Evaluating classifier performance

Szymon Jaroszewicz

Institute of Computer Science
Polish Academy of Sciences
Warsaw, Poland

Ph. D. Programme 2012/2013

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Project co-financed by European Union within the framework of European Social Fund

How good is a classifier?

- We may look at two aspects:
  - Model's accuracy (or equivalently classification error)
  - Whether a model produces a correct *ordering* of cases from best to worst

Let's begin with accuracy.

PhD STUDIES
KAPITAŁ LUDZKI NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA EUROPEJSKI FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

Classification error

- We are interested in as low classification error as possible

$$P_{\mathcal{P}}\{M(X_1, \ldots, X_m) \neq Y\}$$

where $\mathcal{P}$ is the population distribution.

Problem

- We don't know the true distribution $\mathcal{P}$
- Typically only a sample $\mathbf{D}$ is available

PhD STUDIES
KAPITAŁ LUDZKI NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA EUROPEJSKI FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

Resubstitution error, generalization

- Simplest idea: just estimate the error on the training set **D**
- This is called the resubstitution error

Problem

- Resubstitution error is too optimistic
- It uses data which was used to build the model
- ... so it does not take overfitting into account

Project co-financed by European Union within the framework of European Social Fund

Resubstitution error, generalization

- Simplest idea: just estimate the error on the training set **D**
- This is called the resubstitution error

Problem

- Resubstitution error is too optimistic
- It uses data which was used to build the model
- ... so it does not take overfitting into account

Recall the example with PESEL in the tree root:

- 100% accuracy on the training set
- Useless on future data

PhD STUDIES
KAPITAŁ LUDZKI NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA EUROPEJSKI FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

Generalization, overfitting

- By generalization we mean the ability of the learning algorithm to discover true underlying relationships which are applicable to future data.
- Overfitting is in practice always present
- Error on future data is almost always larger then on the data used to build the model

Preventing overfitting is the most important problem in Machine Learning

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

Training and test datasets

- How to estimate the error on future data?
- Recall that we don't know the underlying distribution $\mathcal{P}$
- Main idea (other methods are essentially its variants):

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

Training and test datasets

- How to estimate the error on future data?
- Recall that we don't know the underlying distribution $\mathcal{P}$
- Main idea (other methods are essentially its variants):

- Split data available for training into two parts:
  - training set
  - test set
- The training set is used to build the model
- The test set is used to assess model accuracy

- The test set contains only data which was not used for building the model
- The classes of test examples are known, so the accuracy can be computed

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

Training and test datasets

Advantages:

- Error computed only on data not used for training
- Good statistical properties:
  - test sample is independent from the training sample
  - can apply various statistical tests
  - error estimates are unbiased

Disadvantages:

- Large amount of data is not used for model building
- This may be a big problem for small datasets (10s, even hundreds records)

Project co-financed by European Union within the framework of European Social Fund

Crossvalidation (CV)

k-fold crossvalidation:

1. Split the training dataset $\mathbf{D}$ into k equal parts $\mathbf{D}_i$
2. **For** $i = 1, \ldots, k$:
3. Build a model M_i on $\mathbf{D} \setminus \mathbf{D}_i$
4. $e_i =$ error of M_i on $\mathbf{D}_i$
5. **Return** final error estimate $\frac{1}{k} \sum_{i=1}^{k} e_i$

PhD STUDIES
KAPITAŁ LUDZKI NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA EUROPEJSKI FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

Crossvalidation (CV)

k-fold crossvalidation:

1. Split the training dataset $\mathbf{D}$ into k equal parts $\mathbf{D}_i$
2. **For** $i = 1, \ldots, k$:
3. Build a model M_i on $\mathbf{D} \setminus \mathbf{D}_i$
4. $e_i =$ error of M_i on $\mathbf{D}_i$
5. **Return** final error estimate $\frac{1}{k} \sum_{i=1}^{k} e_i$

- Each part of data is used in turn for testing and the remaining parts for training
- All data is used for both training and testing
- For every record in $\mathbf{D}$ we get a 'test set prediction' which we can compare with the true class

Project co-financed by European Union within the framework of European Social Fund

Crossvalidation (CV)

- Most popular technique, the default choice in most ML systems
- The overall scheme:
  1. Build a model on full data
  2. Use cross-validation's results as the estimate of this model's error
- Typically $k = 10$. Other values such as 3 and 5 are sometimes being used.

Crossvalidation (CV)

Advantages:

- All data is used for both training and testing
- Works very well in practice

Disadvantages:

- Train and test samples in separate folds are not independent
- Standard deviation of error can be computed, but it's hard to give statistical guarantees
- Error estimate is biased (too pessimistic) since training samples are smaller than full datasets
- Longer learning time, need to build the model k times

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

Leave-one-out crossvalidation

- Crossvalidation taken to the extreme
- At each iteration we leave out one of the cases and use it as test set
- In other words: it is crossvalidation with $k = |\mathbf{D}|$ ($|\mathbf{D}|$ is the $\#$ of records in training data)

Advantages:

- Very useful when little data is available
- For some models (e.g. regression) estimates of leave-one-out error can be obtained without actually building all the models

Disadvantages:

- All models are build on very similar data (surprisingly, this turns out not to be a problem in practice)
- Extremely long running time for larger datasets

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

Bootstrap

- Let $N = |\mathbf{D}|$
- Take a random sample of size N from $\mathbf{D}$ with replacement
  - records in the sample are used for training
  - records not in the sample are used for testing
- Repeat the above procedure several times and average the error over all iterations

At each iteration $\approx 1 - e^{-1} \approx 0.632$ of records is used for the training sample

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

Bootstrap

Advantages:

- Every model is built on a sample of size N (the original size)
- Since the iteration is repeated many times we get a large number of error estimates
  - we can draw a histogram of the error distribution
  - we can get confidence intervals on the error (e.g. from the histogram)

Disadvantages:

- The samples are not independent (neither within each training set nor between bootstrap iterations)
- Histograms etc. do not necessarily reflect true sample error distribution

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

Repeated validation

- Each method can be repeated several times and the results averaged
- Typically used with repeated random train-test splits, but CV can also be repeated
- Repeating the validation several times, gives much more stable estimates
- And smoother curves
- But the bias remains... so the approach has clear limits

PhD STUDIES

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Project co-financed by European Union within the framework of European Social Fund

