Evaluation

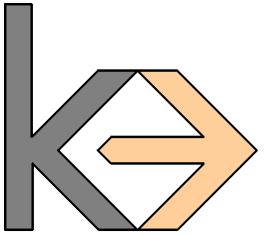
TECHNISCHE UNIVERSITÄT DARMSTADT

Data Mining and Machine Learning: Techniques and Algorithms

Eneldo Loza Mencía

eneldo@ke.tu-darmstadt.de

Knowledge Engineering Group, TU Darmstadt



International Week 2019, 21.1. – 24.1.
University of Economics, Prague

Evaluation of Learned Models

TECHNISCHE UNIVERSITÄT DARMSTADT

- Validation through experts
  - a domain expert evaluates the plausibility of a learned model
    - \+ but often the only option (e.g., clustering)
    - – subjective, time-intensive, costly
- Validation on data
  - evaluate the accuracy of the model on a separate dataset drawn from the same distribution as the training data
    - – labeled data are scarce, could be better used for training
    - \+ fast and simple, off-line, no domain knowledge needed, methods for re-using training data exist (e.g., cross-validation)
- On-line Validation
  - test the learned model in a fielded application
    - \+ gives the best estimate for the overall utility
    - – bad models may be costly

Confusion Matrix

TECHNISCHE UNIVERSITÄT DARMSTADT

	Classified as +	Classified as –	
Is +	true positives (tp)	false negatives (fn)	tp + fn = P
Is –	false positives (fp)	true negatives (tn)	fp + tn = N
	tp + fp	fn + tn	|E| =P + N

- the confusion matrix summarizes all important information
  - how often is class i confused with class j
- most evaluation measures can be computed from the confusion matrix, most prominent ones being:
  - accuracy:
    - percentage of correctly classified examples

$$acc = \frac{tp + tn}{P + N}$$

  - error rate:
    - percentage of incorrectly classified examples

$$err = \frac{fp + fn}{P + N} = 1 - acc$$

Recall and Precision

TECHNISCHE UNIVERSITÄT DARMSTADT

- Accuracy is sometimes not good for evaluation
  - Accuracy must be interpreted relative to *default accuracy* (accuracy of the learner that always predicts majority class)
    - For unbalanced class distributions might be misleading
  - No interpretation of results (low accuracy because classified too cautious/permissive?)
- Alternative:
  - **Recall**: Percentage of *relevant* (=positive) *and predicted* test instances among all *relevant* test instances
  - **Precision**: Percentage of *relevant and predicted* test instances among all *predicted* test instances

$$R = \frac{tp}{tp + fn}$$

$$P = \frac{tp}{tp + fp}$$

	Classified as +	Classified as –
Is +	tp	fn
Is –	fp	tn

F-Measure

- Weighted harmonic mean of recall and precision

$$F_\beta = \frac{(\beta^2+1)\cdot P \cdot R}{\beta^2 \cdot P + R}$$

  - equivalent form for $\alpha = \frac{\beta^2}{\beta^2+1}$ $\quad F_\alpha = \frac{1}{\alpha\frac{1}{R} + (1-\alpha)\frac{1}{P}}$

- The parameter β can be used to trade off the relative importance of recall and precision
  - $F_0 = P$
  - $F_\infty = R$
  - F_1: P and R equally weighted
  - F_2: recall is four times more important than precision
  - $F_{0.5}$: precision is four times more important than recall

Recall and Precision for Multi-Class Problems

TECHNISCHE UNIVERSITÄT DARMSTADT

- For multi-class text classification tasks, recall and precision can be defined for each category separately
- Recall of Class X:
  - How many documents of class X have been recognized as class X?
- Precision of Class X:
  - How many of our predictions for class X were correct?
- Predictions for Class X can be summarized in a 2x2 table
  - z.B:

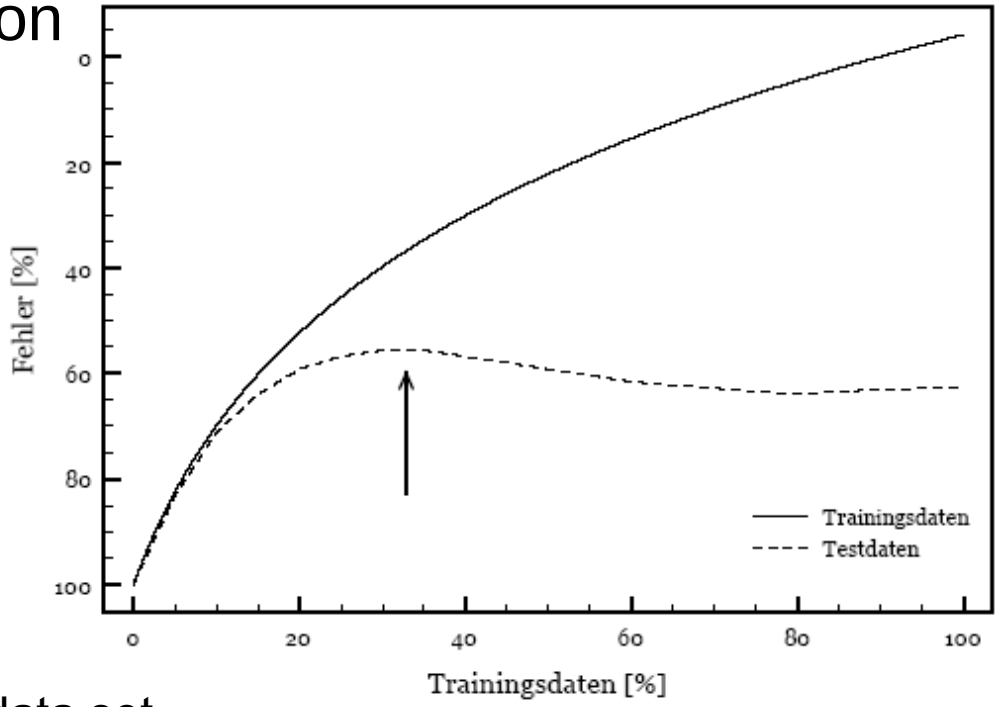
$$X = A, \overline{X} = \{B, C, D\}$$

	classified X	~~classified not X~~	
is X	$n_{X,X}$	$n_{\overline{X},X}$	n_X
~~is not X~~	$n_{X,\overline{X}}$	$n_{\overline{X},\overline{X}}$	$n_{\overline{X}}$
	$\overline{n}_X$	$\overline{n}_{\overline{X}}$	n

Out-of-Sample Testing

TECHNISCHE UNIVERSITÄT DARMSTADT

- Performance cannot be measured on training data
  - over-estimation due to (over)fitting!
- Reserve a portion of the available data for testing
  - typical scenario
    - 2/3 of data for training
    - 1/3 of data for testing (evaluation)
  - a classifier is trained on the training data
  - and tested on the test data
    - e.g., confusion matrix is computed for test data set
- Problems:
  - waste of data
  - labelling may be expensive
  - high variance
    - often: repeat 10 times or → cross-validation

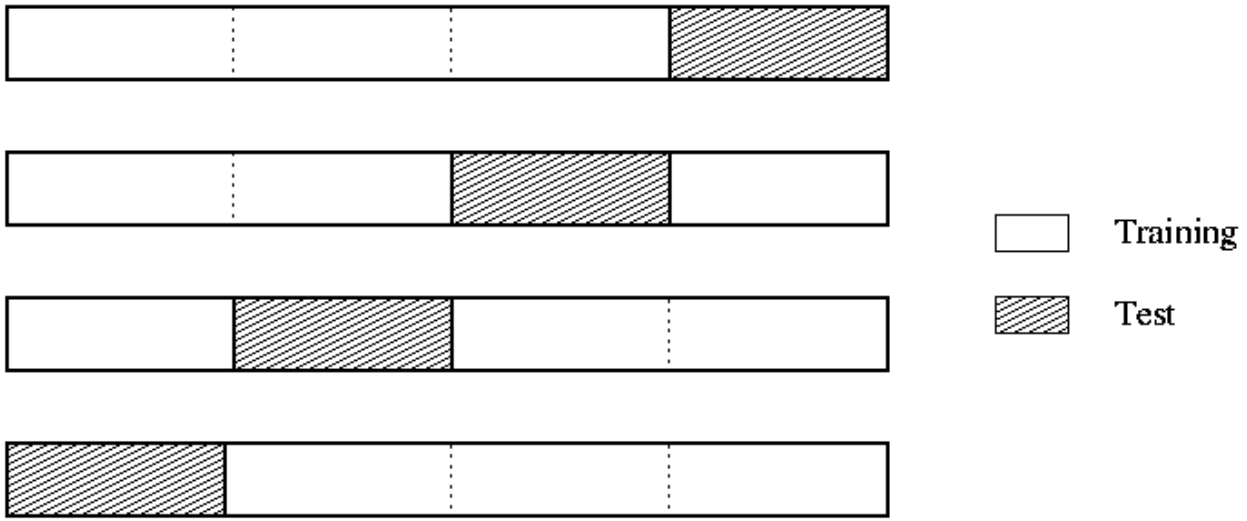


Winkler 2007, according to Mitchell 1997

Cross-Validation

TECHNISCHE UNIVERSITÄT DARMSTADT

- Algorithm:
  - split dataset into x (usually 10) partitions
  - for every partition X
    - use other x-1 partitions for learning and partition X for testing
  - average the results
- Example: 4-fold cross-validation



Leave-One-Out Cross-Validation

TECHNISCHE UNIVERSITÄT DARMSTADT

- n-fold cross-validation
  - where n is the number of examples:
    - use n-1 examples for training
    - 1 example for testing
    - repeat for each example
- Properties:
  - \+ makes best use of data
    - only one example not used for testing
  - \+ no influence of random sampling
    - training/test splits are determined deterministically
  - – typically very expensive
    - but, e.g., not for k-NN (Why?)
  - – bias
    - Why?

