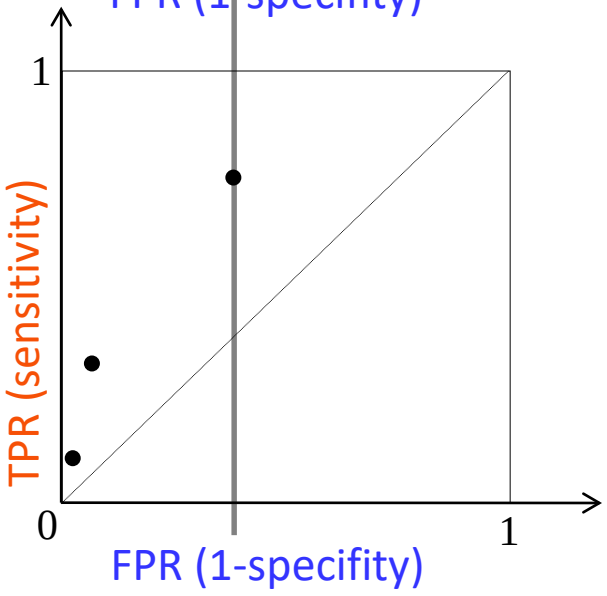

1
TPR (sensitivity)
0
FPR (1-specifity)
1


1
TPR (sensitivity)
0
FPR (1-specifity)
1

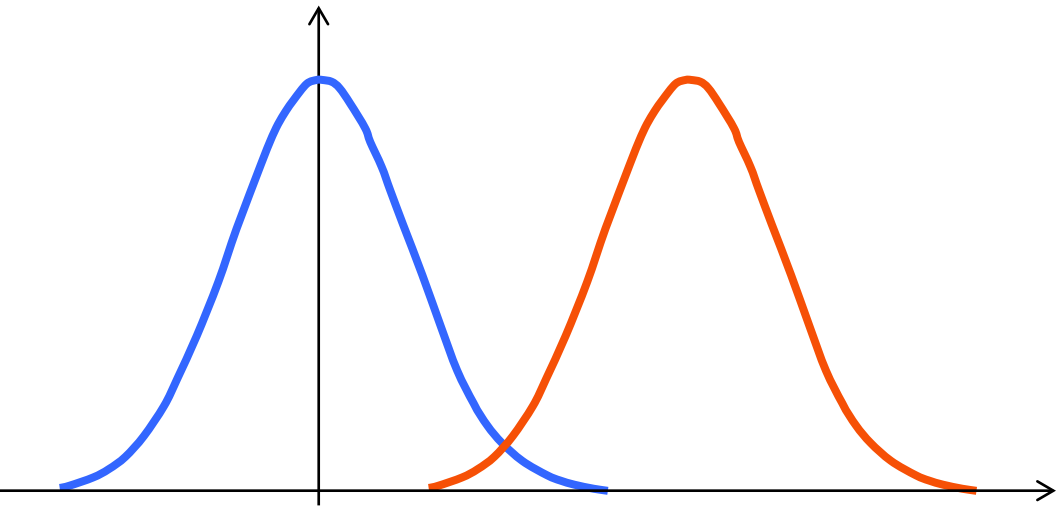
ROC Curve

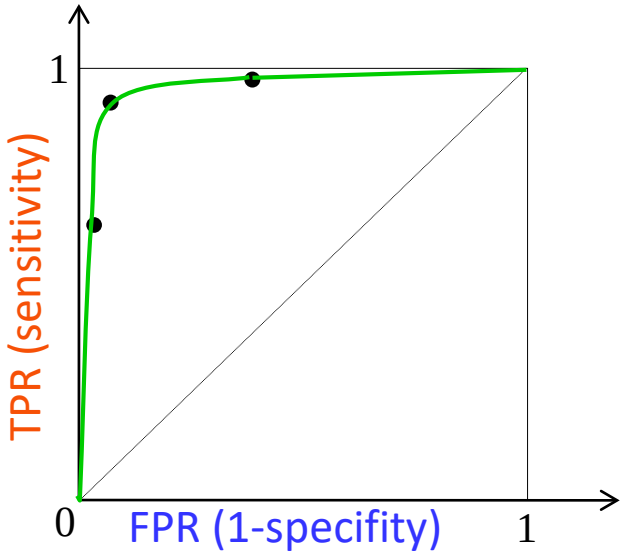

