Remarks on Evaluation

Outline

• Basics of classifier evaluation
  – why we need evaluation
  – working with data
  – sample error and generalization error

• Overfitting
You need thorough evaluation to

1. get a reliable **estimate of the classifier performance**
   - i.e. how it will perform on new – so far unseen – data instances
   - possibly even in the future

2. compare your different classifiers that you have developed
   - to decide which one is “the best”

= **Model assessment and selection**
You need *good* performance
not only on *your* data,
but also on any data that can be *expected*!
All subsets should be selected randomly in order to represent the characteristic distribution of both feature values and target values in the available set of examples.
Evaluation – basic scheme

Diagram:
- Test data
- Classifier
- True classes
- Comparison
- Prediction
- Evaluation
Development working data

Is used both for training your classifier and for evaluation when you tune the learning parameters.

- **Training data**
  is used for *training* your classifier with a particular learning parameter settings when you tune your classifier.

- **Held-out data**
  is used for *evaluating* your classifier with a particular learning parameter settings when you tune your classifier.
Development test set

- the purpose is to simulate the “real” test data
- should be used only for your final development evaluation when your classifier has already been tuned and your learning parameters are finally set
- using it you get an estimate of your classifier’s performance at the end of the development
- is also used for model selection
Sample accuracy and sample error rate

To measure the performance of classification tasks we often use (sample) accuracy and (sample) error rate

**Sample accuracy** is the number of correctly predicted examples divided by the number of all examples in the predicted set

**Sample error rate** is equal to 1 - accuracy

**Training error rate** is the sample error rate measured on the training data set

**Test error rate** is the sample error rate measured on the test data set
Sample error and generalization error

**Sample error** of a hypothesis \( h \) with respect to a data sample \( S \) of the size \( n \) is usually measured as follows

- for **regression**: *mean squared error* \( \text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \)

- for **classification**: *classification error* \( = \frac{1}{n} \sum_{i=1}^{n} I(\hat{y}_i \neq y_i) \)

**Generalization error** (aka “true error” or “expected error”) measures how well a hypothesis \( h \) generalizes beyond the used training data set, to unseen data with distribution \( \mathcal{D} \). Usually it is defined as follows

- for **regression**: \( \text{error}_\mathcal{D}(h) = E (\hat{y}_i - y_i)^2 \)
- for **classification**: \( \text{error}_\mathcal{D}(h) = \text{Pr} (\hat{y}_i \neq y_i) \)
Finding a model that minimizes generalization error

... is one of central goals of the machine learning process