Cost-Sensitive Evaluation

TECHNISCHE UNIVERSITÄT DARMSTADT

- Error rate assumes same costs for all misclassifications, but this is very often not the case in reality

  Examples
  - Loan Applications
    - rejecting an applicant who will not pay back → minimal costs
    - accepting an applicant who will pay back → gain
    - accepting an applicant who will not pay back → big loss
    - rejecting an applicant who would pay back → loss
  - Spam-Mail Filtering
    - rejecting good E-mails (ham) is much worse than accepting a few spam mails
  - Medical Diagnosis
    - failing to recognize a disease is often much worse than to treat a healthy patient for this disease

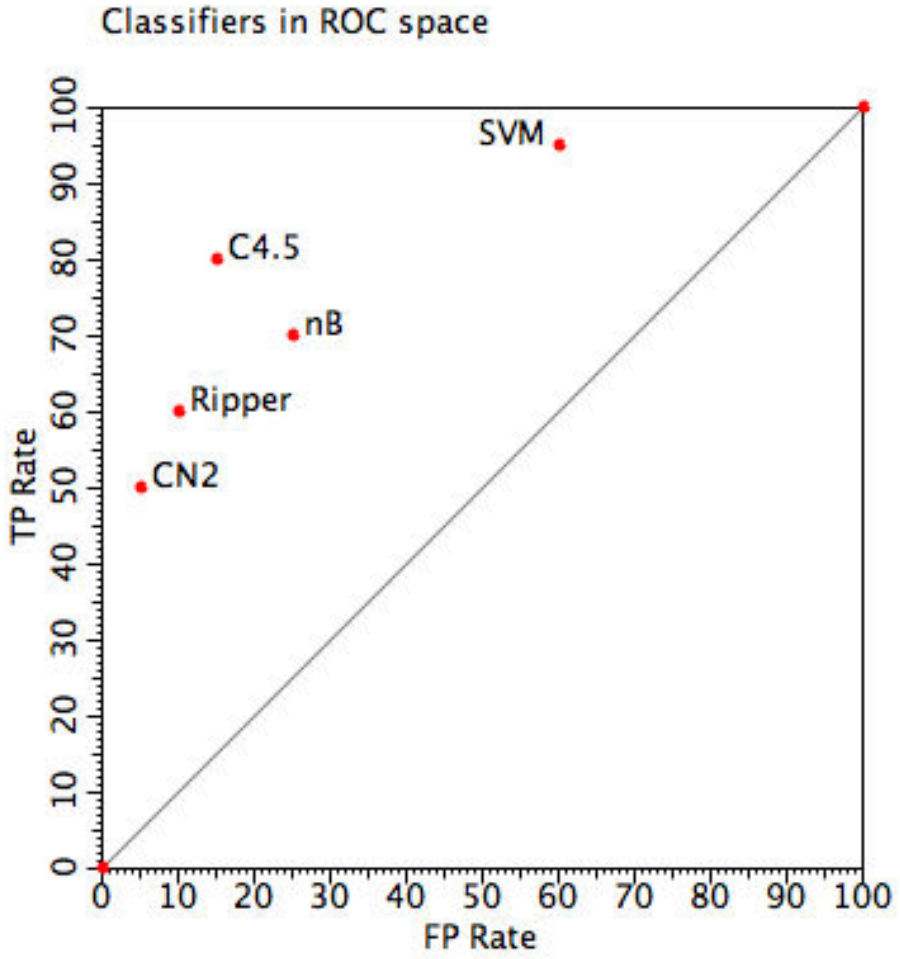
ROC Analysis

TECHNISCHE UNIVERSITÄT DARMSTADT

- Receiver Operating Characteristic
  - origins in signal theory to show tradeoff between hit rate and false alarm rate over noisy channel
- Basic Objective:
  - Determine the best classifier for varying cost models
    - accuracy is only one possibility, where true positives and false positives receive equal weight
- Additional Objective: analyze single classifier
- Method:
  - Visualization in ROC space
    - each classifier is characterized by its measures
      - false positive rate fpr (y-axis)
      - true positive rate tpr (x-axis)

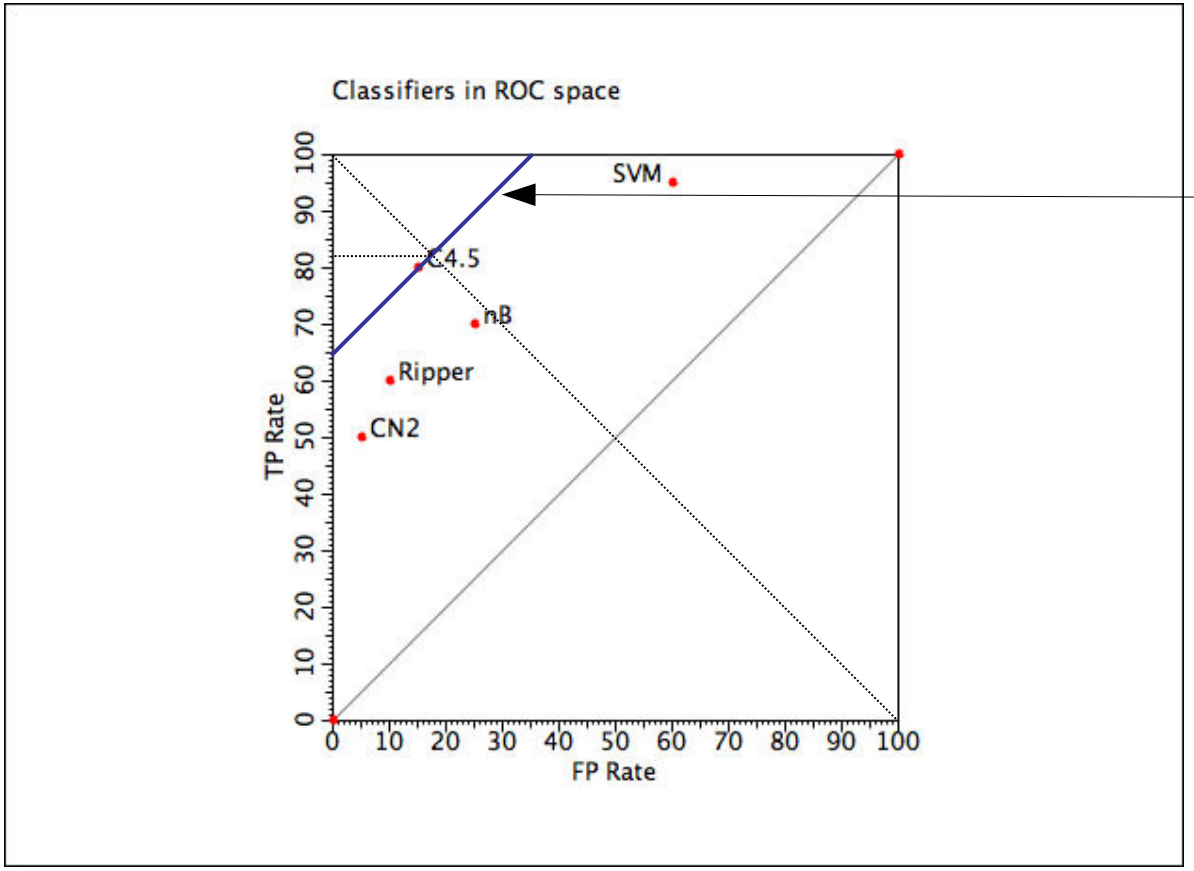
$$fpr = \frac{fp}{fp + tn} \qquad tpr = \frac{tp}{tp + fn}$$