d_0 border

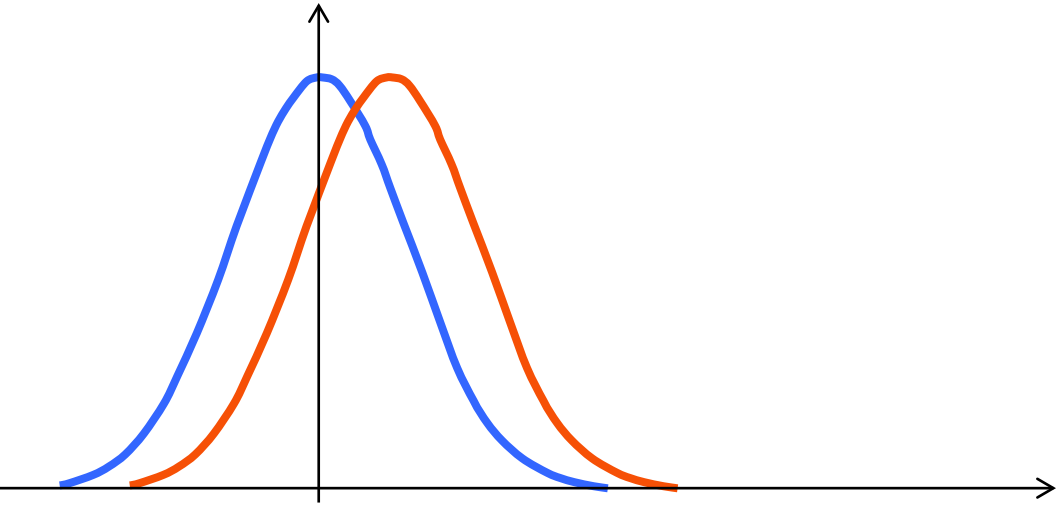

1
TPR (sensitivity)
0
1
FPR (1-specifity)

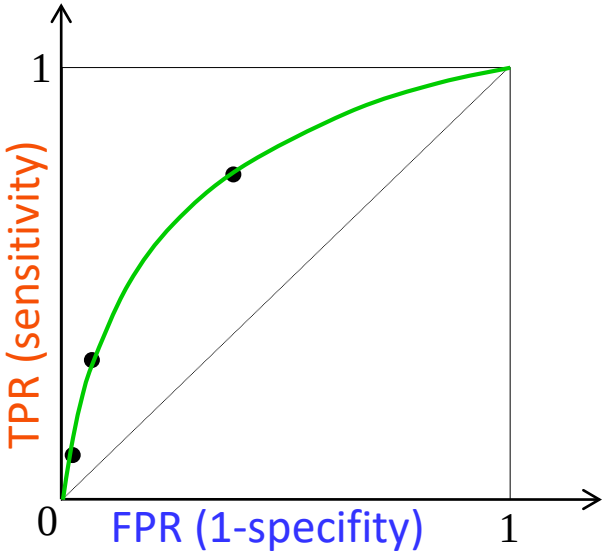

1
TPR (sensitivity)
0
1
FPR (1-specifity)

ROC Curve




1
TPR (sensitivity)
0