Training / test / validation datasets

- Models have many parameters which are impossible to choose in advance

A typical mistake

1. Build many models with different parameters
2. Pick the one which is best on the test set
3. Use the test set to estimate the generalization error of the best model

- We are dealing here with a problem of 'secondary' overfitting, when the model (through parameter tuning) fits a specific test set very well but will not generalize to future data

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

Training / test / validation datasets

- Solution: split the available data into 3 parts:
  - training set used to build the models
  - validation set used to tune model parameters
  - test set used for final accuracy estimation

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

Training / test / validation datasets

- Solution: split the available data into 3 parts:
  - training set used to build the models
  - validation set used to tune model parameters
  - test set used for final accuracy estimation
- We have sees this arrangement already in decision trees:
  - build the tree on the training set
  - prune the tree on the validation set
    - it was not used for building the tree
    - this limits the effects of overfitting during pruning
  - estimate the error of the pruned tree on the test set

Confusion matrix

- If > 2 classes, many types of mistakes possible
- The confusion matrix shows us what types of errors were made
  - rows correspond to true classes
  - columns correspond to predicted classes
  - entries correspond to numbers of cases
- Example:
  - classes **a**, **b**, **c**
  - 50 cases for each class, 150 cases total

predicted class →	**a**	**b**	**c**	
	50	0	0	**a**
	0	44	6	**b**
	0	3	47	**c**
				↑
				true class

Project co-financed by European Union within the framework of European Social Fund

Confusion matrix

predicted class →	a	b	c	
	50	0	0	**a**
	0	44	6	**b**
	0	3	47	**c**
				↑ true class

- We can see that
  - all cases of class **a** are predicted correctly
  - 6 cases of class **b** were classified as **c**
  - 3 cases of class **c** were classified as **b**
- Diagonal elements correspond to correct predictions
  Off-diagonal entries to incorrect predictions

Confusion matrix

$$
\begin{array}{lccccl}
\text{predicted class} \rightarrow & \mathbf{a} & \mathbf{b} & \mathbf{c} & & \\
 & 50 & 0 & 0 & \mathbf{a} & \\
 & 0 & 44 & 6 & \mathbf{b} & \\
 & 0 & 3 & 47 & \mathbf{c} & \\
 & & & & \uparrow & \\
 & & & & \text{true class} &
\end{array}
$$

- Confusion matrix can be computed
    - on the test set
    - by summing matrices obtained during each fold of CV
    - by averaging over bootstrap samples
    - ...
- The matrix may contain probabilities instead of counts

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

What about the money?

- If a classification model is applied in a business setting, mistakes cost money
- Costs of mistakes are specified using a cost matrix $\mathbf{C}$:
  - entry $\mathbf{C}_{ij}$ gives the cost of classifying a case belonging to class i as class j
  - $\mathbf{C}_{ij} \geq 0$
  - $\mathbf{C}_{ii} = 0$
- Analogously we can use the benefit matrix and maximize benefits instead of minimizing costs.

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

What about the money?

- If a classification model is applied in a business setting, mistakes cost money
- Costs of mistakes are specified using a cost matrix $\mathbf{C}$:
    - entry $\mathbf{C}_{ij}$ gives the cost of classifying a case belonging to class i as class j
    - $\mathbf{C}_{ij} \geq 0$
    - $\mathbf{C}_{ii} = 0$
- Analogously we can use the benefit matrix and maximize benefits instead of minimizing costs.

- Exercise: given that for inputs $\mathbf{x}$ a model M predicts class y_i and given $P_{\mathcal{P}}(y|\mathbf{x})$ find the expected misclassification cost
- Exercise: Given a cost matrix $\mathbf{C}$ and the joint distribution $\mathcal{P}$ of $X_1, \ldots, X_m, Y$ find the optimal (least costly) classifier

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

Cost sensitive learning

- Costs can be taken into account after a model is built to assess its performance or pick the cutoff threshold
- Costs can be taken into account while building a model
- E.g. Breiman et al. suggest:
    - a modified Gini index for selecting splits in decision trees
    - assigning weights to training records based on costs

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

Cost sensitive learning

- Costs can be taken into account after a model is built to assess its performance or pick the cutoff threshold
- Costs can be taken into account while building a model
- E.g. Breiman et al. suggest:
  - a modified Gini index for selecting splits in decision trees
  - assigning weights to training records based on costs

- Cost sensitive learning is an important research topic

- But the benefits are not always worthwhile
- If you model $P(Y|X_1, \ldots, X_m)$ well, you can adjust for arbitrary cost matrix *after* building the model

Project co-financed by European Union within the framework of European Social Fund

Ordering based model quality measures. Curves.

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

Score

Score

- Assume a two-class problem: $\mathrm{Dom}(Y) = \{+, -\}$
- Call $+$ the positive class
- (if there are more classes, pick one of them and call it positive)
- A score is a value indicating model's confidence that the class is $+$
- The higher the score, the more confident the model is that the class is positive

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

Rankers

Ranker

A ranker is a model which, given X's, returns a score

I.e. it *ranks* examples based on its certainty that they belong to the positive class

- Note the difference from classifiers which return the predicted class Y
- In fact the difference is much less pronounced
  - all models we are going to see in the course are actually rankers or can easily be converted to rankers
  - every ranker can be converted to a classifier
  - (in fact a ranker corresponds to a series of classifiers)

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

Classifiers as rankers

How can a classifier also be a ranker?

- Very often the score is simply the probability $P(Y = +|\mathbf{x})$ of the positive class
- So every model which predicts class probabilities is automatically a ranker

- But this does not have to be the case
- All that matters is the ordering: higher score = more likely the $+$ class
- A suitable value can be found also in non-probabilistic classifiers
  - e.g. distance from the decision boundary in SVMs

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

Decision trees as rankers

- How can a decision tree assign scores?
- After pruning, most leaves 'contain' several training cases
- Estimate class probabilities using those cases and use them as score
- Laplace correction (we'll discuss it later) may be used to smooth the probabilities

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

Rankers as classifiers

How to convert a ranker to a classifier?

- Pick a threshold θ
- Classify all cases with *score* $\geq \theta$ as $+$ and the remaining ones as $-$

- Different thresholds correspond to different classifiers
- The higher the value of θ, the less inclined we are to classify an example as positive

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

Rankers as classifiers

How to pick the threshold?