ROC Analysis
Example plot



ROC plot produced by ROCon (http://www.cs.bris.ac.uk/Research/MachineLearning/rocon/) Slide © P. Flach, ICML-04 Tutorial on ROC

ROC Analysis
Selecting the optimal classifier



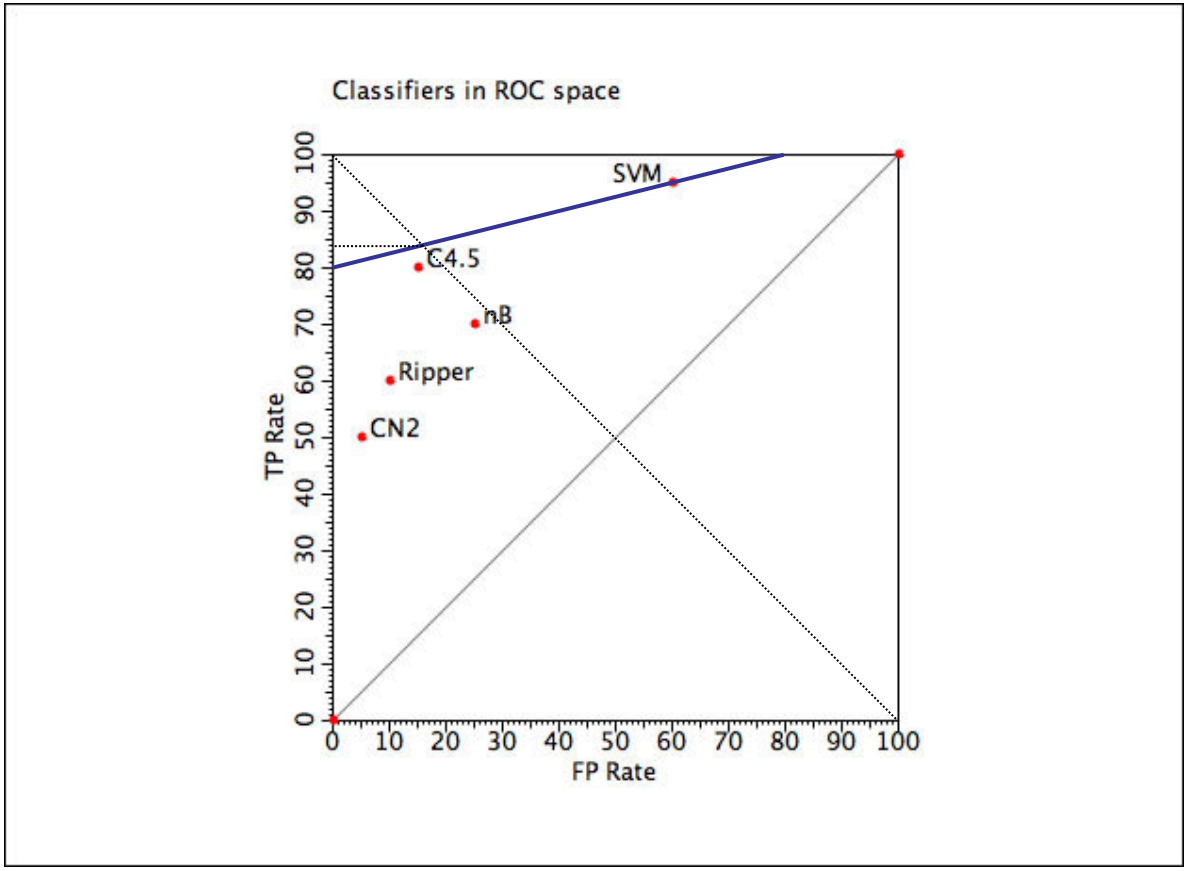
Iso-cost lines connects ROC points with the same costs

For costs “false positive is as bad as a false negative”

→ C4.5 is optimal

Slide adapted from P. Flach, ICML-04 Tutorial on ROC

ROC Analysis
Selecting the optimal classifier

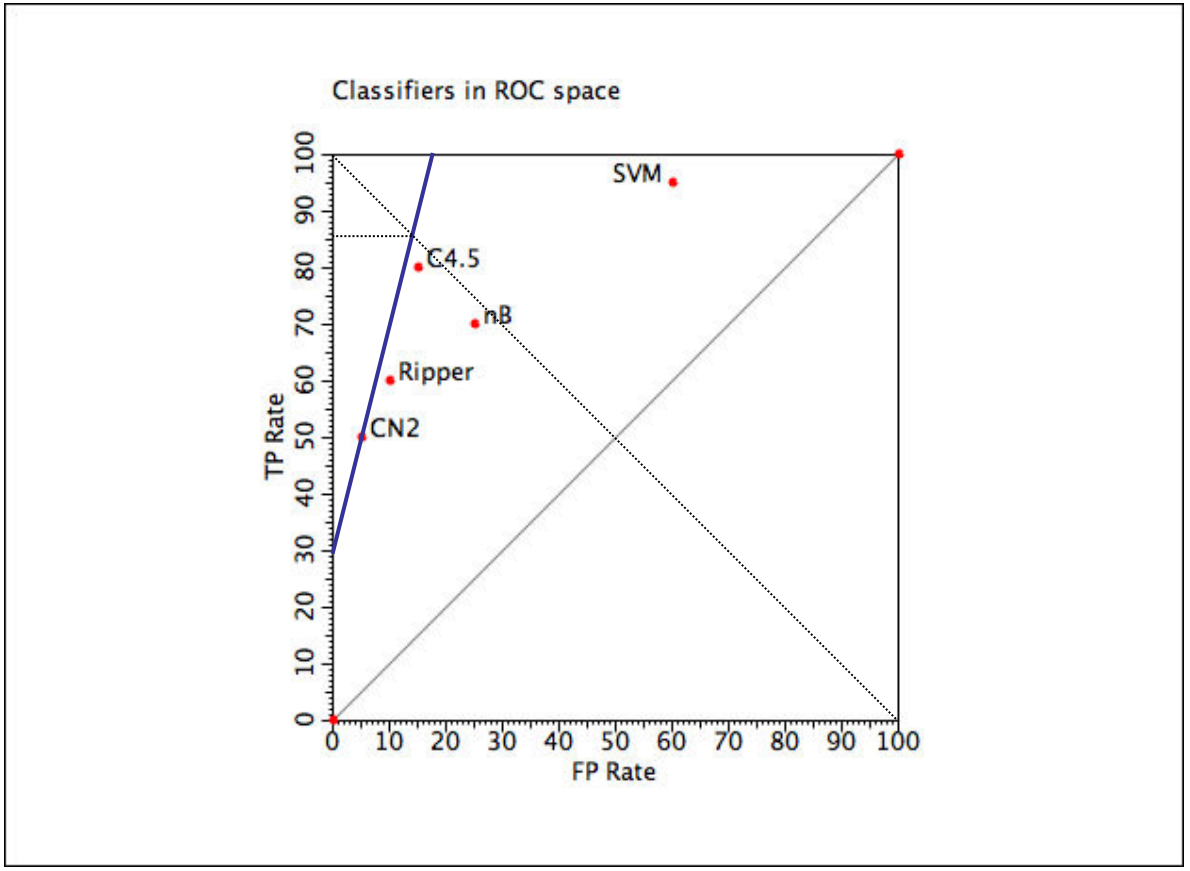


For costs “false positive is for times less worse than a false negative”

