A Gentle Introduction to Machine Learning in Natural Language Processing using R

ESSLLI '2013
Düsseldorf, Germany

http://ufal.mff.cuni.cz/mlnlpr13

Barbora Hladká
hladka@ufal.mff.cuni.cz

Martin Holub
holub@ufal.mff.cuni.cz

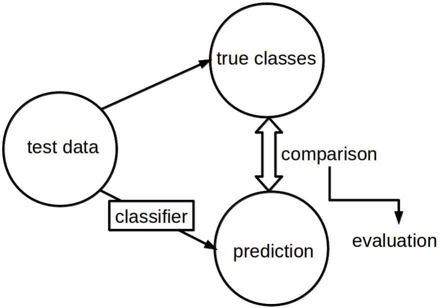
Charles University in Prague,
Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics

Day 5

- 5.1 Cross-validation and confidence intervals
- 5.2 13 points you cannot miss on the way to ML
- 5.3 Overview of the course

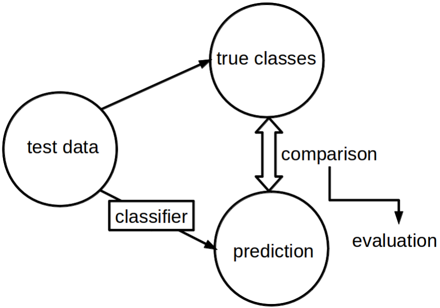
Block 5.1
Cross-validation and confidence intervals

The evaluation process



Block 5.1
Cross-validation and confidence intervals

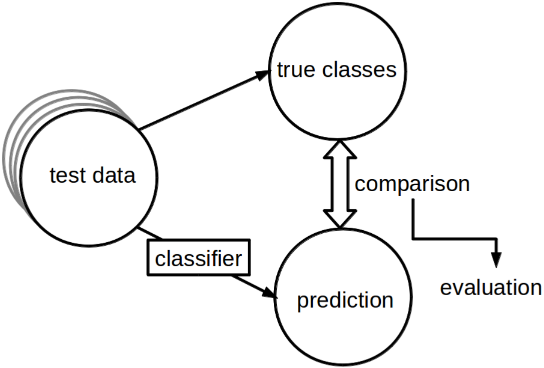
The evaluation process



Is it enough to test your classifier on one test set?
You can get a good/bad result by chance!

The ideal evaluation

The more test data, the more confident evaluation . . .

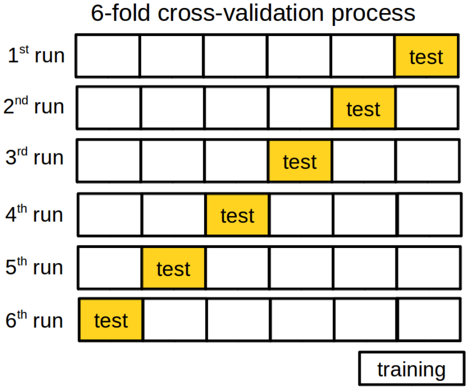


k-fold cross-validation

Development working data is partitioned into k **subsets** of equal size. Then you do k iterations.

In the i-th step of the iteration, the i-th subset is used as a test set, while the remaining parts form the training set.

Example



Using t-tests

When you get k different results from the cross-validation experiment, what can you conclude then?

1. **One Sample t-test**
   to test if the mean of a (normally distributed) population is equal to a given value

2. **Paired Two-Sample t-test**
   to test if the difference of the means of two populations is equal to a given value, assuming that the given sample contains paired individuals

Example: one sample t-test

You have two models, **A** and **B**, and for each of them 10 results – accuracies obtained from 10-fold cross-validation experiment.

```
> A.acc
 [1] 0.853 0.859 0.863 0.871 0.832 0.848 0.863 0.860 0.850 0.849
> mean(A.acc)
 [1] 0.8548

> B.acc
 [1] 0.851 0.848 0.862 0.871 0.835 0.836 0.860 0.859 0.841 0.843
> mean(B.acc)
 [1] 0.8506
```

The average accuracy of **A** is 85.48 %, while the average accuracy of **B** is only 85.06 %.

Is the model A *really* better than the model B?

To test if the difference between the models **A** and **B** is **statistically significant** we will check **confidence intervals** for the mean accuracy.

```
### Could the true mean of A accuracy be 0.8506?
> t.test(A.acc, mu=0.8506)
    One Sample t-test

data:  A.acc
t = 1.2195, df = 9, p-value = 0.2537
alternative hypothesis: true mean is not equal to 0.8506
95 percent confidence interval:
 0.8470088 0.8625912
sample estimates:
mean of x
   0.8548
```

We cannot reject the null hypothesis that the mean of A accuracy is equal to 0.8506. The t-test says that the true mean of A accuracy could be between 0.8470088 and 0.8625912, which is the confidence interval at the significance level $\alpha = 5\,\%$.

Example: two-sample t-test

```
### Could the true mean of the difference be equal to zero?
> t.test(A.acc, B.acc)
    Welch Two Sample t-test

data:  A.acc and B.acc
t = 0.8157, df = 17.803, p-value = 0.4254
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.006625999  0.015025999
sample estimates:
mean of x mean of y
   0.8548    0.8506
```

We cannot reject the null hypothesis that the mean of the difference between A accuracy and B accuracy is equal to 0.
The t-test says that the true mean of the difference could be between -0.006625999 and 0.015025999, which is the confidence interval at the significance level $\alpha = 5\,\%$.