- If the score is the class probability, $\theta = \frac{1}{2}$ is equivalent to picking the more probable class

- In practice, costs are often involved. Then, picking a threshold becomes a business decision. A quote I heard at a conference:

  > *"The model should just order the cases from best to worst, it's business people's task to pick the threshold"*

- Different types of curves are used to help such decisions

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

Curves

All (almost) types of curves describing classifier performance are drawn as follows

1. Sort the test cases by decreasing score
2. For every possible threshold compute two statistics A and B describing classifier performance
3. Draw a point with
   - x coordinate equal to A
   - y coordinate equal to B

PhD STUDIES
KAPITAŁ LUDZKI NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA EUROPEJSKI FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

Curves

By taking different statistics for the axes we get different types of curves

- ROC curves – more popular in ML research and e.g. medicine
- Cumulative gains curves (a.k.a. lift curves) – used mainly in business applications
- Other types:
  - Lift curves
  - Precision / recall charts
  - Calibration curves
  - ...

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

True positives, false negatives etc.

		True class +	True class −
Predicted class	+	True positive	False positive
	−	False negative	True negative

- True positive: a case predicted to belong to class + which really belongs to class +
- False positive: a case predicted to belong to class + which really belongs to class −
- True negative: a case predicted to belong to class − which really belongs to class −
- False negative: a case predicted to belong to class − which really belongs to class +

PhD STUDIES

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Project co-financed by European Union within the framework of European Social Fund

True positive rate, etc.

- Define quantities:

$$\begin{aligned} TP(\theta) &= |\{\mathbf{x}, y \in \mathbf{D} : M(\mathbf{x}) = + \wedge y = +\}| \\ &= |\{\mathbf{x}, y \in \mathbf{D} : score(\mathbf{x}) \geq \theta \wedge y = +\}| \\ TPR(\theta) &= \frac{TP(\theta)}{|\{\mathbf{x}, y \in \mathbf{D} : y = +\}|} \end{aligned}$$

- TPR measures the percentage of all positive responses which have been picked out by the model
- Also called sensitivity

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

True positive rate, etc.

- Analogously:

$$\begin{aligned} FP(\theta) &= |\{\mathbf{x}, y \in \mathbf{D} : M(\mathbf{x}) = + \wedge y = -\}| \\ &= |\{\mathbf{x}, y \in \mathbf{D} : score(\mathbf{x}) \geq \theta \wedge y = -\}| \\ FPR(\theta) &= \frac{FP(\theta)}{|\{\mathbf{x}, y \in \mathbf{D} : y = -\}|} \end{aligned}$$

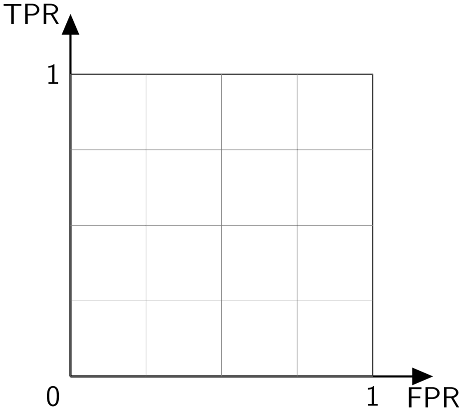
- FPR measures the percentage of all negative responses which the model wrongly predicted as positive

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

ROC curves

- For each threshold θ:
  - y-axis = True Positive Rate (TPR)
  - x-axis = False Positive Rate (FPR)
- So we plot the number of positive cases we picked vs. the number of negative cases we had to pick to achieve this
- ROC – Receiver Operating Characteristic. First used for radar signals, thus the name

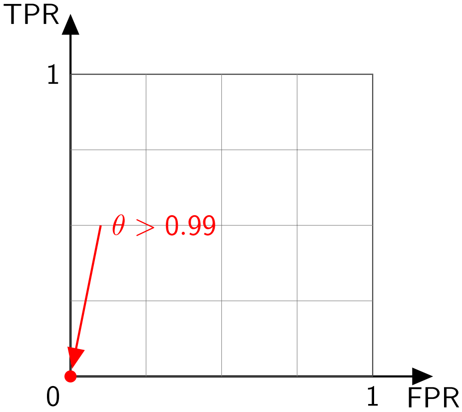
ROC curves – example



score	Y
0.99	+
0.88	+
0.71	−
0.55	+
0.32	−
0.15	+
0.08	−
0.01	−

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

ROC curves – example

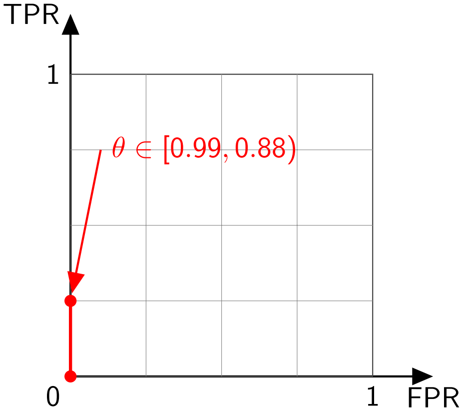


score	Y
0.99	+
0.88	+
0.71	−
0.55	+
0.32	−
0.15	+
0.08	−
0.01	−

Initial point – no cases classified as +

PhD STUDIES
KAPITAŁ LUDZKI NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA EUROPEJSKI FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

ROC curves – example

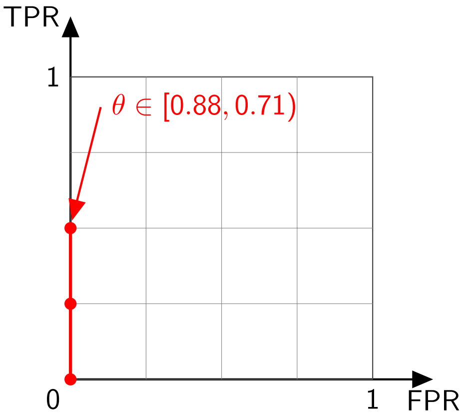


score	Y
→ 0.99	$+$
0.88	$+$
0.71	$-$
0.55	$+$
0.32	$-$
0.15	$+$
0.08	$-$
0.01	$-$

First case is positive, move up

PhD STUDIES
KAPITAŁ LUDZKI NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA EUROPEJSKI FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

