Notes on the Final Project and Evaluating Methods

David M. Blei

March 14, 2012

The final project

- We would like you to work together in groups of two or three.

- Most final projects will be a piece of applied data analysis. For example, this is one way to execute a good final project.
  - Consider a problem to solve. This problem might be *predictive* (like classification or recommendation) or *descriptive* (like exploring a clustering).
  - Identify a data set. (We'll provide links for where to look.)
  - Consider an algorithm (or a couple) that make sense for solving the problem.
  - Study the algorithm on your chosen problem
    * Tune the knobs with train/test/validation (see below)
    * Consider why you see the behavior you do
    * *Analyze and explain individual errors*
  - Discuss what you learned about the data and the algorithm

- The project proposal is due on April 5. This is a two page document, one per group. It has four sections.
  1. **Problem** What is the problem that you are working on? Why is it important or interesting? For example, "We are going to predict what movies a user likes based on the other movies she has seen."
  2. **Data** What data are you going to analyze? How is it relevant to solving the problem that you proposed to work on?
  3. **Methods** What method or methods will you study? How do they relate to the data and the problem?
  4. **Evaluation** How do you plan to evaluate your method? What are the natural measures of evaluation? What do you expect to find?

  And, please include citation to relevant papers and books.

Evaluation

How do we evaluate the methods we are studying?

We divide our data into two sets:

- Training
- Testing

The training data is what we fit our models to. The testing data is what we use to evaluate held-out performance.

It is not so simple. Most algorithms—k-means, mixtures, classification models—have parameters that we cannot fit with MLE. For example, consider the number of components in a mixture or the smoothing parameter in the text classification model. Algorithms have these parameters too, such as the convergence criterion of EM.

The validation set can be used as a way to calibrate these settings, without "cheating." That is, at the end of the study we report the sensitivity on the validation set and then the final results on the test set.

An alternative is to use cross validation. Create many validation sets (called folds), covering the whole collection. For each fold, predict the data in that fold from a model fit to the out-of-fold data. This gives us a prediction for every data point in the collection. Evaluate that prediction.

It is still a good idea to have a true heldout test set—cross validation can be used to calibrate the parameters. Then fit to the whole training set and report final performance on the test set.