→ SVM is optimal

Slide adapted from P. Flach, ICML-04 Tutorial on ROC

ROC Analysis
Selecting the optimal classifier



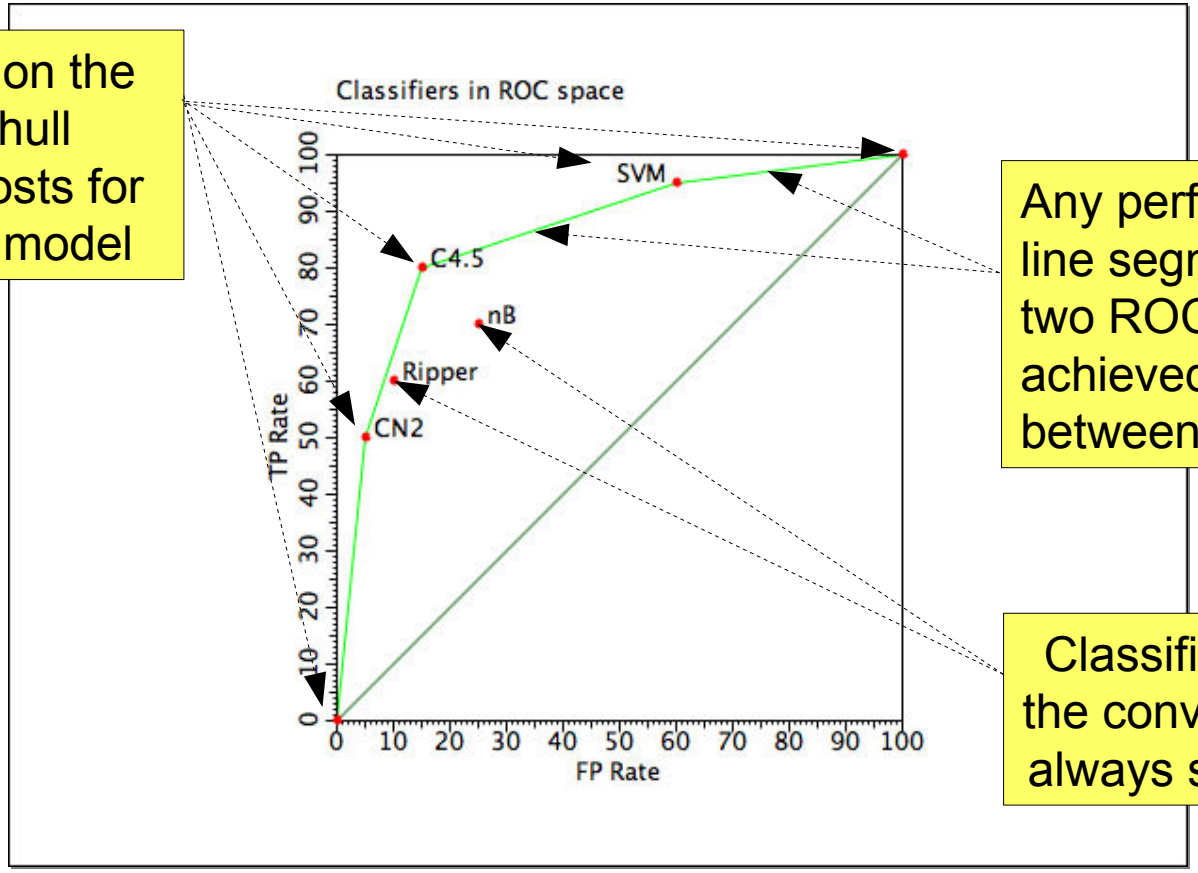
For costs “false positive is for times worse than a false negative”
→ CN2 is optimal

Slide adapted from P. Flach, ICML-04 Tutorial on ROC

ROC Analysis
The ROC convex hull

TECHNISCHE UNIVERSITÄT DARMSTADT

Classifiers on the convex hull minimize costs for some cost model

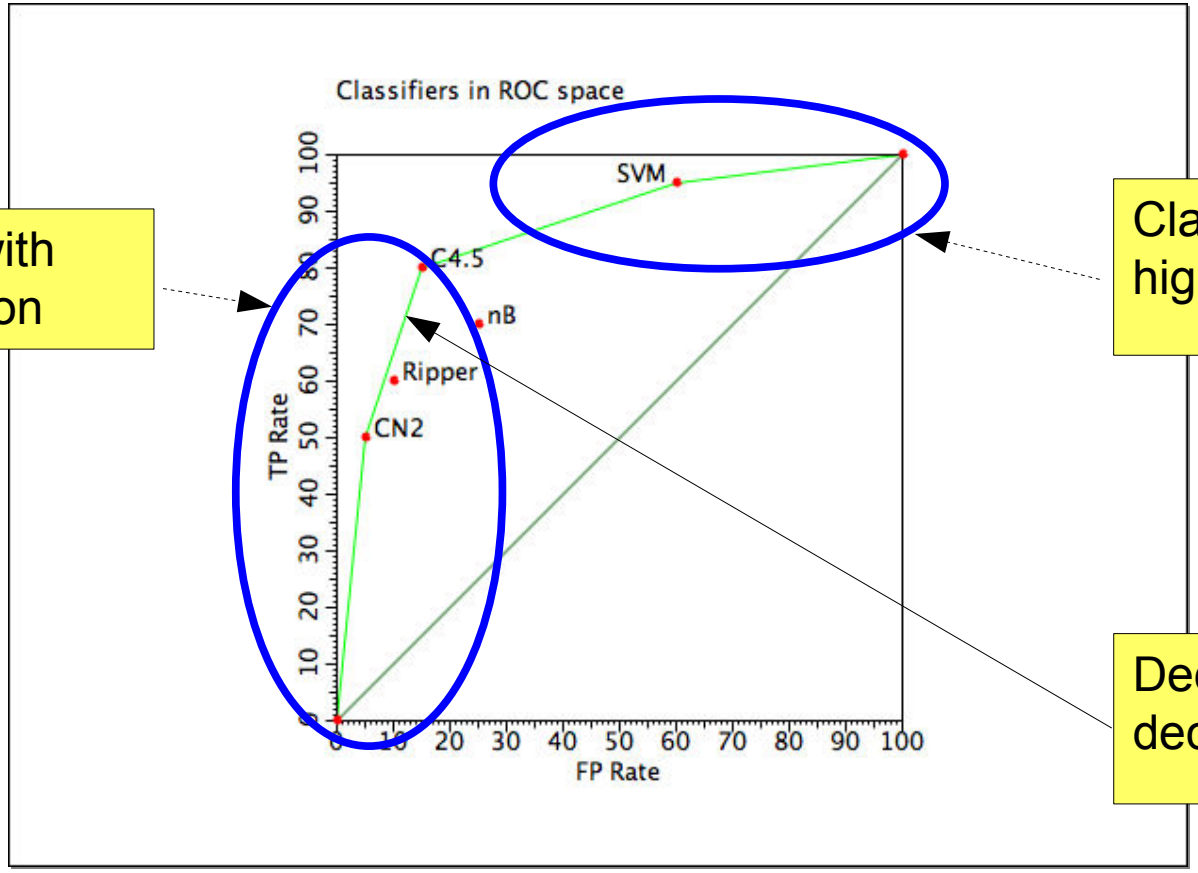


Any performance on a line segment connecting two ROC points can be achieved by interpolating between the classifiers

Classifiers below the convex hull are always suboptimal

Slide adapted from P. Flach, ICML-04 Tutorial on ROC

ROC Analysis
The ROC convex hull



Classifiers with high precision

Classifiers with high recall

Decreasing slope → decreasing precision

Slide adapted from P. Flach, ICML-04 Tutorial on ROC

Rankers and Classifiers

TECHNISCHE UNIVERSITÄT DARMSTADT

- A scoring classifier outputs scores $f(x,+)$ and $f(x,-)$ for each class
  - e.g. estimate probabilities $P(+ \mid x)$ and $P(- \mid x)$
  - scores don't need to be normalised
- $f(x) = f(x,+) / f(x,-)$ can be used to rank instances from most to least likely positive
  - e.g. odds ratio $P(+ \mid x) / P(- \mid x)$
- Rankers can be turned into classifiers by setting a threshold on $f(x)$
- Example:
  - Naïve Bayes Classifier for two classes is actually a ranker
  - that has been turned into classifier by setting a probability threshold of 0.5 (corresponds to a odds ratio treshold of 1.0)
    - $P(+ \mid x) > 0.5 > 1 - P(+ \mid x) = P(- \mid x)$ means that class $+$ is more likely