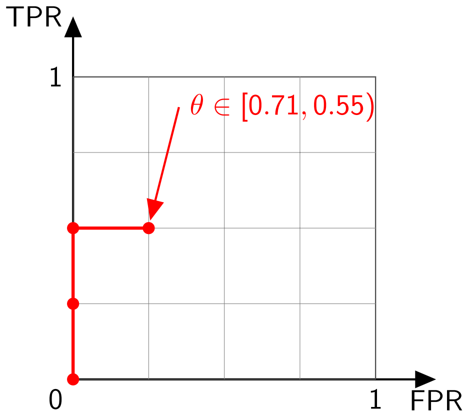
ROC curves – example



	score	Y
	0.99	+
$\rightarrow$	0.88	+
	0.71	−
	0.55	+
	0.32	−
	0.15	+
	0.08	−
	0.01	−

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

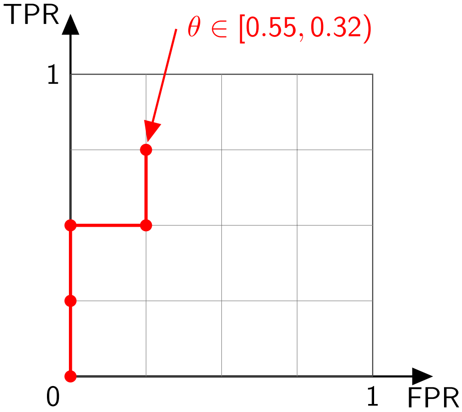
ROC curves – example



score	Y
0.99	$+$
0.88	$+$
$\rightarrow$ 0.71	$-$
0.55	$+$
0.32	$-$
0.15	$+$
0.08	$-$
0.01	$-$

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

ROC curves – example



	score	Y
	0.99	$+$
	0.88	$+$
	0.71	$-$
$\rightarrow$	0.55	$+$
	0.32	$-$
	0.15	$+$
	0.08	$-$
	0.01	$-$

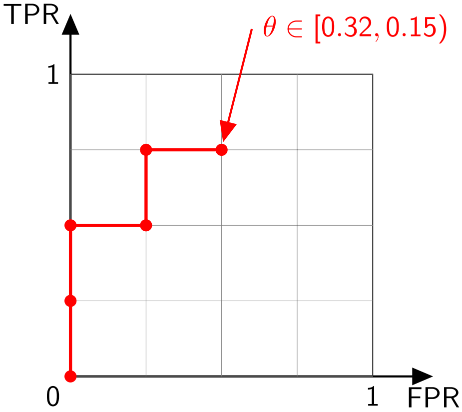
PhD STUDIES

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Project co-financed by European Union within the framework of European Social Fund

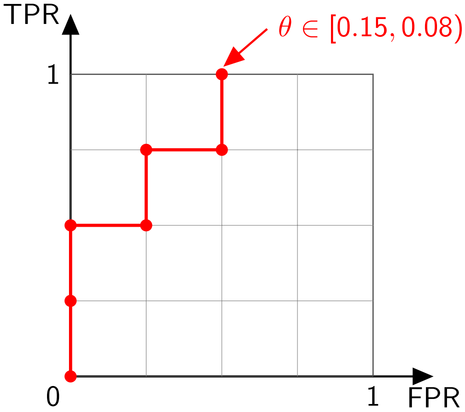
ROC curves – example



	score	Y
	0.99	+
	0.88	+
	0.71	−
	0.55	+
→	0.32	−
	0.15	+
	0.08	−
	0.01	−

PhD STUDIES
KAPITAŁ LUDZKI NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA EUROPEJSKI FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

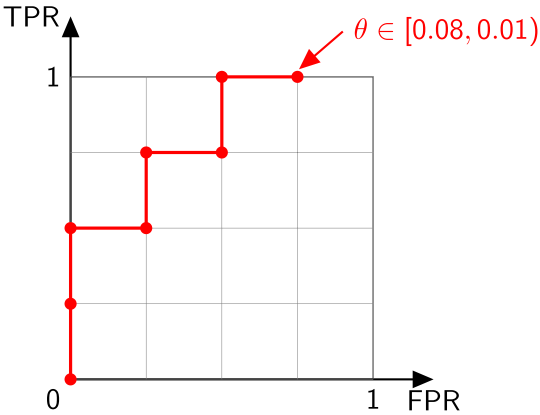
ROC curves – example



	score	Y
	0.99	+
	0.88	+
	0.71	–
	0.55	+
	0.32	–
$\rightarrow$	0.15	+
	0.08	–
	0.01	–

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

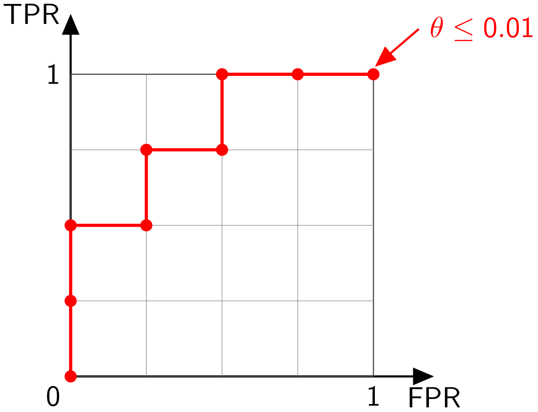
ROC curves – example



	score	Y
	0.99	+
	0.88	+
	0.71	−
	0.55	+
	0.32	−
	0.15	+
$\rightarrow$	0.08	−
	0.01	−

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

ROC curves – example



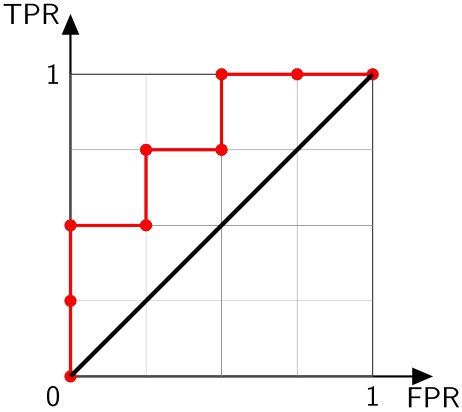
	score	*Y*
	0.99	$+$
	0.88	$+$
	0.71	$-$
	0.55	$+$
	0.32	$-$
	0.15	$+$
	0.08	$-$
$\rightarrow$	0.01	$-$