FPR (1-specifity)
1

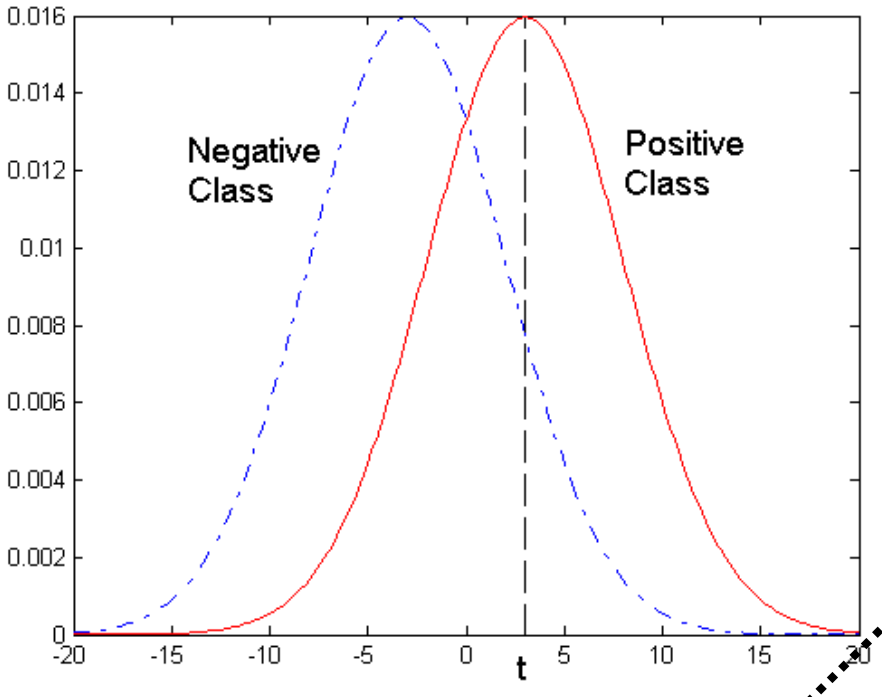


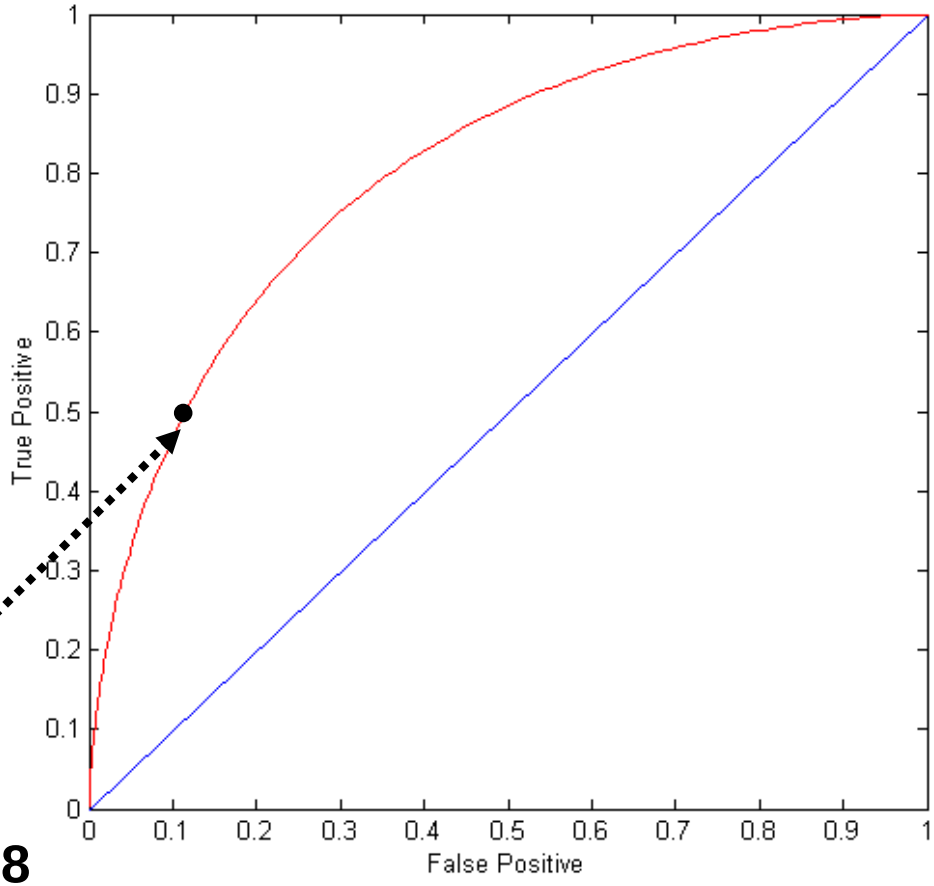

1
TPR (sensitivity)
0
FPR (1-specifity)
1

ROC Curve

1-dimensional data set containing 2 classes (positive and negative)