A simple classifier that estimates probabilities $P(c|x)$ for each class c

Slide adapted from P. Flach, ICML-04 Tutorial on ROC

Drawing ROC Curves for Rankers

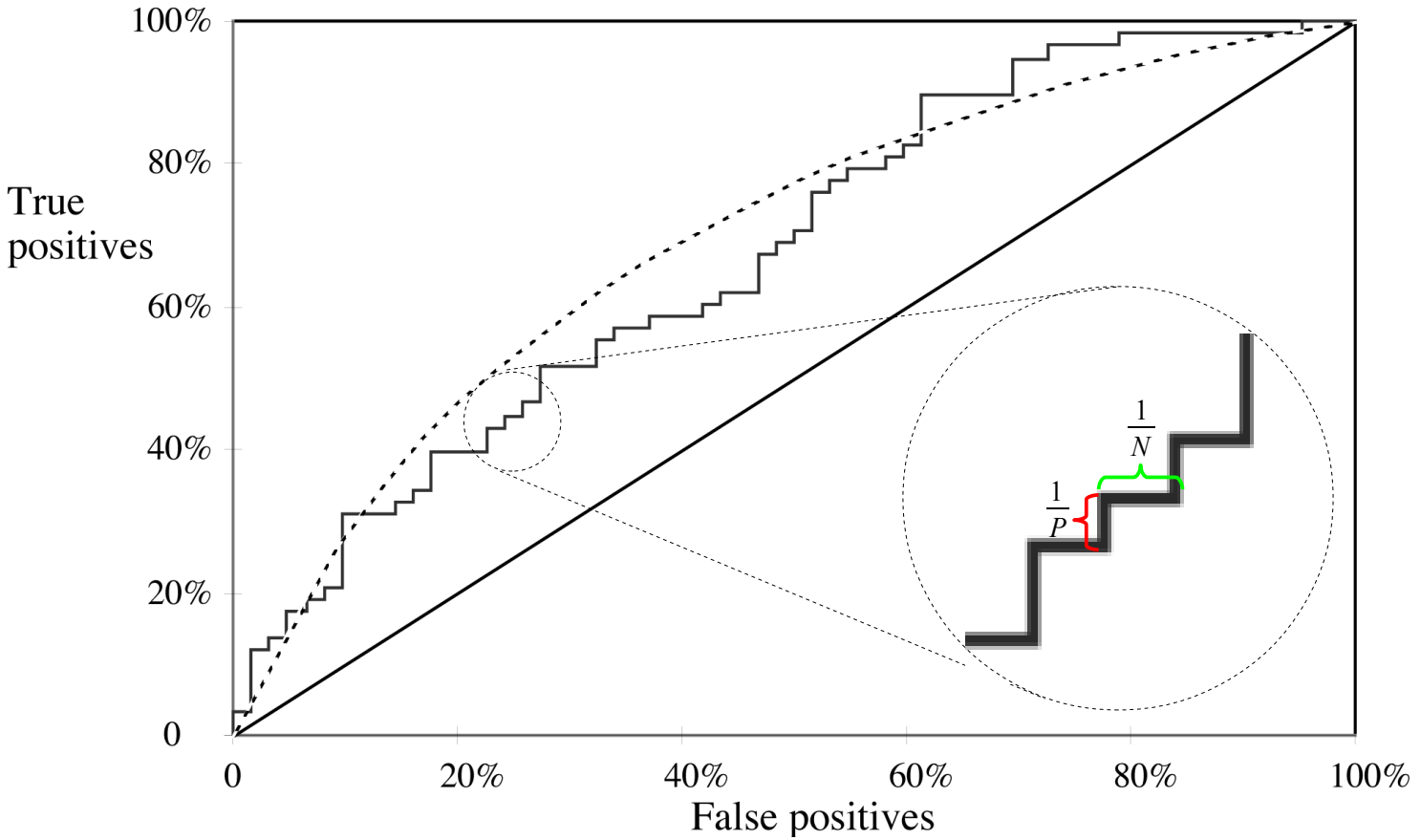
TECHNISCHE UNIVERSITÄT DARMSTADT

Performance of a ranker can be visualized via a ROC curve

- Naïve method:
  - consider all possible thresholds
    - only $k+1$ thresholds between the k instances need to be considered
  - each threshold corresponds to a new classifier
  - for each classifier
    - construct confusion matrix
    - plot classifier at point (fpr,tpr) in ROC space
- Practical method:
  - rank test instances on decreasing score $f(x)$
  - start in $(0,0)$
    - if the next instance in the ranking is +: move $1/P$ up
    - if the next instance in the ranking is –: move $1/N$ to the right
    - make diagonal move in case of ties

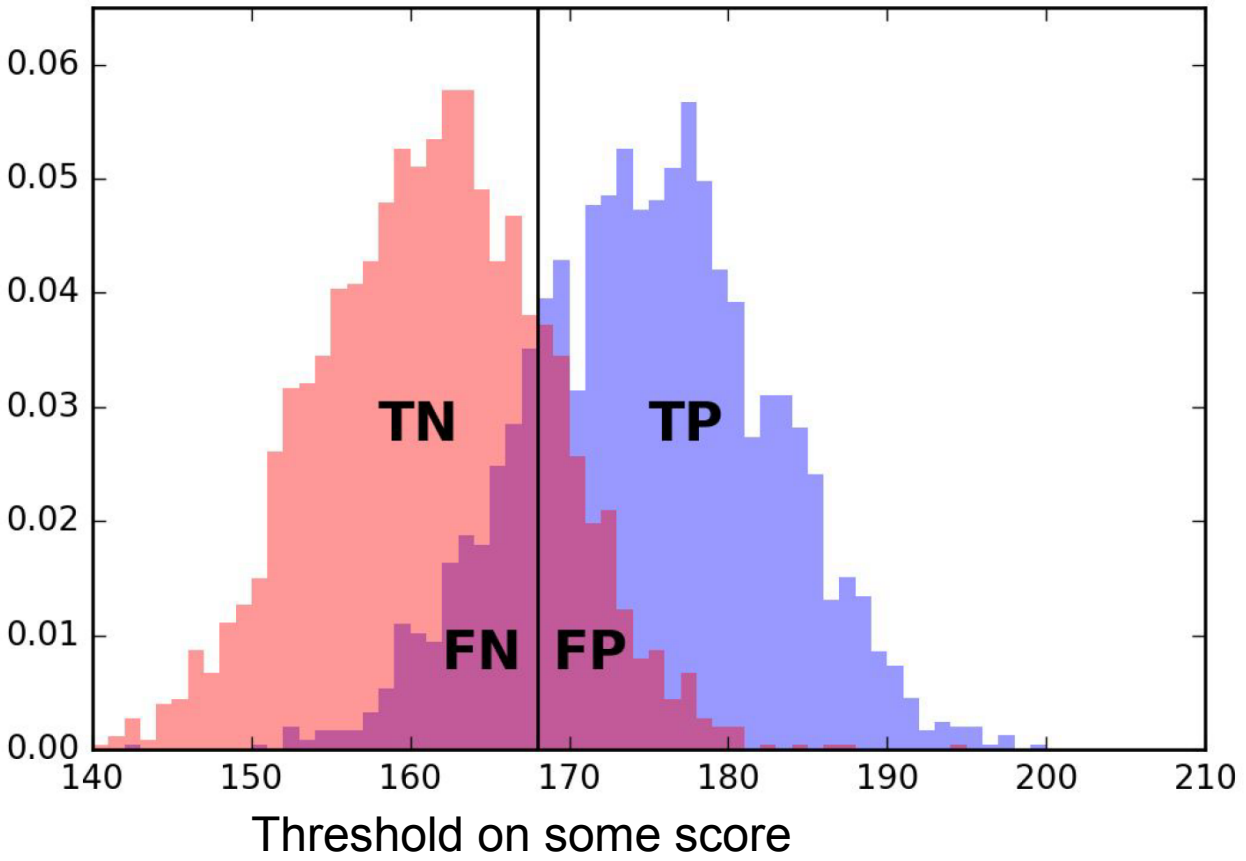
Slide adapted from P. Flach, ICML-04 Tutorial on ROC

Drawing ROC Curves for Rankers
A sample ROC curve



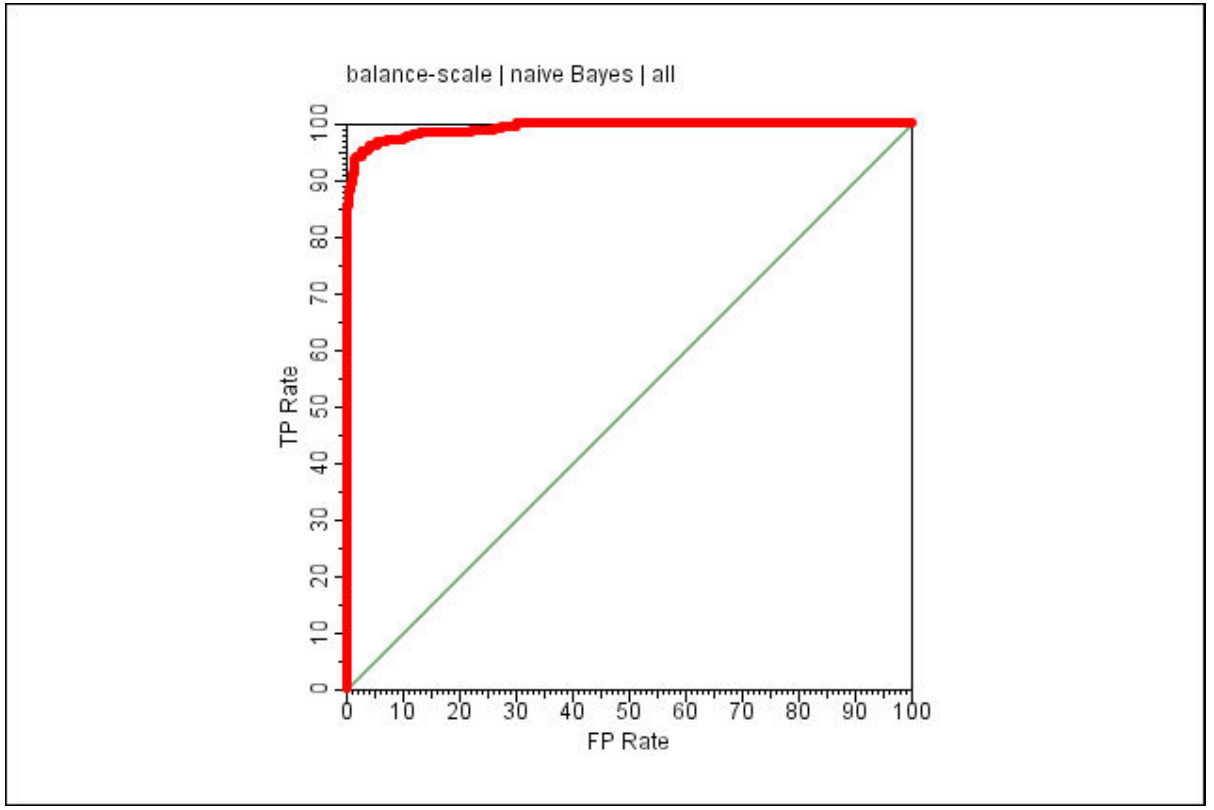
Slide adapted from Witten/Frank, Data Mining