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Project co-financed by European Union within the framework of European Social Fund

ROC curves – example

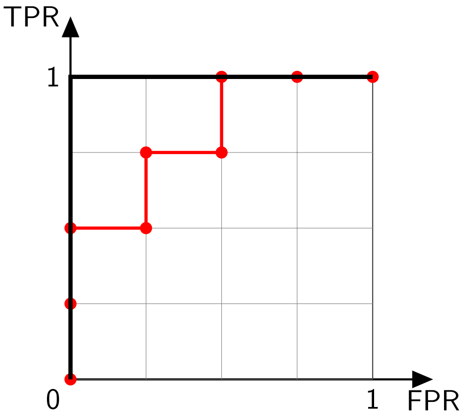


score	Y
0.99	$+$
0.88	$+$
0.71	$-$
0.55	$+$
0.32	$-$
0.15	$+$
0.08	$-$
0.01	$-$

Model assigning a class at random corresponds to the diagonal line

Project co-financed by European Union within the framework of European Social Fund

ROC curves – example

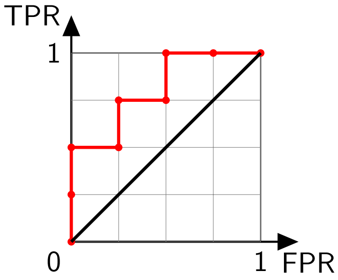


score	Y
0.99	$+$
0.88	$+$
0.71	$-$
0.55	$+$
0.32	$-$
0.15	$+$
0.08	$-$
0.01	$-$

Perfect classification: goes through upper left corner

PhD STUDIES
KAPITAŁ LUDZKI NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA EUROPEJSKI FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

ROC curves – interpretation



- The further the curve from the diagonal, the better the model
- The more true positives we pick up, the more false positives we have to accept
  - we can get 50% of true positives with no false positives (never happens in practice)
  - we can get 75% of true positives, but we have to 'pay' by accepting 25% of the false positives

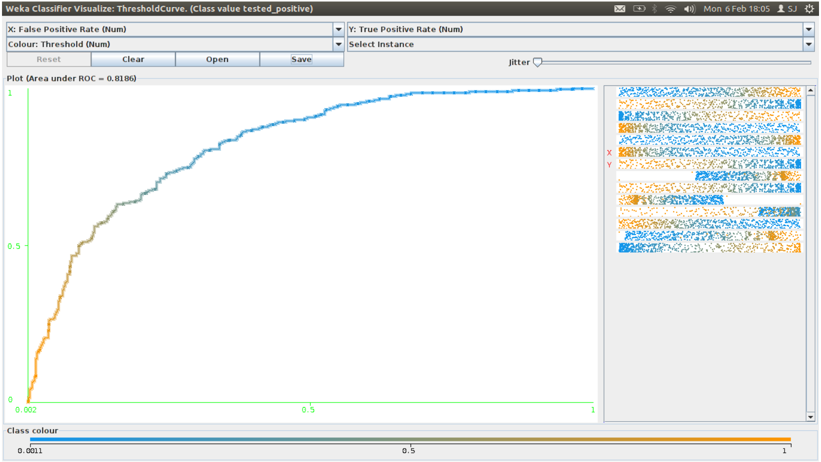
PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

ROC curves – a remark

- The ROC curve above was a series of steps
- This is an effect of a small data sample
- As the number of training examples N grows, the curve becomes smoother and smoother
- This is true also for other types of curves we are going to see

A real ROC curve (from Weka)

PIMA-Indians-diabetes dataset



Project co-financed by European Union within the framework of European Social Fund

ROC curves – Interpretation

- Example: classifier used as a medical test for a disease
- Positive test outcome = disease present
- High thresholds:
  - we catch few people with the disease
  - results are reliable, there are few false alarms
- Low thresholds:
  - we catch most people with the disease
  - results are not reliable, there are many false alarms
- The decision depends on the disease and should be made by a doctor

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

Statistical interpretation

- A classifier can be interpreted as a statistical test with
  - H_0: the given case belongs to class $-$
  - H_1: the given case belongs to class $+$
- In an ROC curve:
  - the x-axis indicates the type-I error
  - the y-axis indicates the power of the test $= 1 -$ type-II error
- The curve allows us to visualize the power vs. type-I error for various classification thresholds

Project co-financed by European Union within the framework of European Social Fund

Area Under the ROC Curve (AUC)

- The picture of classifier performance given by the ROC curve cannot really be summarized by a single number
- But there is always a strong desire to do so
- The Area Under the ROC curve (AUC) is one such (useful) attempt
  - the larger the area, the further the curve is likely to be from the diagonal
  - the best possible model (and only this model) has AUC $= 1$

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

Area Under the ROC Curve (AUC)

Other interpretations:

- In the population or sample case:

$$AUC = P(score_{+} > score_{-}),$$

  where $score_{+} \sim P(score|Y = +)$, $score_{-} \sim P(score|Y = -)$
- i.e. the probability that the model gives a higher score to a random case from the $+$ class than to a random case from the $-$ class

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

Area Under the ROC Curve (AUC)

Computation for a sample:

- Apply the trapezoidal rule to the ROC curve systematically underestimates the AUC
- Use the formula

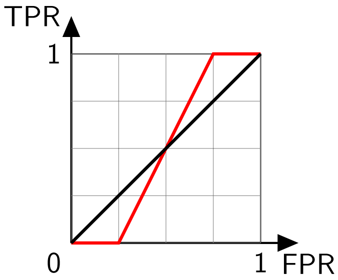
$$\widehat{AUC} = \frac{1}{N_+ N_-} \sum_{\mathbf{x}_+ \in \mathbf{D}_+} \sum_{\mathbf{x}_- \in \mathbf{D}_-} I[score(\mathbf{x}_+) > score(\mathbf{x}_-)]$$

  where N_+ is the number of positive cases,
  $\mathbf{D}_+ = \{\mathbf{x}, y \in \mathbf{D} : y = +\}$
  N_-, $\mathbf{D}_-$ analogous
- Sample version is equivalent to the Mann-Whittney U statistic used in some nonparametric tests
- Confidence intervals based on normal approximation can be obtained

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

Area Under the ROC Curve (AUC)

Single number is not enough:



- Both classifiers have AUC $= 0.5$
- The black classifier is useless
- The red classifier may be quite useful

ROC curves – practical considerations

- Drawing the ROC curve based on crossvalidation
  - Simply combine test data from all the folds and draw a single curve
- Computing AUC based on crossvalidation
  - Combine all data and compute a single AUC
  - or: Average AUC from all folds + get a confidence interval

PhD STUDIES

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Project co-financed by European Union within the framework of European Social Fund

Averaging ROC curves

- Suppose we want to average several ROC curves (e.g. from different train/test splits)
- Simply averaging the y-axis values for each x does not make much sense:
  - We would average TPR for given FPR, but in practice we cannot control the FPR of a classifier
- It is better to average the TPR and FPR for a given threshold θ
- We get two confidence intervals: for TPR and for FPR

PhD STUDIES

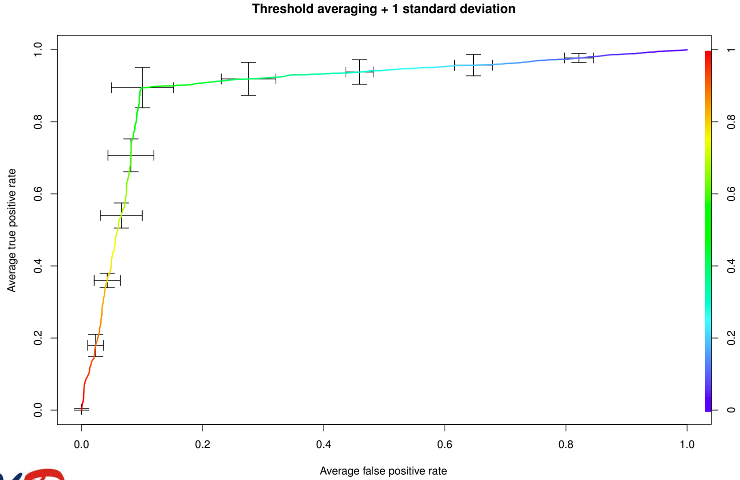
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Project co-financed by European Union within the framework of European Social Fund

Averaging ROC curves

Example in R



Cumulative gains curves (a.k.a. lift curves)

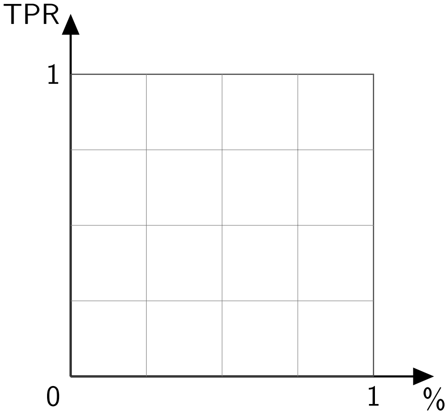
- y-axis = True Positive Rate (TPR)
- x-axis = number (or percentage) of cases classified as $+$ by the model

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI FUNDUSZ SPOŁECZNY

Project co-financed by European Union within the framework of European Social Fund

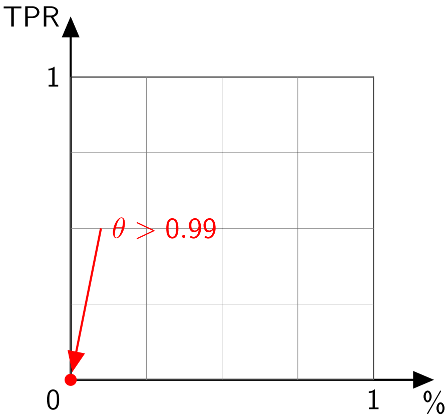
Cumulative gains curves (a.k.a. lift curves) – example



score	Y
0.99	$+$
0.88	$+$
0.71	$-$
0.55	$+$
0.32	$-$
0.15	$+$
0.08	$-$
0.01	$-$

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

Cumulative gains curves (a.k.a. lift curves) – example

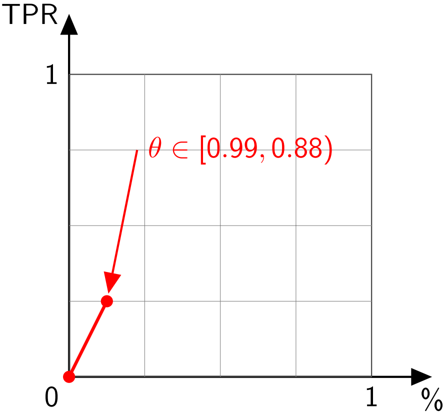


score	Y
0.99	$+$
0.88	$+$
0.71	$-$
0.55	$+$
0.32	$-$
0.15	$+$
0.08	$-$
0.01	$-$

Initial point – no cases classified as $+$

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

Cumulative gains curves (a.k.a. lift curves) – example

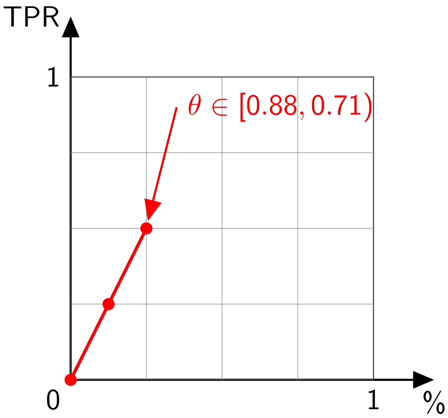


	score	Y
$\rightarrow$	0.99	$+$
	0.88	$+$
	0.71	$-$
	0.55	$+$
	0.32	$-$
	0.15	$+$
	0.08	$-$
	0.01	$-$

First case is positive, move up and left

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

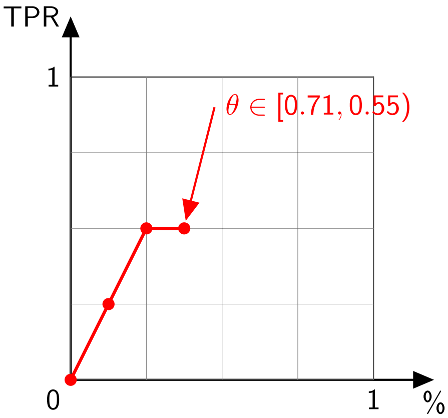
Cumulative gains curves (a.k.a. lift curves) – example



	score	Y
	0.99	+
$\rightarrow$	0.88	+
	0.71	−
	0.55	+
	0.32	−
	0.15	+
	0.08	−
	0.01	−

