Evaluation of Learning Models
Overview

• Motivation

• Metrics for Classifier’s Evaluation

• Methods for Classifier’s Evaluation
  • Holdout
  • K-fold
  • Stratification
  • Leave-one-out
Motivation

- It is important to evaluate classifier’s generalization performance in order to:
  - Determine whether to employ the classifier
  - Compare classifiers
  - Optimize the classifier
Metrics for Classifier’s Evaluation

- **Accuracy** = \( \frac{TP+TN}{P+N} \)
- **Error** = \( \frac{FP+FN}{P+N} \)
- **Precision** = \( \frac{TP}{TP+FP} \)
- **Recall/TP rate** = \( \frac{TP}{P} \)
- **FP rate** = \( \frac{FP}{N} \)
## Metrics for Classifier’s Evaluation

<table>
<thead>
<tr>
<th>Actual class</th>
<th>Predicted class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pos</td>
<td>Pos</td>
</tr>
<tr>
<td>Pos</td>
<td>TP</td>
</tr>
<tr>
<td>Neg</td>
<td>FP</td>
</tr>
</tbody>
</table>
The $F_1$ score is the harmonic average of the precision and recall:

$$F_1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$
How to Estimate the Metrics?

• We can use:
  • Training data;
  • Independent test data;
  • Hold-out method;
  • $k$-fold cross-validation method;
  • Leave-one-out method;
  • Bootstrap method;
  • And many more…
The accuracy/error estimates on the training data are not good indicators of performance on future data.

Q: Why?
A: Because new data will probably not be exactly the same as the training data!

The accuracy/error estimates on the training data measure the degree of classifier’s under-/over-fitting.
Estimation with Independent test data

- Estimation with independent test data is used when we have plenty of data and there is a natural way to forming training and test data.
Hold-out Method

- The hold-out method splits the data into training data and test data. Then we build a classifier using the train data and test it using the test data.

- The hold-out method is usually used when we have thousands of instances, including several hundred instances from each class.
The test data can’t be used for parameter tuning!
Making the Most of the Data

- Once evaluation is complete, *all the data* can be used to build the final classifier.

- Generally, the **larger the training data** the **better the classifier** (but returns diminish).

- The **larger the test data** the **more accurate** the error estimate.
Stratification

• The *holdout* method reserves a certain amount for testing and uses the remainder for training.

• For “unbalanced” datasets, samples might not be representative.
  • *Few or none instances of some classes.*

• **Stratified sample**: balancing the data.
  • *Make sure that each class is represented with approximately equal proportions in both subsets.*
**k-Fold Cross-Validation**

- **k-fold cross-validation**:  
  - *First step*: data is split into $k$ subsets of equal size;  
  - *Second step*: each subset in turn is used for testing and the remainder for training.
- The estimates are averaged to yield an overall estimate.
More on Cross-Validation

• Standard method for evaluation: stratified 10-fold cross-validation.

• Stratification reduces the estimate’s variance.

• Even better: repeated stratified cross-validation:
  • E.g. ten-fold cross-validation is repeated ten times and results are averaged (reduces the variance).
K-folds from what?

- Which dataset should we do k-fold on? Training? Training+validation? Everything?

- The *test set* should be put aside and it cannot be used for ANY purpose (including data prep).

- The *validation set* is used to test performances of the models at the various stages of the model development.
  - Different models, various versions of the same model, etc.
  - A good practice is to the use the validation set as little as possible (to avoid "mental" overfit)

- *Train set* is used for training. This includes cross validation.
Leave-One-Out Cross-Validation

- Leave-One-Out is a particular form of cross-validation:
  - Set number of folds to number of training instances;
  - I.e., for \( n \) training instances, build classifier \( n \) times.

- Makes best use of the data.

- Involves no random sub-sampling.

- Very computationally expensive.
Leave-One-Out Cross-Validation and Stratification

- A disadvantage of Leave-One-Out-CV is that stratification is not possible:
  - It guarantees a non-stratified sample because there is only one instance in the test set!

- Extreme example - random dataset split equally into two classes:
  - Best inducer predicts majority class;
  - 50% accuracy on fresh data;
  - Leave-One-Out-CV estimate is 100% error!
Pre-processing steps on validation set?

- Whatever data prep steps done on training data, should be able to be applied (not redesign) at the final test on unseen data.
  - For example, if normalizing a feature with a mean, the mean is fixed at the data prep and not recalculated. Otherwise, a model that was trained and tested on data that used the former mean may not be adequate.

Q: So should we use the validation set for pre-processing?
A: There are quite a few opinions and no consensus. Both options are valid and the decision is mainly based on the stability of the statistics (measures) that can be extracted from Train vs. Train+Validation.

- Data preparation steps are mostly based on descriptive statistics. So a larger sample that provides a stable statistics is an advantage.
- By not using the validation set, we have a better estimate of the “true error”, which we measure by applying pre-processing steps to test data.
- By using the validation set, we separate the effect of the pre-processing from that of the classifier.