any points located at x > t is classified as positive

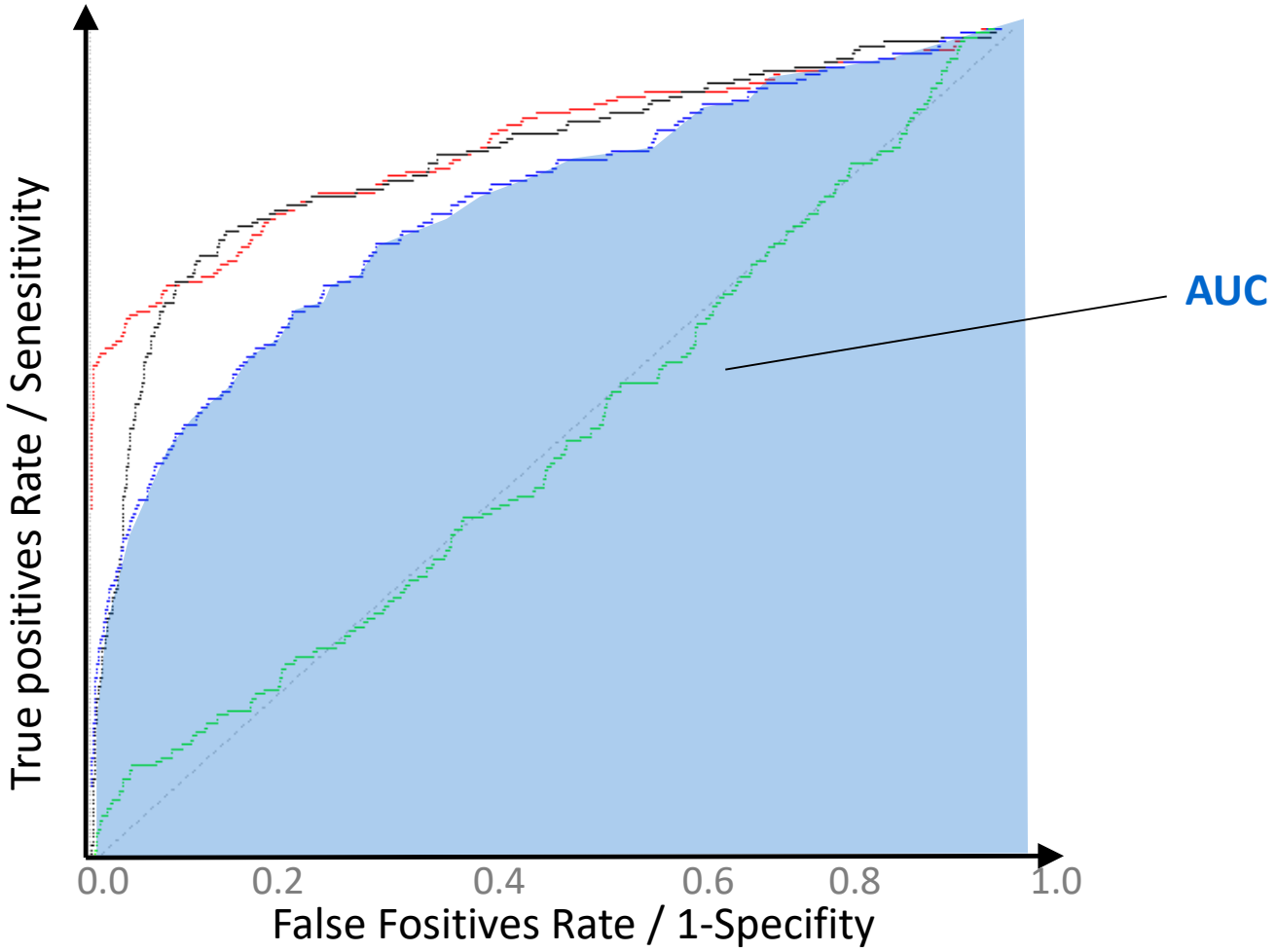




At threshold t:

TP=0.5, FN=0.5, FP=0.12, FN=0.88

ROC Curve



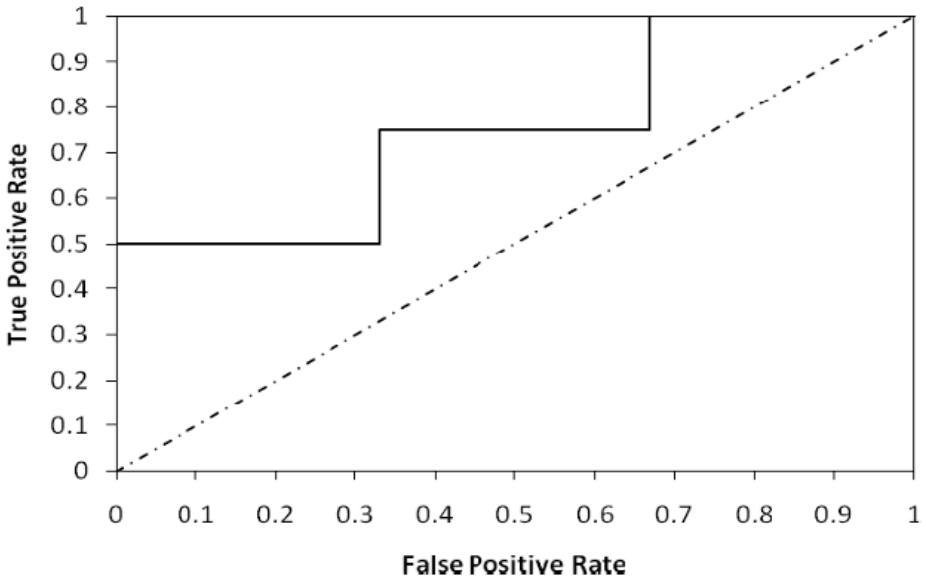
Integral (area under curve = AUC)