Drawing ROC Curves for Rankers Alternative View



Skiena: CSE 519

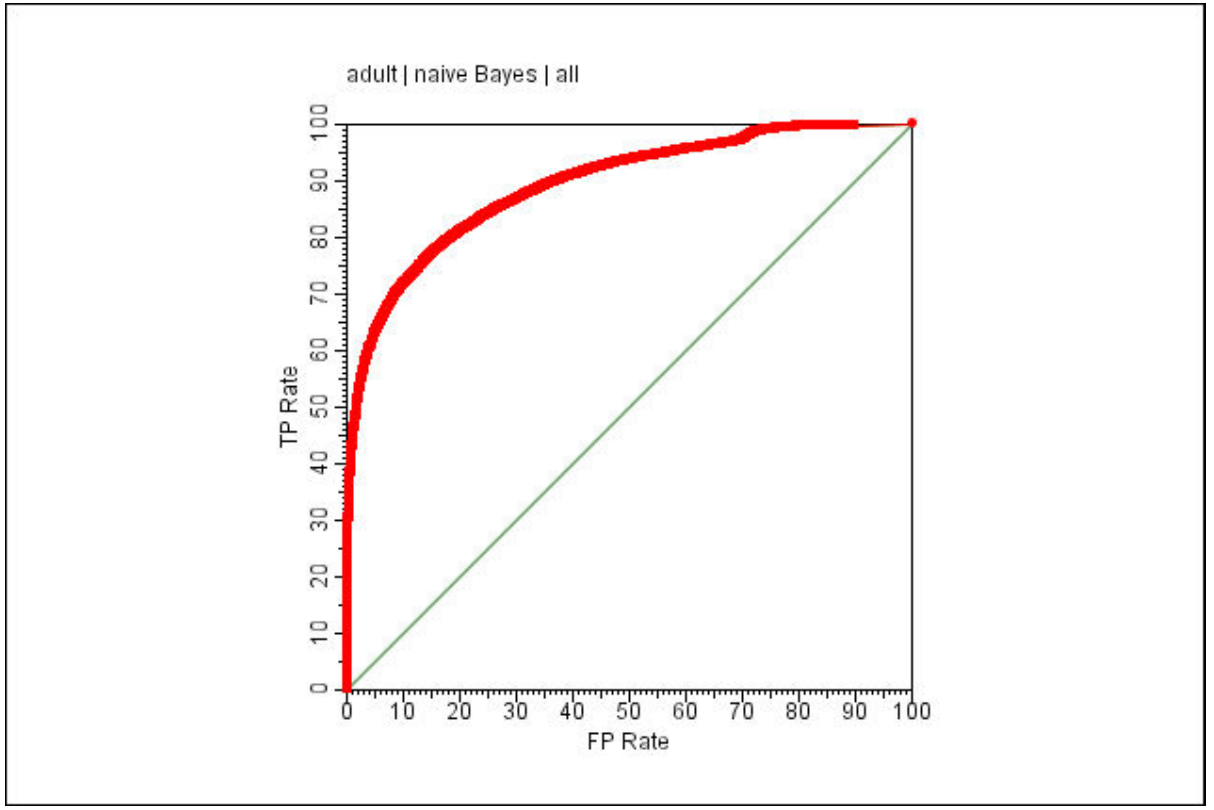
ROC analysis
Example curves for ranker



- Good separation between classes, convex curve

Slide adapted from P. Flach, ICML-04 Tutorial on ROC

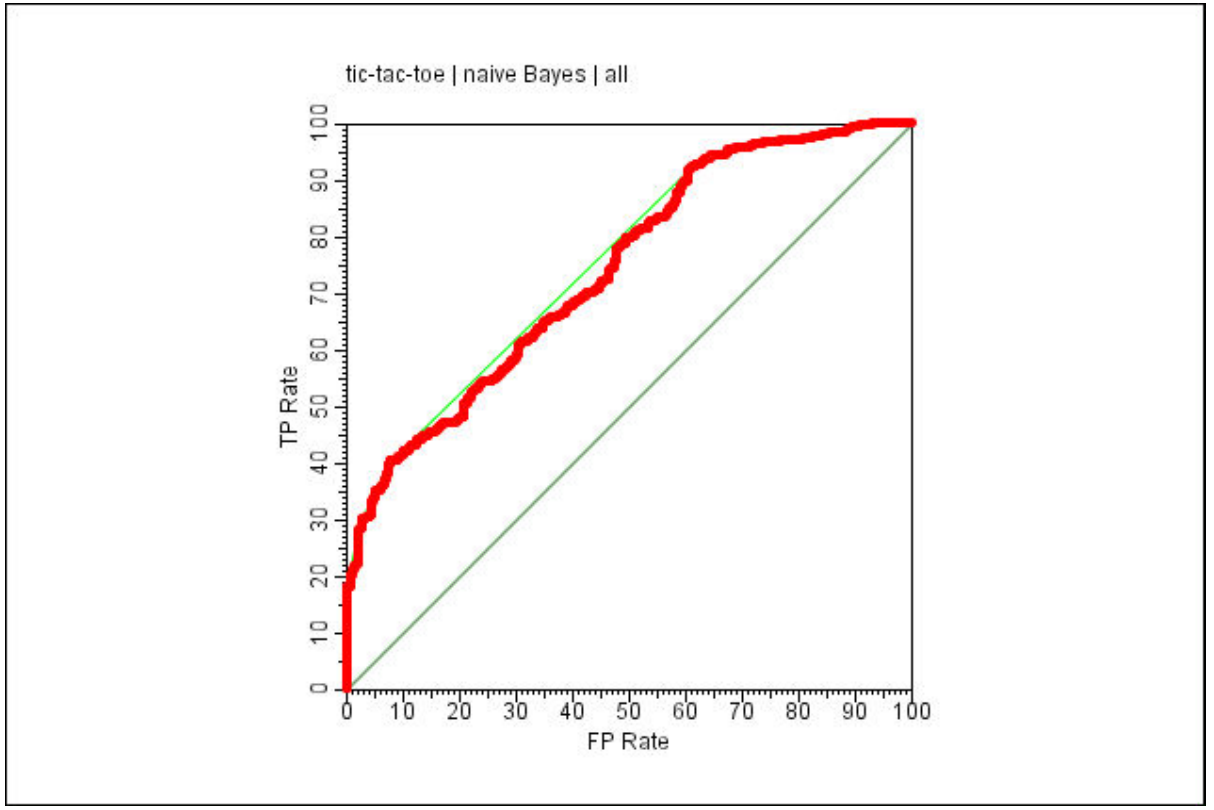
ROC analysis
Example curves for ranker



- Reasonable separation, mostly convex

Slide adapted from P. Flach, ICML-04 Tutorial on ROC

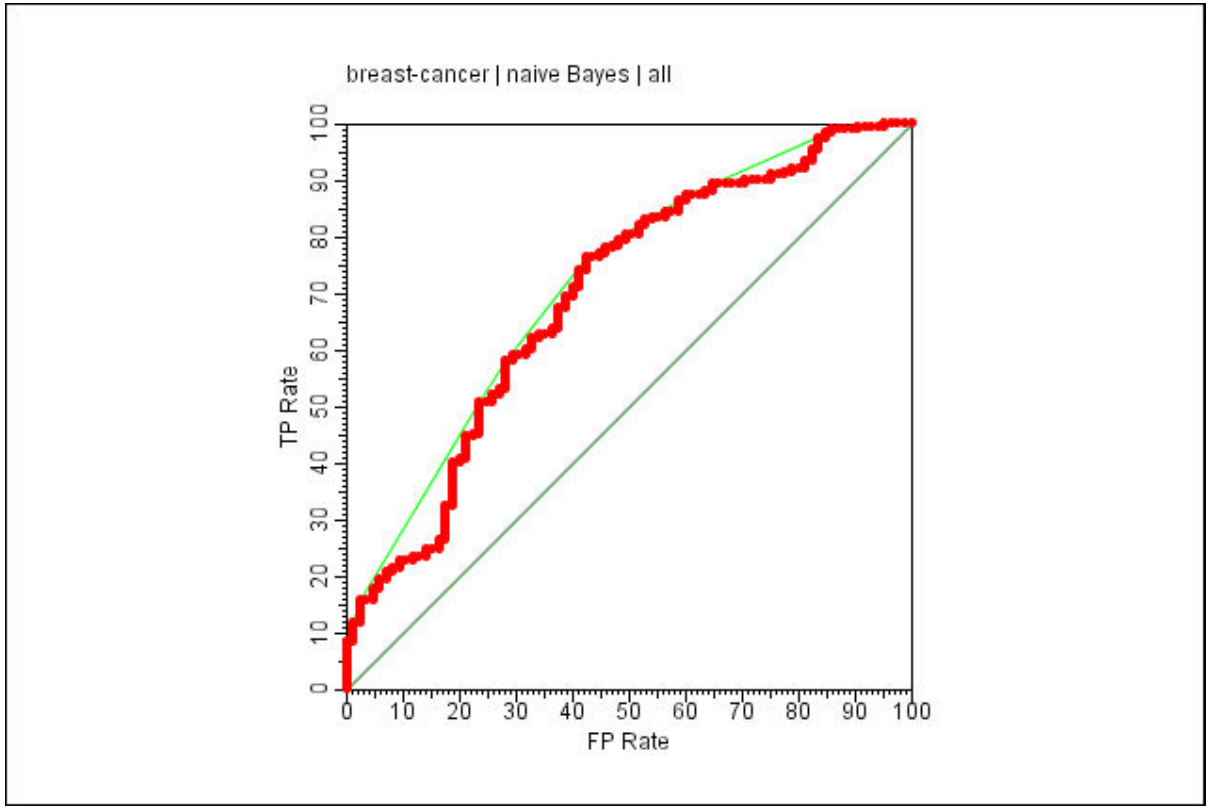
ROC analysis
Example curves for ranker



- Fairly poor separation, mostly convex

Slide adapted from P. Flach, ICML-04 Tutorial on ROC

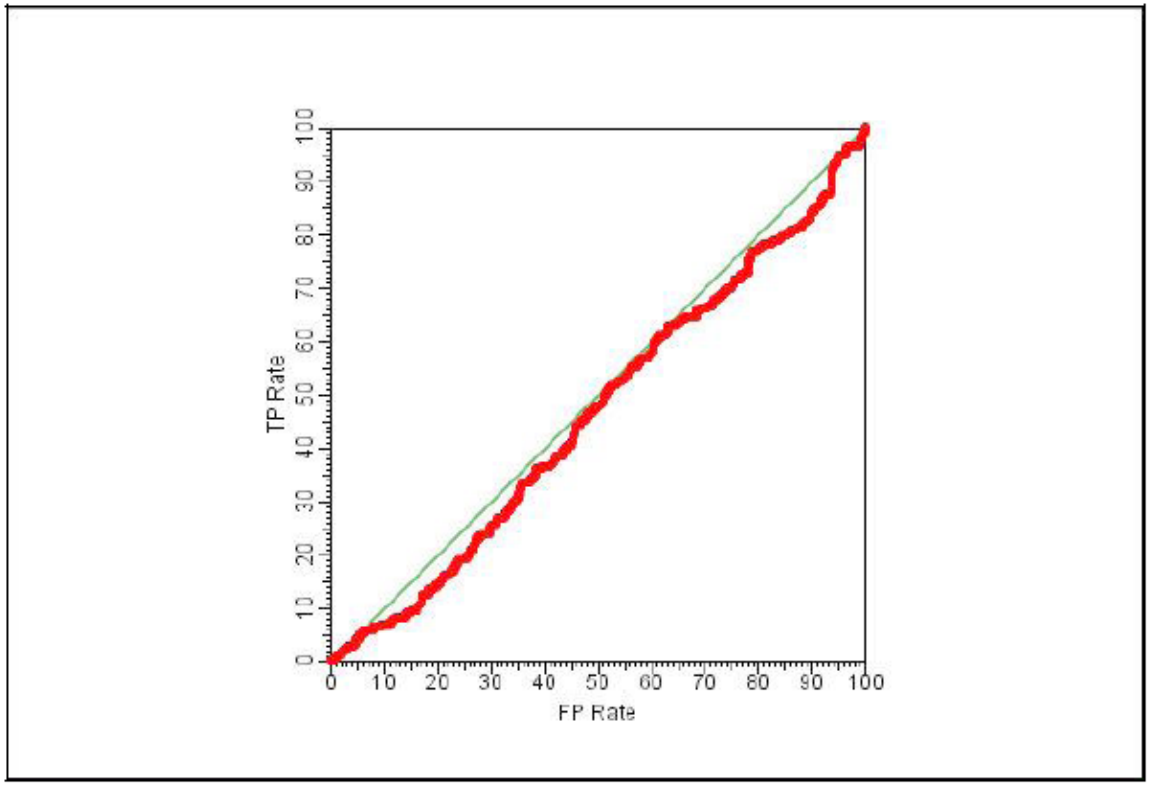
ROC analysis
Example curves for ranker



- Poor separation, large and small concavities

Slide adapted from P. Flach, ICML-04 Tutorial on ROC

