Lecture 5

Spring 2018

Faglærer: Magne Hallstein Johnsen,

Institutt for elektronikk og telekommunikasjon, NTNU

NTNU

Lecture content

- How to evaluate classifiers
- The importance of generalization
- Evaluating the found/empirical error rate using confidence
- Different classifiers and significance
- Leave-one-out technique for small data sets
- The confusion matrix

How to evaluate classifiers

- Both the BDR classifier and the corresponding **minimum** error rate are unknown for any problem

- Even the **true** error rate for any chosen/implemented classifier is unknown

- Given a finite labeled test set :
  - The **empirical/estimated** error rate EER for this test set can be found by just counting the errors
  - However **another test set** (even of same size) will give another EER!
  - How to estimate the true error rate based on a **single and finite** test set...?

- The same problems arise if we train the same classifier with two **different** training sets of same size.

A good classifier will generalize well

- Assume finite labeled development/train and test sets of size N and M respectively
- Calculate the empirical error rate EER_T for the test set for a chosen classifier.
- The big question : is this EER_T far from the unknown/true error rate TER for the classifier?
- We apply a "suboptimal" strategy for the question:
  - Calculate the empirical error rate EER_D for the **development** set
  - If the difference $EER_T - EER_D$ is small , we assume the same applies to $EER_T - TER$
- We call the above property for a **generalization** ability
- The above generalization is not valid if the data set sizes N and/or M are small.

Evaluating the found/empirical error rate using confidence

- The empirical error rate is equal to the true error rate only when the test set size $M = \infty$
- Thus the difference $EER_T - TER$ should decrease as M increases.
- But still two different test sets of same size will give different error rates.
- This leads to the so called confidence strategy for the error rate :
  - For a given M calculate an error rate interval centered on the empirical error rate
  - With 95% confidence the true error rate should be within the interval.
  - The interval is a function of EER_T and M
  - The interval decreases as the test set size M increases

Different classifiers and significance

- Assume two different classifier structures
- The same training set is used for both classifiers
- The same test set is used for both classifiers resulting in two error rates EER_{T1} and EER_{T2}.
- Assume $EER_{T1} < EER_{T2}$, can we claim the first structure is better than the other?
- Again we should use the 95% confidence strategy :
  - Calculate the interval centered on the lowest error rate EER_{T1}
  - If EER_{T2} is outside the interval we can say that the difference is **significant**, i.e. the first classifier is better.
  - If EER_{T2} is inside the interval other train/test sets can be used to confirm or change the evaluation result

The Leave-one-out technique

- Sometimes data is hard to aquire. Thus splitting the data of size R into train and test parts is difficult.
- Too small training and/or test sets are not representative.
- The Leave-one-out technique can to some extent compensate for this.
- For $i = 1, \ldots, R$
  - Use all R except sample number i as training set
  - Train the classifier
  - Test the classifier with the single sample number i
- Use the mean of the test sample results as EER_T.

The confusion matrix

- Given $C > 1$ classes a test sample from class ω_i is classified as ω_j
- The EER only tells us how often a sample is misclassified
- Often the class informations will be of interest as well, i.e. which pair of classes that are most/least confusable
- Elements in a confusion matrix $\mathbf{A}$ will hold this information, i.e. $A[i, j]$ shows the number of times the classiifer claims $x \in \omega_i$ while true is $x \in \omega_j$
- The confusion matrix will give the most confusable classes, which then can be further investigated for improvement
- The confusion matrixes can show significant differences from classifier to classifier. This lead to fusioning of different classifiers outcomes.