Perfect classiciation = 1 , random = 0.5

Example for creating an ROC curve

Table 3.4. Computations for drawing an ROC curve

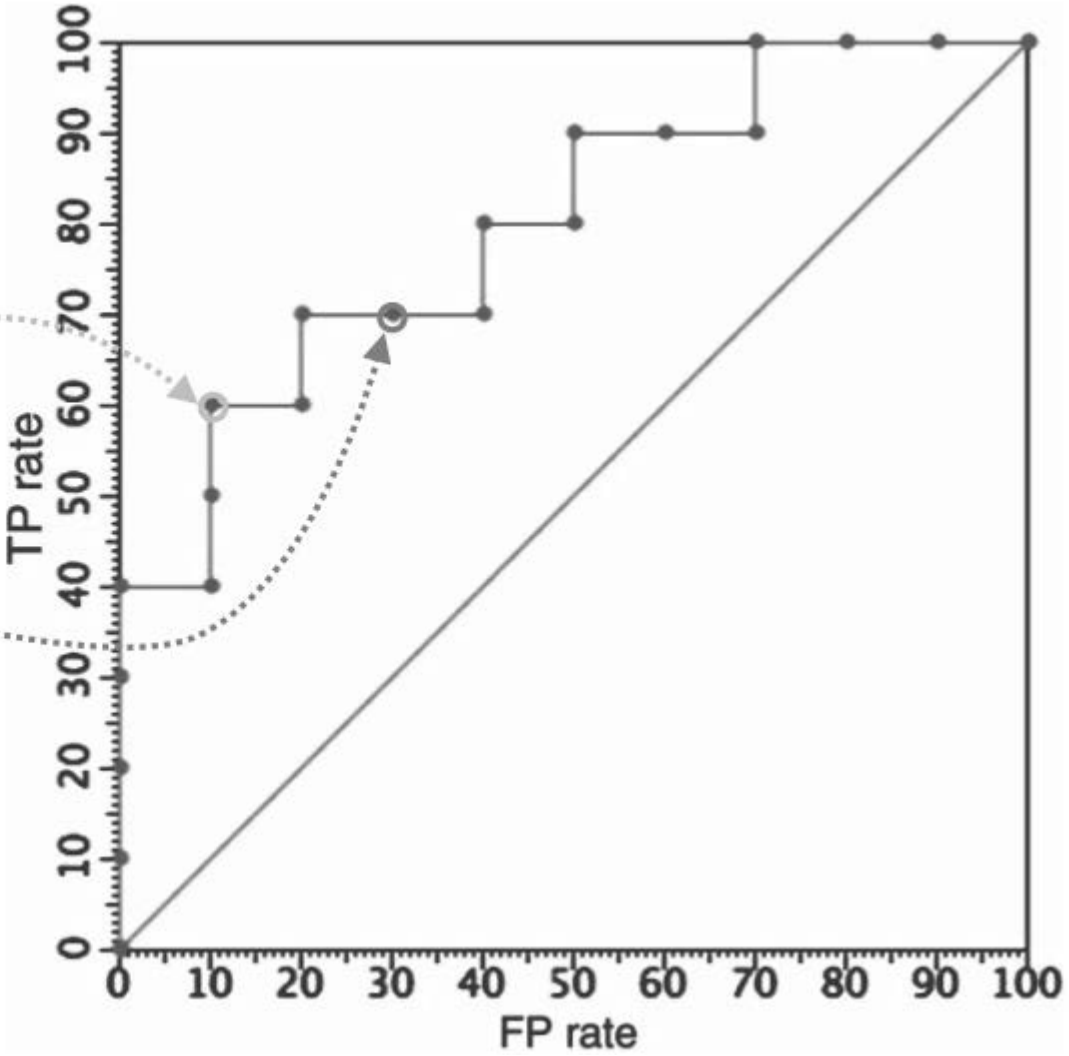
Rank		1	2	3	4	5	6	7	8	9	10
Actual class		+	+	–	–	+	–	–	+	–	–
TP	0	1	2	2	2	3	3	3	4	4	4
FP	0	0	0	1	2	2	3	4	4	5	6
TN	6	6	6	5	4	4	3	2	2	1	0
FN	4	3	2	2	2	1	1	1	0	0	0
TPR	0	0.25	0.5	0.5	0.5	0.75	0.75	0.75	1	1	1
FPR	0	0	0	0.17	0.33	0.33	0.50	0.67	0.67	0.83	1



Liu, B. (2011). *Web data mining: Exploring hyperlinks, contents, and usage data* (2nd ed). *Data-centric systems and applications*. Heidelberg, New York: Springer.

Example for creating an ROC curve

Class	Score
+	0.98
+	0.93
+	0.87
+	0.84
–	0.79
+	0.73
+	0.67
–	0.62
+	0.57
–	0.54
–	0.48
+	0.43
–	0.37
+	0.34
–	0.28
–	0.24
+	0.18
–	0.12
–	0.09
–	0.03