ROC analysis
Example curves for ranker

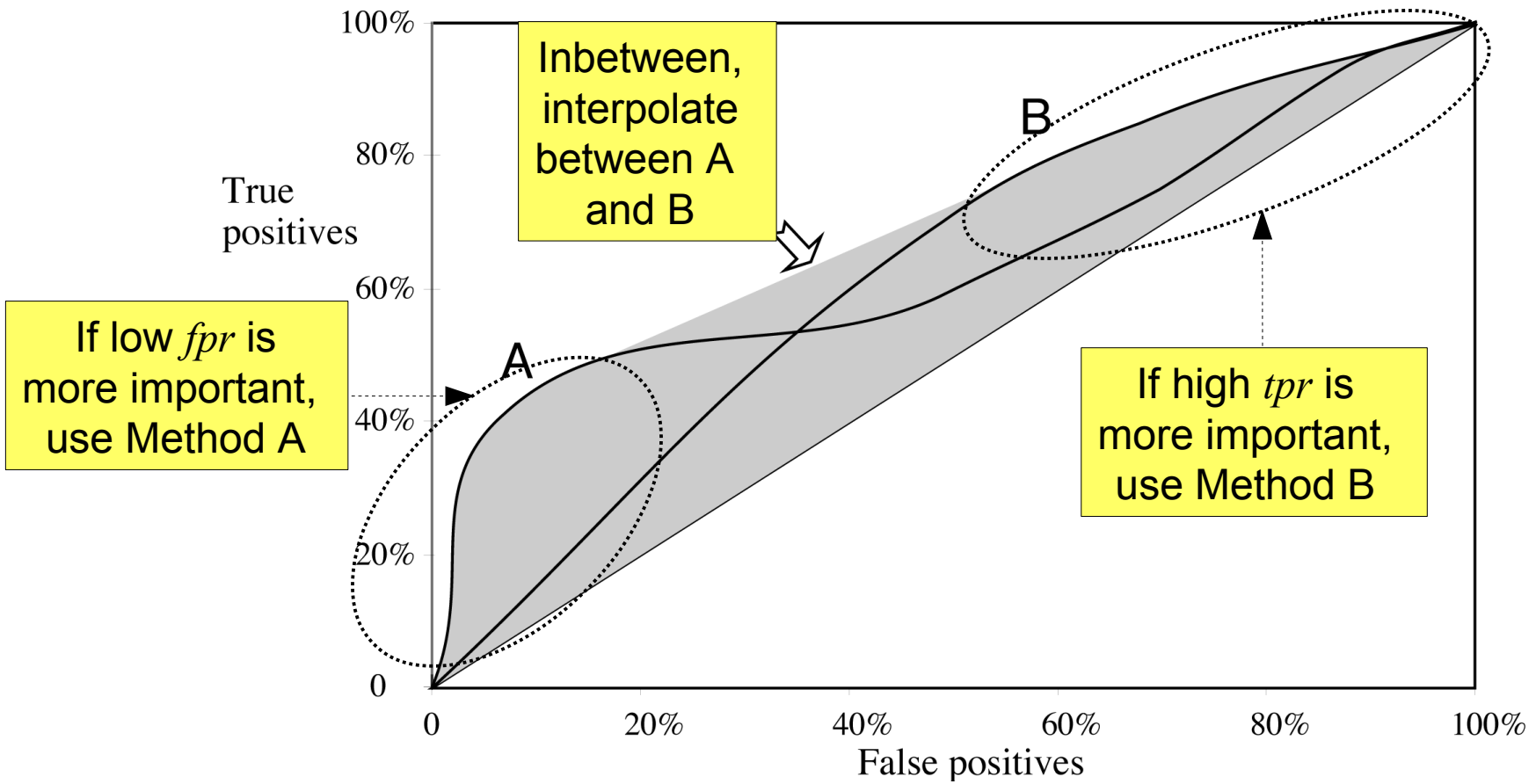


- Random performance

Slide adapted from P. Flach, ICML-04 Tutorial on ROC

ROC analysis
Comparing Rankers with ROC Curves

TECHNISCHE UNIVERSITÄT DARMSTADT



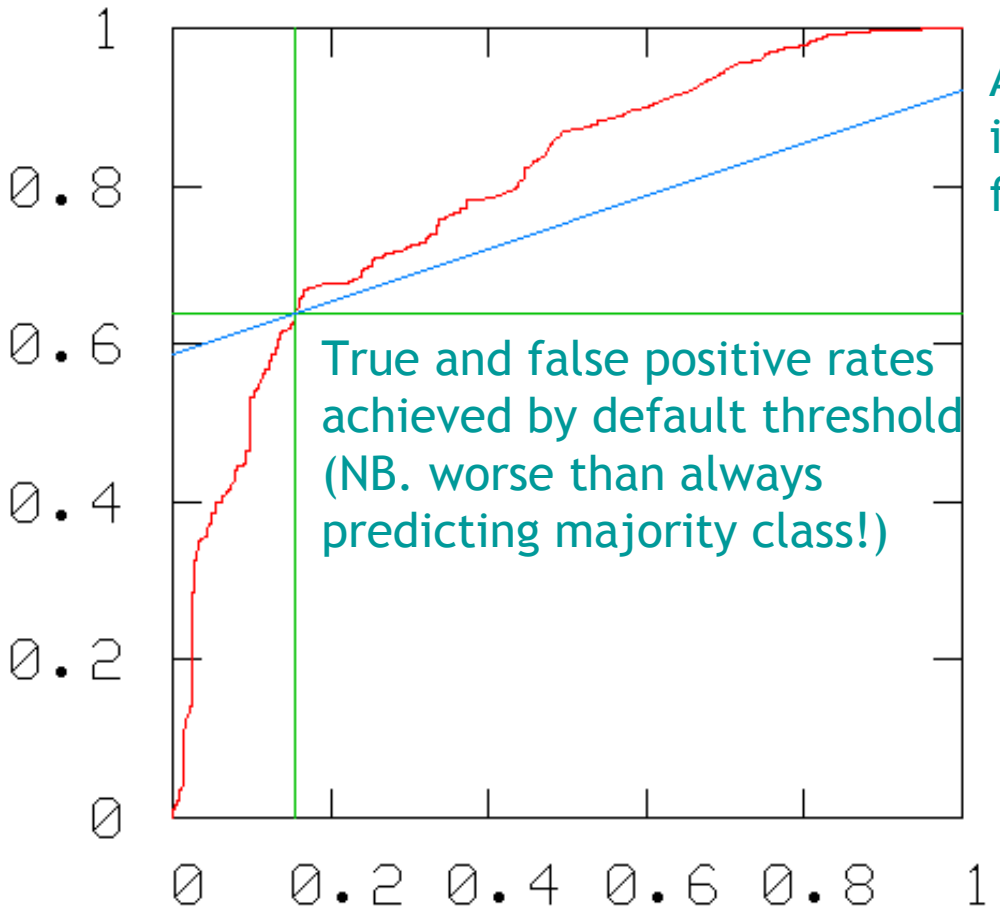
Slide adapted from Witten/Frank, Data Mining

Calibrating a Ranking Classifier

- What is the right threshold of the ranking score $f(x)$ if the ranker does not estimate probabilities?
  - classifier can be *calibrated* by choosing appropriate threshold that minimizes costs
  - may also lead to improved performance in accuracy if probability estimates are bad (e.g., Naïve Bayes)
- Easy in the two-class case:
  - calculate cost for each point/threshold while tracing the curve
  - return the threshold with minimum cost
