Model evaluation

- qualitative – following the definition of data mining (Piatetski-Shapiro, Fayaad, 90th):

  how new, interesting, useful and understandable the model is

  (not) corresponding to expectations (common sense), to knowledge of an expert
- quantitative

Model evaluation

- qualitative – following the definition of data mining (Piatetski-Shapiro, Fayaad, 90th):

  how new, interesting, useful and understandable the model is

  (not) corresponding to expectations (common sense), to knowledge of an expert
- quantitative

Evaluation for different machne learning task

- clustering – is the number of clusters and the structure appropriate
- associations – which rule is interesting
- outlier detection – top N outliers
- classification and regression

Evaluation for different machne learning task

- clustering – is the number of clusters and the structure appropriate
- associations – which rule is interesting
- outlier detection – top N outliers
- classification and regression

Classification

Training set

|

Learning algorithm

|

input atributes of a test instance

– – –> Model/Hypothesis/Classifier – – –>

predicted class label

- accuracy [celková správnost] – how often returns the correct class label
- speed – learning, testing
- robustness – to make correct predictions given noisy data or data with missing values
- scalability – efficient for large amounts of data

Classification

Training set

|

Learning algorithm

|

input atributes of a test instance

– – –> Model/Hypothesis/Classifier – – –>

predicted class label

- accuracy [celková správnost] – how often returns correct class label
- speed – learning, testing
- robustness – to make correct predictions given noisy data or data with missing values
- scalability – efficient for large amounts of data

Classification

main criterion – how succesful Model is on data

a principal decision – what data to use for the most accurate prediction of model accuracy

Most common (but correct?)

- learning data
- test set
- cross-validation
- leave-one-out

Is there any other possibility, maybe better? bootstraping, splitting data into disjunctive parts, ...

Classification

main criterion – how succesful Model is on data

a principal decision – what data to use for the most accurate prediction of model accuracy

Most common (but correct?)

- learning data
- test set
- cross-validation
- leave-one-out

Is there any other possibility, maybe better? bootstraping, splitting data into disjunctive parts, ...

Classification

main criterion – how succesful Model is on data.

a principal decision – what data to use for the most accurate prediction of model accuracy

Most common (but correct?)

- learning data
- test set
- cross-validation
- leave-one-out

Is there any other possibility, maybe better? bootstraping, splitting data into disjunctive parts, ...

Classification

main criterion – how succesful Model is on data.

a principal decision – what data to use for the most accurate prediction of model accuracy

Most common (but correct?)

- learning data
- test set
- cross-validation
- leave-one-out

Is there any other possibility, maybe better? bootstraping, splitting data into disjunctive parts, ...

Classification

main criterion – how succesful Model is on data.

a principal decision – what data to use for the most accurate prediction of model accuracy

Most common (but correct?)

- learning data
- test set
- cross-validation
- leave-one-out

Is there any other possibility, maybe better? bootstraping, splitting data into disjunctive parts, ...

Classification

main criterion – how succesful Model is on data.

a principal decision – what data to use for the most accurate prediction of model accuracy

Most common (but correct?)

- learning data
- test set
- cross-validation
- leave-one-out

Is there any other possibility, maybe better? bootstraping, splitting data into disjunctive parts, ...

Confusion matrix

Predicted class

Actual class		*yes*	*no*	Total
	yes	TP	FN	P
	no	FP	TN	N
	Total	P'	N'	$P+N$

TP, TN, FP, FN ... the number of true positive, true negative, false positive, false negative

P, N ... cardinality of positive and negative samples

Evaluation measures

(overall) accuracy [celková správnost]

$$Acc = \frac{TP+TN}{TP+TN+FP+FN}$$

error rate, (misclassification rate) [chyba]

$$Err = 1 - Acc = \frac{w_{FP} * FP + w_{FN} * FN}{TP+TN+FP+FN}$$

w_{FP}, w_{FN} ... weight of FP and FN errors

default $w_{FP}, w_{FN} = 1$

precision

$$\frac{TP}{TP+FP}$$

sensitivity, true positive rate, recall

$$\frac{TP}{TP+FN}$$

specificity, true negative rate

$$\frac{TN}{TN+FP}$$

Evaluation measures

(overall) accuracy [celková správnost]

$$Acc = \frac{TP+TN}{TP+TN+FP+FN}$$

error rate, (misclassification rate) [chyba]

$$Err = 1 - Acc = \frac{w_{FP} * FP + w_{FN} * FN}{TP+TN+FP+FN}$$

w_{FP}, w_{FN} ... weight of FP and FN errors

default $w_{FP}, w_{FN} = 1$

precision

$$\frac{TP}{TP+FP}$$

sensitivity, true positive rate, recall

$$\frac{TP}{TP+FN}$$

specificity, true negative rate

$$\frac{TN}{TN+FP}$$

Evaluation measures

Accuracy for a class P, N

F-measures combines precision and recall

F, F1, F-score = hramonic mean of precision and recall

$$F_1 = \frac{2 * precision * recall}{precision + recall}$$

$$F_\beta = \frac{(1+\beta^2) precision * recall}{\beta^2 * precision + recall}$$

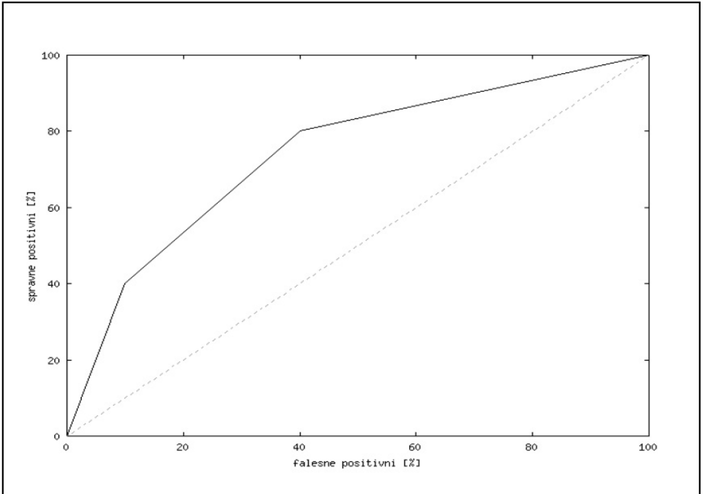
β ... a non-negative real number

Evaluation measures for comparing classifiers

Learning curve

Accuracy as a function of number of iterations

ROC curve – relation between TP and FP



Sampling

- holdout – split data randomly to learning and test data, e.g. 2/3 vs. 1/3
  stratified sampling – preserve relative frequency of classes in samples
- Random (sub)sampling – holdout method is repeated k times
  The overall accuracy estimate is taken as the average of the accuracies obtained from each iteration.
- bootstraping
- undersampling/oversampling of a class – for processing imbalanced data