Metrics

Lecture 7

Doug Downey, Northwestern EECS 395/495

Winter 2010

What's a "good" probabilistic model?

- Say we've learned a probabilistic model
  - Maybe we wrote it down (as in HW2) or learned it from data (HW3)
- How do we decide if our model is good?

Measuring Model Performance

- Two steps:
  - Obtain model probabilities on example data
    - Test-set data not used in training
  - Compute a **metric** of model performance
- Which metric?
  - Depends on the task

- And others -- depends on the task

Example Metrics (1 of 3)

- Log Likelihood
  - For test set T= {(***y***, ***e***)} and model M
    LL(T, M) = $\Sigma_T \log P_M(\boldsymbol{y} \mid \boldsymbol{e})$
  - Higher is better
  - Somewhat "uninterpretable"
    - If I tell you average log likelihood improved from -1.4 to -1.2, how happy will you be?
- More interpretable: accuracy of MAP output
  - Acc(T, M) = (Σ_T 1 if ***y*** = arg max$_{\boldsymbol{y'}}$ $P_M(\boldsymbol{y'} \mid \boldsymbol{e})$) / |T|

Example Metrics (2 of 3)

- But what about the following tasks:
  - Is a randomly drawn Web document about baseball?
  - Is record M in DB1 the same as record M' in DB2?

- …simply answering "no" all the time ensures high (99%+) accuracy.

Example Metrics (3 of 3)

- Precision and Recall, in terms of:
  - # positives(M) = number of test cases where MAP output is $y = 1$
  - # true positives = number of positives where test case output is really $y = 1$
  - # positives(T) = number of test cases with $y=1$
- Precision = #true positives / #positives(M)
- Recall = #true positives / # positives(T)
- F1 = Harmonic Mean(Precision, Recall)

More Interesting Cases

- Example: stock market prediction

Optimizing on Training Set

- Cross-validation
  - Partition data set into *k* pieces (="folds")
  - For each piece *p*
    - train on all pieces but *p,* test on *p*
- Homework: use 2-fold cross validation on training set
  - How well will this predict your test set performance at the end of the quarter?