- Non-trivial in the multi-class case

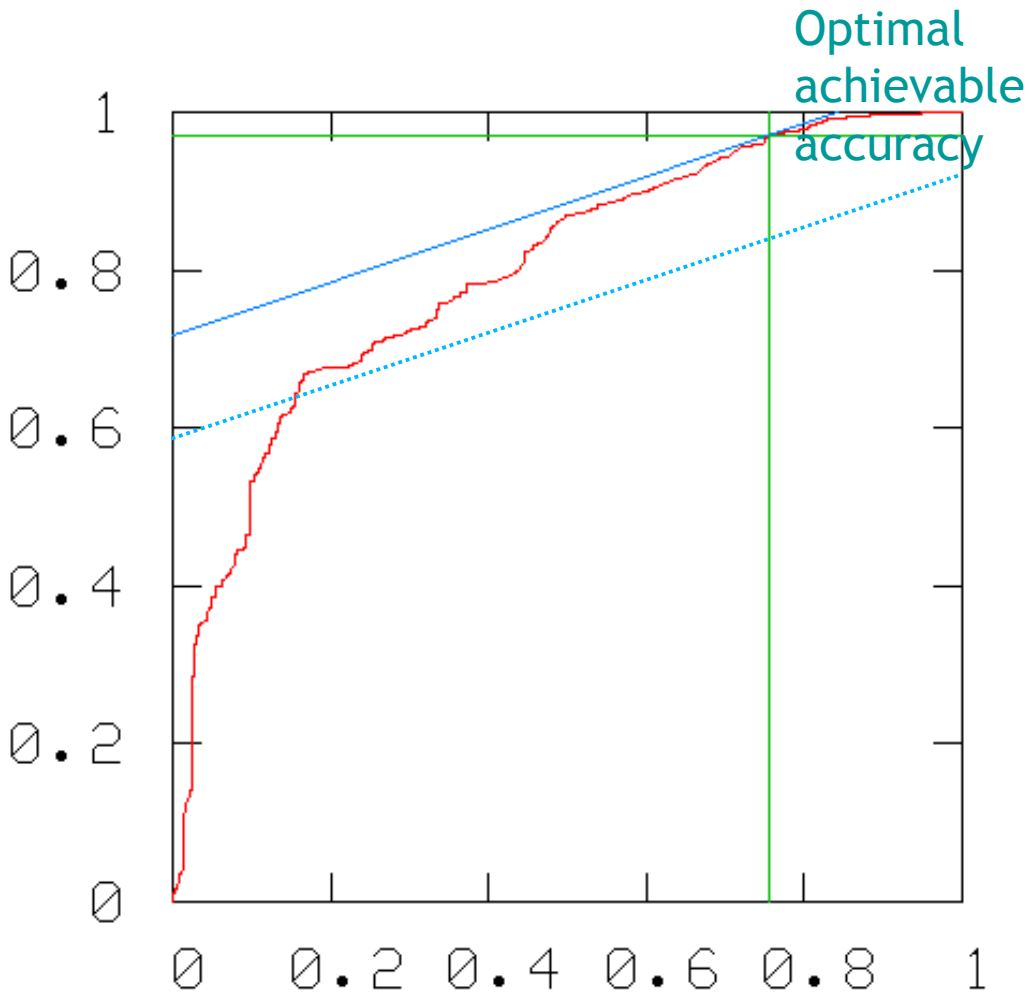
Note: threshold selection is part of the classifier training and must therefore not be performed on the test data!

Example: Uncalibrated threshold



Slide adapted from P. Flach, ICML-04 Tutorial on ROC

Example: Calibrated threshold



Slide adapted from P. Flach, ICML-04 Tutorial on ROC

What is missing?

TECHNISCHE UNIVERSITÄT DARMSTADT

- Comparison of classifiers
  - Aggregation of results and statistical tests
  - Hyper parameter optimization
  - Proper separation of dataset: train, validation, test
- Cost-sensitive
  - transformation into cost-sensitive classifiers
- Evaluation on other target spaces
  - Multi-class classification
  - Multi-label classification
  - Rankings
  - Regression