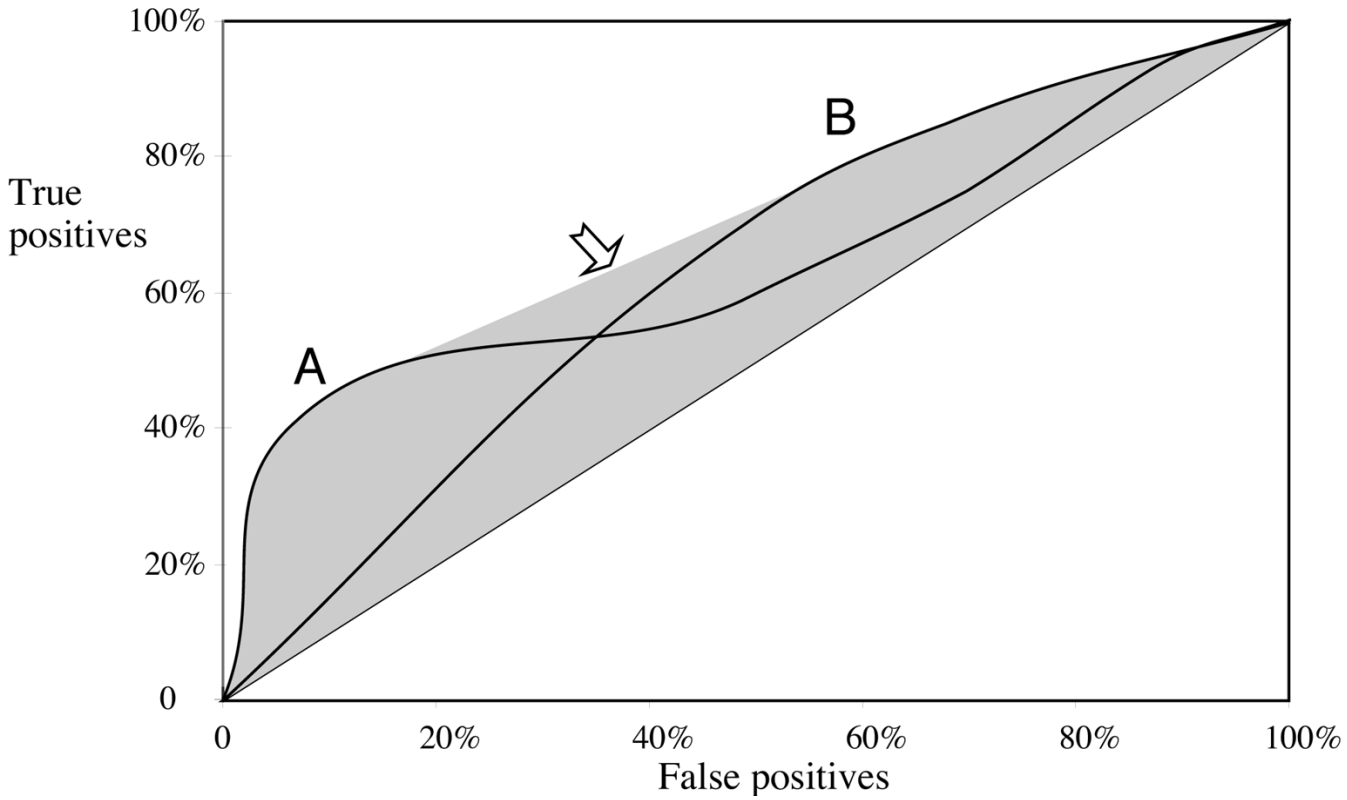


ROC-Kurven für zwei Verfahren

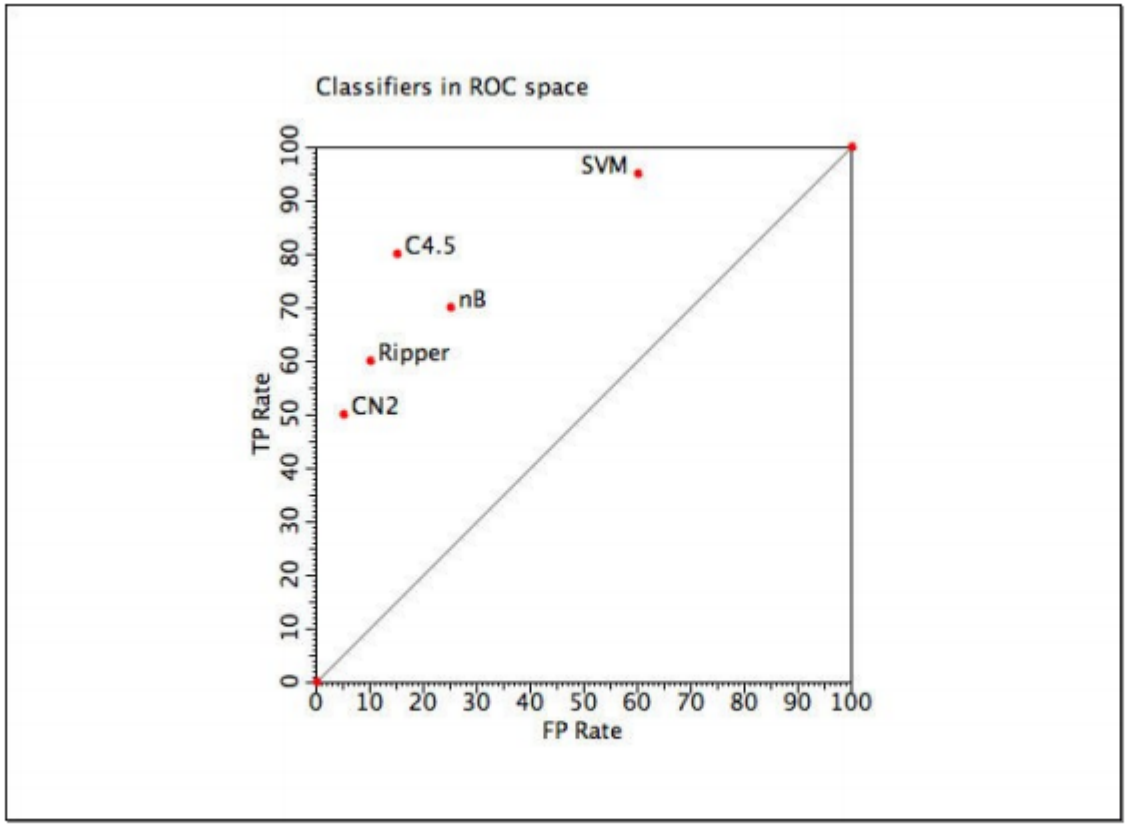
For a small, selected quantity, use method A

