Experimental methodology

INF5830
Fall 2015

Machine learning experiments

Definition: A computer program is said to **learn** from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E (Tom M. Mitchell: “Machine Learning”)

Machine learning experiments

Examples:

- Word Sense Disambiguation:
  - Task T: assigning the correct sense to ambiguous words
  - Performance measure P: percentage of correctly assigned word-sense pairs
  - Training experience E: corpus of words with correct sense
- Transition-based dependency parsing:
  - Task T: assigning a parser action to parser configurations
  - Performance measure P: Labeled Accuracy Score
  - Training experience E: treebank (transformed to parser configurations paired with parser actions)

Designing a learning system

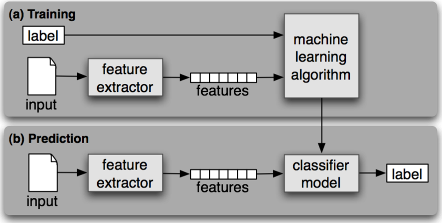
- Choosing the learning experience (data):
  - choice of labels or formal representations
  - representative for task
- Choosing the target function:
  - WSD: w $\rightarrow$ s
  - DepPars: config $\rightarrow$ action
- Choosing a representation for the input (features)
  - w $= [w_1, w_{i-1}, pos(w_1), pos(w_{i-1}) \ldots]$
  - config $= [w(S_1), w(I_1), pos(S_1), pos(I_1), etc \ldots]$
- Choosing a machine learning algorithm

NLP experiments

- Task/Problem: a function mapping inputs to outputs
- Data: instances mapping individual inputs to particular outputs
- Representation: representation for the task (target function)
- Acquisition: learn a model
- Evaluation: how does the acquired model perform?

Acquisition

- Acquisition deals with learning a model which approximates the target function
- choice of algorithm which optimizes the mapping from inputs to outputs
  - use data containing inputs for the task to be solved



(NLTK book)

Evaluation

- Internal evaluation: compare **accuracy** of model output to gold standard
- External evaluation (task-based evaluation):
  - quantify whether model output improves performance on a dependent task

Evaluation

- Precision: how accurate is the model?

$$P = \frac{tp}{tp + fp}$$

- Recall: how good is its coverage?

$$R = \frac{tp}{tp + fn}$$

- F-score: combined measure

$$F = \frac{2 \times P \times R}{P + R}$$

Evaluation: data-driven dependency parsing

evaluation scores:

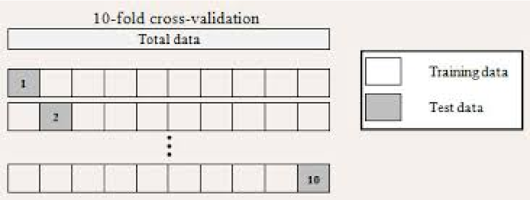
- *Attachment score*
  percentage of words that have the correct head (and label)
- For single dependency types (labels):
  - *Precision*
  - *Recall*
  - *F measure*

Experimental conditions

- Have data with inputs and outputs
- Need data to train, develop and evaluate models
- Many machine learning methods have additional parameters which may need to be optimized
- Fundamental rule: **never evaluate your model on data which it has seen**
  - part of the data must be held out for testing
- Also: do not develop your model by tracking performance on the test data
  - part of the data must be held out for development
  - used for setting various parameters of machine learning algorithms
- Use the rest of the material for training

Experimental conditions

- Very little data: *n*-fold cross-validation
  - vary the training and testing material
  - usually done with 10 folds



Experimental conditions

- Comparing systems
  - need to have similar conditions
  - ideally compare on the same test set
  - significance testing
- Common to establish a **baseline**: performance of a simple system
  - most frequent label
  - system with default settings
  - system without some special feature

Experimental conditions

Example: data-driven dependency parsing

- compare systems on the same test set
- default system compared to system obtained by
  - varying parsing algorithm
  - varying machine learning algorithm
  - varying feature model

Experimental conditions

Example: data-driven dependency parsing

- statistical significance: Bikel's randomized parsing evaluator (`compare.pl`)
  - given null hypothesis of no difference between two sets of results
  - shuffling should produce a difference equal to or greater than original
  - if the two sets differ significantly shuffling should rarely result in greater difference
  - shuffling repeated 10,000 times, total number of differences equal or larger recorded
  - relative frequency = significance of difference
- differences taken to be significant if $p<0.05$.