Block 5.2
13 points you cannot miss on the way to ML

Task and data management

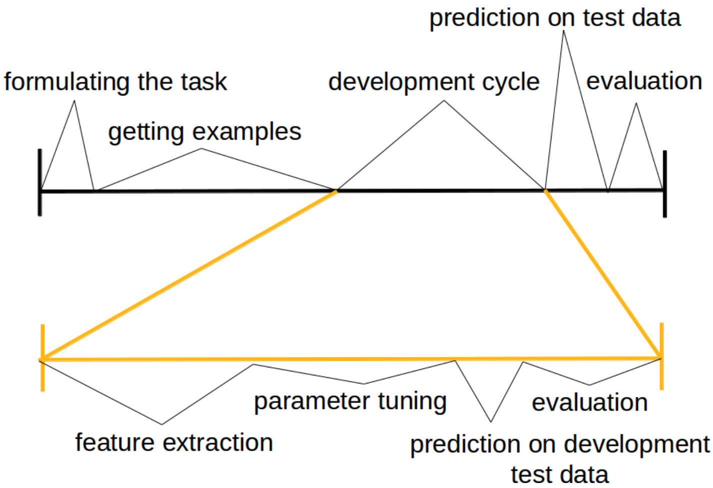
1. Time management
2. Formulating the task
3. Getting data
4. The more data, the better
5. Feature engineering
6. Curse of dimensionality

Methods and evaluation

7. Learning algorithms
8. Development cycle
9. Evaluation
10. Optimizing learning parameters
11. Overfitting
12. The more classifiers, the better
13. Theoretical aspects of ML

(1) Time management

How much time do particular steps take?



(2) Formulating the task

- Precise formulation of the task
- What are the objects of the task?
- What are the target values of the task?

(3) Getting data

- Gather data
- Assign true classification
- Clean it
- Preprocess it

(4) The more data, the better

If we don't have enough data

- **cross-validation** – The data set *Data* is partitioned into subsets of equal size. In the i-th step of the iteration, the i-th subset is used as a test set, while the remaining parts from the training set.
- **bootstrapping** – New data sets $Data_1$, ..., $Data_k$ are drawn from *Data* with replacement, each of the same size as *Data*. In the i-th iteration, $Data_i$ forms the training set, the remaining examples in *Data* form the test set

(5) Feature engineering

- Understand the properties of the classified objects
  - How they interact with the target class
  - How they interact each other
  - How they interact with a given ML algorithm
  - Domain specific
- Feature selection manually
- Feature selection automatically: generate large number of features and then filter some of them out

(6) Curse of dimensionality

- A lot of features $\longrightarrow$ high dimensional spaces
- The more features, the more difficult to extract useful information
- Dimensionality increases $\longrightarrow$ predictive power of classifier reduces
- The more features, the harder to train a classifier
- **Remedy**: feature selection, dimensionality reduction

(7) Learning algorithms

Which one to choose?

First, identify appropriate learning paradigm

- Classification? Regression?
- Supervised? Unsupervised? Mix?
- If classification, are class proportions even or skewed?

In general, **no learning algorithm dominates all others on all problems**.

(8) Development cycle

- Test developer's expectation
- What does it work and what doesn't?

(9) Evaluation

Model assessment

- **Metrics** and **methods** for performance evaluation
  How to evaluate the performance of a classifier? How to obtain reliable estimates?
- **Classifier comparison**
  How to compare the relative performance among competing classifiers?
- **Classifier selection**
  Which classifier should we prefer?

(10) Optimizing learning parameters

Searching for the best classifier, i.e.

- adapting ML algorithms to the particulars of a training set
- optimizing classifier performance

Optimization techniques

- Greedy search
- Beam search
- Grid search
- Gradient descent
- Quadratic programming
- ...

(11) Overfitting

To avoid it using

- cross-validation
- feature engineering
- parameter tuning
- regularization – a standard method to penalize classifiers with more complex structure

(12) The more classifiers, the better

- **Build an ensemble of classifiers** using
  - different learning algorithm
  - different training data
  - different features
- **Analyze** their performance: complementarity implies potential improvement
- **Combine** classification results (e.g. majority voting).

Examples of ensemble techniques

- **bagging** works by taking a bootstrap sample from the training set
- **boosting** works by changing weights on the training set

(13) Theoretical aspects

Computational learning theory aims to understand fundamental issues in the learning process. Mainly the issues on

- How computationally hard is the learning problem?
- How much data do we need to be confident that good performance on that data really means something?

Block 5.3
Overview of the course

Day 1

- 1.1 Relation between NLP and ML
- 1.2 Course outline
- 1.3 Non-technical view on ML
- 1.4 Dealing with data
- 1.5 Intro to R
- Summary

Day 2

- 2.1 A few necessary R functions
- 2.2 Mathematics
- 2.3 Decision tree learning – Theory
- 2.4 Decision tree learning – Practice
- Summary

Day 3

- 3.1 Formal foundations of ML
- 3.2 Naive Bayes learning – Theory
- 3.3 Naive Bayes learning – Practice
- 3.4 Evaluation of a classifier
- Summary

Day 4

- 4.1 Information Theory and Feature Selection
- 4.2 SVM learning – Theory
- 4.3 SVM learning – Practice

Day 5

- 5.1 Cross-validation and confidence intervals
- 5.2 13 points you cannot miss on the way to ML
- 5.3 Overview of the course

Experiments

COL task

Features	Algorithm	Accuracy (%)
$A_1, ..., A_{11}$		
	DT	87.8
	NB	85.3
	SVM (kernel=linear, cost=10)	86.8
	SVM (kernel=linear, cost=100)	86.8
$A_1, ..., A_{10}$		
	DT	85.6
	NB	85.6
	SVM (kernel=linear, cost=10)	85.4
	SVM (kernel=linear, cost=100)	85.4

The course is over

You are at the very beginning . . . Good luck!!!