For larger quantities, use method B

Between: choose between A and B with appropriate probabilities



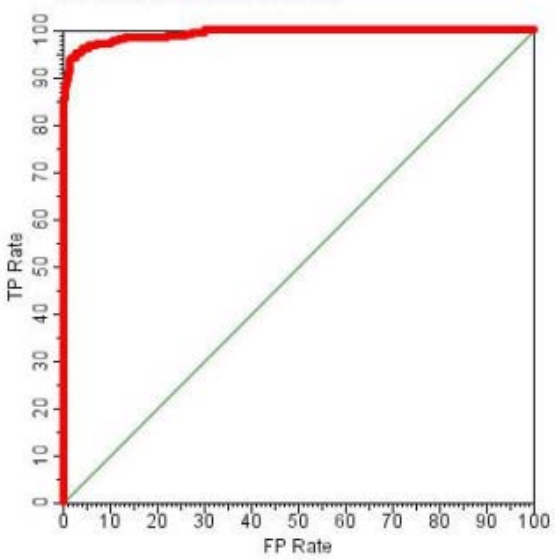
Example

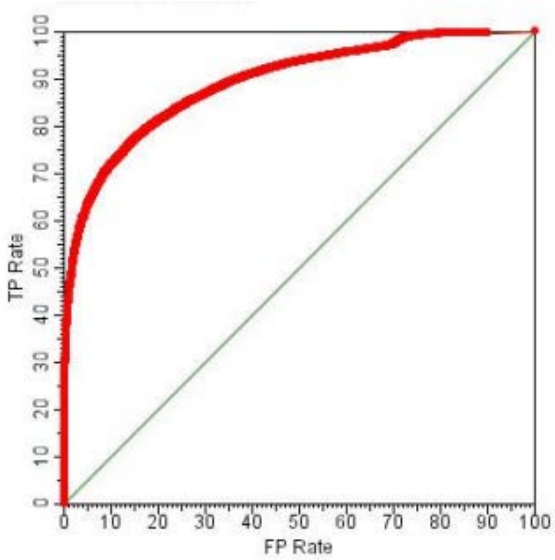


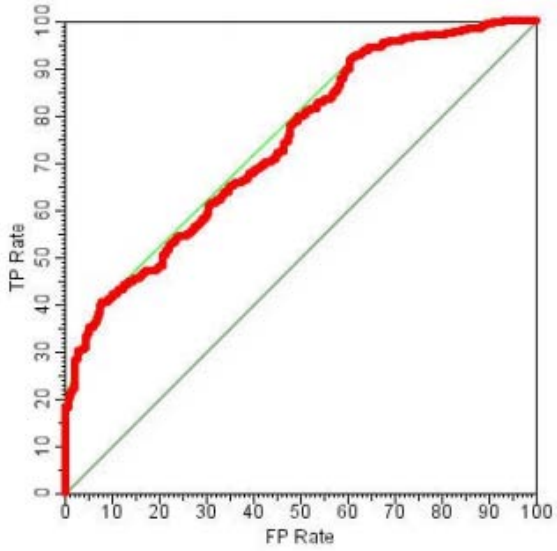
ROC plot produced by ROCon (http://www.cs.bris.ac.uk/Research/MachineLearning/rocon/)

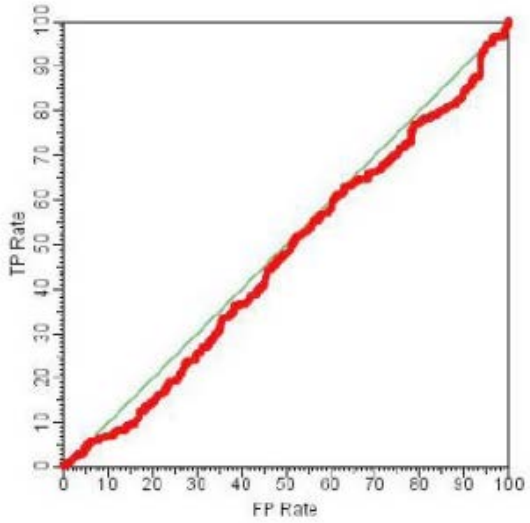
Slide © P. Flach, ICML-04 Tutorial on ROC