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

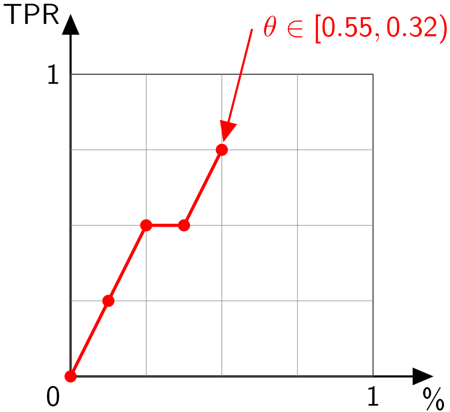
Cumulative gains curves (a.k.a. lift curves) – example



score	Y
0.99	$+$
0.88	$+$
$\rightarrow$ 0.71	$-$
0.55	$+$
0.32	$-$
0.15	$+$
0.08	$-$
0.01	$-$

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

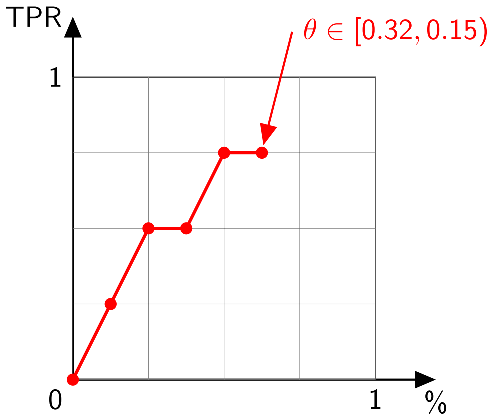
Cumulative gains curves (a.k.a. lift curves) – example



score	Y
0.99	$+$
0.88	$+$
0.71	$-$
$\rightarrow$ 0.55	$+$
0.32	$-$
0.15	$+$
0.08	$-$
0.01	$-$

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

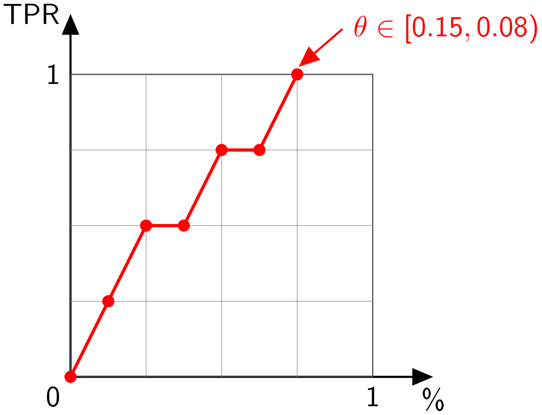
Cumulative gains curves (a.k.a. lift curves) – example



score	Y
0.99	$+$
0.88	$+$
0.71	$-$
0.55	$+$
$\rightarrow$ 0.32	$-$
0.15	$+$
0.08	$-$
0.01	$-$

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

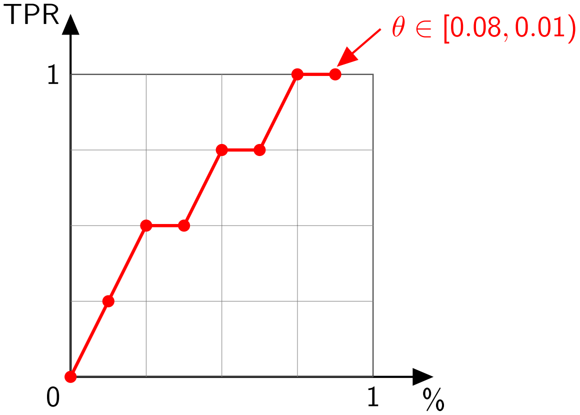
Cumulative gains curves (a.k.a. lift curves) – example



score	Y
0.99	$+$
0.88	$+$
0.71	$-$
0.55	$+$
0.32	$-$
$\rightarrow$ 0.15	$+$
0.08	$-$
0.01	$-$

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

Cumulative gains curves (a.k.a. lift curves) – example



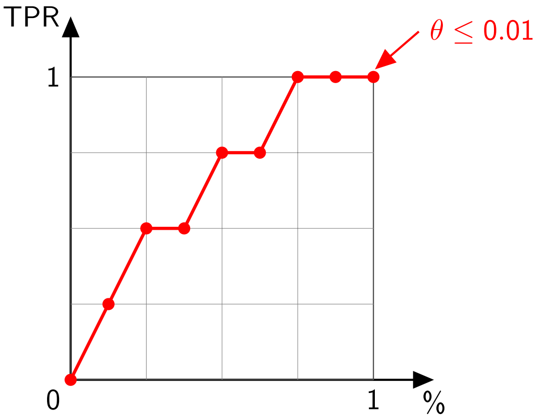
	score	Y
	0.99	$+$
	0.88	$+$
	0.71	$-$
	0.55	$+$
	0.32	$-$
	0.15	$+$
$\rightarrow$	0.08	$-$
	0.01	$-$

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

Project co-financed by European Union within the framework of European Social Fund

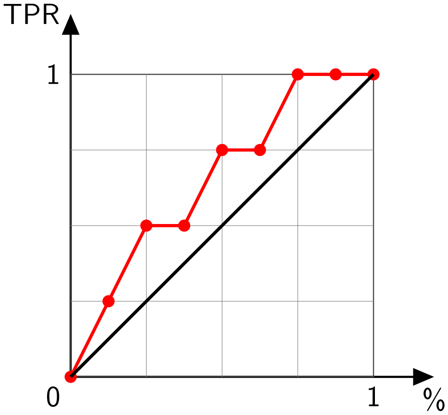
Cumulative gains curves (a.k.a. lift curves) – example



	score	Y
	0.99	$+$
	0.88	$+$
	0.71	$-$
	0.55	$+$
	0.32	$-$
	0.15	$+$
	0.08	$-$
$\rightarrow$	0.01	$-$