Examples

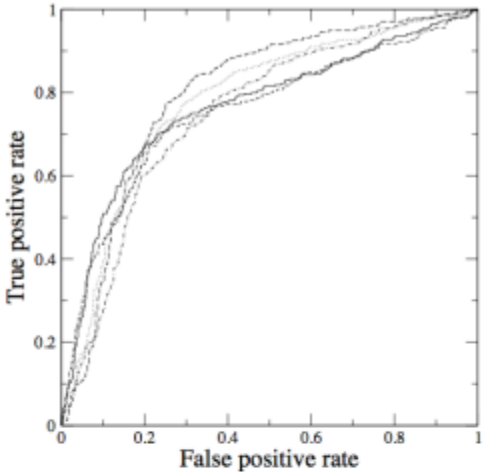

TP Rate
FP Rate
0 10 20 30 40 50 60 70 80 90 100


TP Rate
FP Rate
0 10 20 30 40 50 60 70 80 90 100

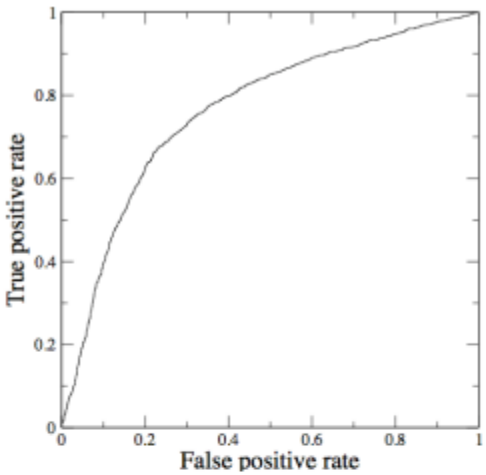

TP Rate
FP Rate
0 10 20 30 40 50 60 70 80 90 100


TP Rate
FP Rate
0 10 20 30 40 50 60 70 80 90 100

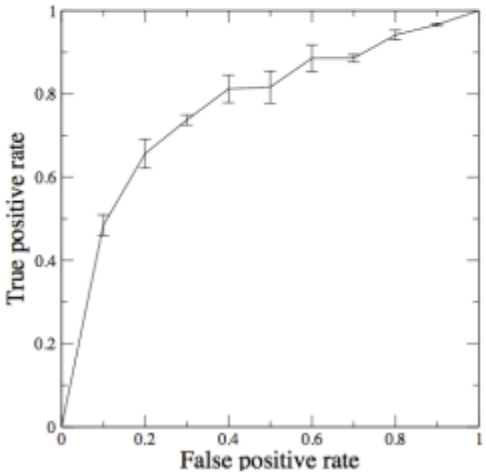
Averaging ROC curves



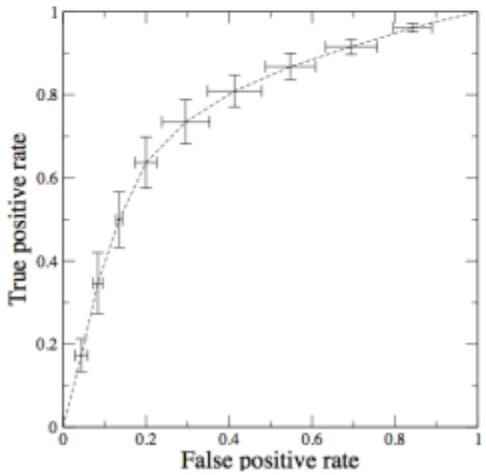
(a) ROC curves from five test samples



(b) ROC curve from combining the samples



(c) Vertical averaging, fixing fpr



(d) Threshold averaging

https://www.cs.bris.ac.uk/~flach/ICML04tutorial/ROCtutorialPartI.pdf

Multi-Class Problem and ROC

ROC works for a binary classification problem.

Often you have multi-class problems.

Two ideas

1. The idea is generally to carry out pairwise comparison (one class vs. all other classes, one class vs. another class).
   Allwein, E. L., Schapire, R. E., & Singer, Y. (2001). Reducing multiclass to binary: a unifying approach for margin classifiers. J. Mach. Learn. Res., 1, 113–141.

   Landgrebe, T. C. W., & Duin, R. P. W. (2007). Approximating the multiclass ROC by pairwise analysis. Pattern Recogn. Lett., 28(13), 1747–1758. doi:10.1016/j.patrec.2007.05.001
2. Compute all precision and recall of all the classes, then average them to get a single real number measurement.

Outline

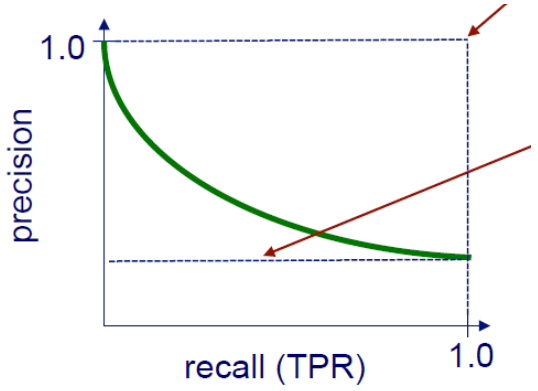
- Methods for Performance Evaluation
- Metrics for Performance Evaluation
- Method for Model Comparison
  Receiver Operating Characteristic
- Summary

Summary ROC

- insensitive to changes in class distribution (ROC curve does not change if the proportion of positive and negative instances in the test set are varied)
- can identify optimal classification thresholds for tasks with differential misclassification costs

Alternative precision/recall curves

Plots the precision vs. recall (TP-rate) as a threshold on the confidence of an instance being positive is varied



- show the fraction of predictions that are false positives
- well suited for tasks with lots of negative instances

Summary

- Use Cross-Validation
- Confusion Matrix

Predicted Class \ True Class	Yes	No
Yes	TP = True Positive	FP = False Positive
No	FN = False Negative	TN = True Negative

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

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

$$\text{Error rate} = \frac{FP + FN}{P + N}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

- *Residual Sum of Squares* $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

		Predicted condition			
	Total population	Predicted Condition positive	Predicted Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	
True condition	condition positive	**True positive**	**False Negative** (Type II error)	True positive rate (TPR), Sensitivity, Recall, probability of detection $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$
	condition negative	**False Positive** (Type I error)	**True negative**	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	True negative rate (TNR), Specificity (SPC) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$
	Accuracy (ACC) $= \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Test outcome positive}}$	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Test outcome negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$
		False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Test outcome positive}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Test outcome negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$	

Source: Wikipedia