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

Cumulative gains curves (a.k.a. lift curves) – example



score	*Y*
0.99	+
0.88	+
0.71	−
0.55	+
0.32	−
0.15	+
0.08	−
0.01	−

Model assigning a class at random corresponds to the diagonal line

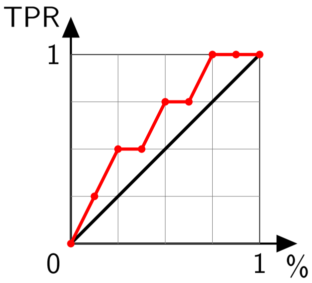
PhD STUDIES

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Project co-financed by European Union within the framework of European Social Fund

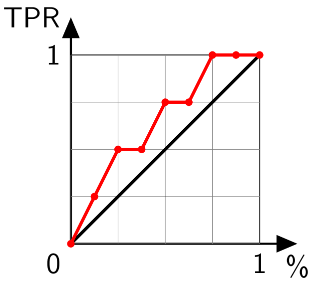
Cumulative gains curves – Interpretation



- The further from the diagonal, the better the model
- Lowering the threshold, more and more cases are classified as positive
- In business settings positive cases are e.g. mailed a marketing offer

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

Cumulative gains curves – Interpretation



- E.g. targeting 25% best customers we catch 50% of those who will respond
- The higher the percentage, the less we gain
- E.g. targeting 75% best customers we catch all good ones, sending more brings no gain

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

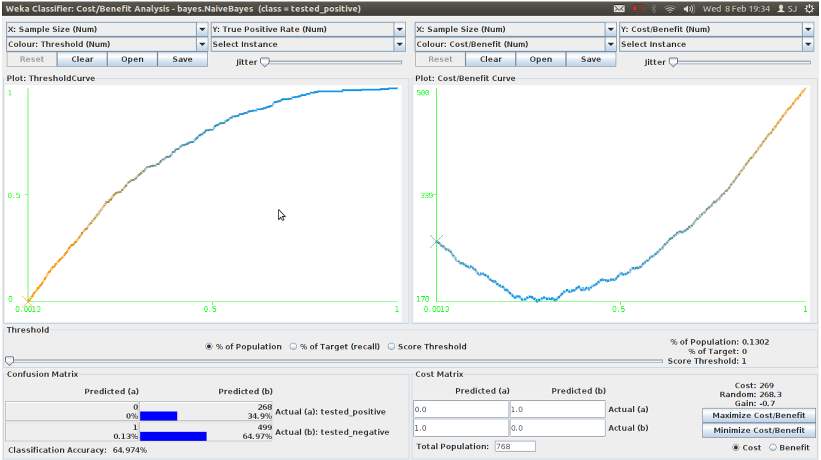
Cumulative gains curves – adding costs

- Real costs / benefits can be taken into account while drawing the curve
- The y-axis shows monetary gain
- The curve is no longer increasing

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

A Cumulative gains curve with and without costs (from Weka)

PIMA-Indians-diabetes dataset



KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Project co-financed by European Union within the framework of European Social Fund

Other types of curves

- Many other types of curves (less frequently used)
- Precision/recall charts
  - y-axis: precision $= P(Y = +|M(\mathbf{x}) = +)$
  - x-axis: recall $=$ TPR
- Lift curves (different from cumulative gains curves)
  - y-axis: lift $= \frac{P(Y=+|M(\mathbf{x})=+)}{P(Y=+)}$
  - x-axis $=$ number (or percentage) of cases classified as $+$ by the model
- Cost curves
- Calibration curves

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

Drawing the curves in R

- The `ROCR` package
- Very powerful:
  - many kinds of statistics can be used for both axes
  - various kinds of averaging, confidence intervals
  - nice plots

Project co-financed by European Union within the framework of European Social Fund

Statistical comparisons of classifiers

- How to reliably compare classifiers?
- On a single dataset:
    - Error bounds (binomial confidence intervals) on the test set
    - Error bounds for crossvalidation (usually not statistically rigorous)

PhD STUDIES

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Project co-financed by European Union within the framework of European Social Fund

Statistical comparisons of classification algorithms

- How to reliably compare classification algorithms?
- (On multiple datasets)

Assumption

The available datasets form a representative sample from a 'population' of all datasets on which the algorithms are to be used

- An approach proposed by Demšar in 2006:
  - the procedure is based on the rankings of algorithms on each dataset
  - any performance measure can be used
    - accuracy
    - AUC
    - ...

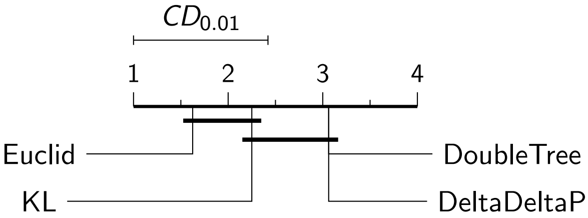
  As long as meaningful comparisons are possible

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

- The approach proposed by Demšar in 2006:
  1. First use the Friedman's test to check if there are any significant differences between the algorithms
     - Friedman's test is a nonparametric (rank based) equivalent of ANOVA
  2. If there are significant differences, use the Nemenyi test to check how the algorithms compare to each other
     - for a given significance level, the Nemenyi test gives a critical difference CD_α
     - if the average ranks of two algorithms differ by more than CD_α, their performance is significantly different

Statistical comparisons of classification algorithms

Differences are best visualized using Critical Difference Plots



- Average rank of each algorithm is marked on the horizontal scale
- Thick lines connect algorithms which are not significantly different
- The size of CD_α is marked on top

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
Project co-financed by European Union within the framework of European Social Fund

Papers to read

1. Tom Fawcett, *ROC Graphs: Notes and Practical Considerations for Researchers*, home.comcast.net/~tom.fawcett/public_html/papers/ROC101.pdf [A thorough introduction to ROC curves]
2. Foster Provost and Tom Fawcett, *Analysis and Visualization of Classifier Performance: Comparison under Imprecise Class and Cost Distributions.* KDD-97 (Third International Conference on Knowledge Discovery and Data Mining) http://home.comcast.net/~tom.fawcett/public_html/papers/KDD-97.pdf [Convex hulls of ROC curves – interesting but not that important]
3. Janez Demšar, *Statistical comparisons of classifiers over multiple data sets*, Journal of Machine Learning Research, 2006(7), pp. 1–30, jmlr.csail.mit.edu/papers/volume7/demsar06a/demsar06a.pdf [statistical comparisons of classifiers]

Project co-financed by European Union within the framework of European Social Fund