CSE 515: Statistical Methods in Computer Science

Homework #1

Due at noon on April 15, 2009

Guidelines: You can brainstorm with others, but please solve the problems and write up the answers by yourself. You may use textbooks (Koller & Friedman, Russel & Norvig, etc.), lecture notes, and standard programming references (e.g., online Java API documentation). Please do NOT use any other resources or references (e.g., example code, online problem solutions, etc.) without asking.

Submission instructions: Submit this assignment by email to Daniel Lowd `<lowd@cs>`. Attachments should include: A PDF containing written answers; source code for the mixture model; and a README explaining how to compile and run the source code under Linux (e.g., tricycle) or Windows (e.g., aqua).

1. Suppose that 0.001% of the U.S. population are wanna-be terrorists, 90% of them are on the no-fly list, and that 0.1% of the overall U.S. population is on the no-fly list. Under these assumptions, what fraction of people on the no-fly list are wanna-be terrorists?

2. Koller & Friedman, Exercise 2.9. (When there is insufficient information to compute the conditional probability, prove the insufficiency with a counterexample.)

3. Consider a biased die. Given a prior distribution and a series of rolls, we want to estimate the probability of seeing each of the six faces on the next roll. Assuming that the rolls are independent, this can be modeled as a categorical distribution with six parameters, one for each face:

   $$\{\theta_1, \ldots, \theta_6\}; \theta_i \geq 0; \sum_i \theta = 1$$

   Suppose the prior distribution is defined as Dirichlet$(1, 1, 1, 1, 1, 1)$.

   (a) According to this prior, what is the expected value of each θ_i?

   (b) According to this prior, what is the most likely set of parameter values?

   (c) Suppose that, in 30 rolls of the die, you see 1 one, 2 twos, 3 threes, 4 fours, 5 fives, and 6 sixes. According to the given prior and Bayesian statistics, describe the updated belief density function, $P(\{\theta_1, \ldots, \theta_6\} |$ 1 one, $\ldots$, 6 sixes)

   (d) According to the updated beliefs, what is the expected value of each θ_i?

   (e) According to the updated beliefs, what is the most likely set of parameter values?

   (f) According to the updated beliefs, how likely is a uniform distribution (i.e., $\theta_i = 1/6$) relative to the most likely parameter values? (Just report the ratio of the probability density functions for these two sets of parameters.)

4. **Programming project.** Implement the EM algorithm for mixtures of Gaussians in your choice of programming language. (C, C++, Java, Perl, Python, and OCaml are all fine. Please ask about any others.) Assume that means, covariances, and cluster priors are all unknown. For simplicity, you can assume that covariance matrices are diagonal (i.e., all you need to estimate is the variance of each variable). Initialize the cluster priors to a uniform distribution and the standard deviations to a fixed fraction of the range of each variable. Your algorithm should run until the relative change in the log likelihood of the training data falls below some threshold (e.g., stop when log likelihood improves by $< 0.1\%$). The program should be run on the command line with the following arguments:

   ```
   ./gaussmix <# of clusters> <data file> <model file>
   ```

   It should read in data files in the following format:

   ```
   <# of examples> <# of features>
   <ex.1, feature 1> <ex.1, feature 2> ... <ex.1, feature n>
   <ex.2, feature 1> <ex.2, feature 2> ... <ex.2, feature n>
   ...
   ```

   And output a model file in the following format:

   ```
   <# of clusters> <# of features>
   <clust1.prior> <clust1.mean1> <clust1.mean2> ... <clust1.var1> ...
   <clust2.prior> <clust2.mean1> <clust2.mean2> ... <clust2.var1> ...
   ...
   ```

   Train and evaluate your model on the Wine dataset, available from the course Web page. Each data point represents a wine, with features representing chemical characteristics including alchohol content, color intensity, hue, etc. We provide a single default train/test split with the class removed to test generalization. You can find the full dataset and more information in the UCI repository (and linked from the course Web page). Start by using 3 clusters, since the Wine dataset has three different classes. Evaluate your models on the test data.

   Two recommendations:

   - To avoid underflows, work with logs of probabilities, not probabilities.
   - To compute the log of a sum of exponentials, use the "log-sum-exp" trick:

   $$\log \sum_i \exp(x_i) = x_{max} + \log \sum_i \exp(x_i - x_{max})$$

   Answer the following questions with both numerical results and discussion.

   (a) Plot train and test set likelihood vs. iteration. How many iterations does EM take to converge?

(b) Experiment with two different methods for initializing the mean of each Gaussian in each cluster: random values (e.g., uniformly distributed from some reasonable range) and random examples (i.e., for each cluster, pick a random training example and use its feature values as the means for that cluster). Does one method work better than the other or do the two work approximately the same? Why do you think this is? (Use whichever version works best for the remaining questions.)

(c) Run the algorithm 10 times with different random seeds. How much does the log likelihood change from run to run?

(d) Infer the most likely cluster for each point in the training data. How does the true clustering (see `wine-true.data`) compare to yours?

(e) Graph the training and test set log likelihoods, varying the number of clusters from 1 to 10. Discuss how the training set log likelihood varies and why. Discuss how the test set log likelihood varies, how it compares to the training set log likelihood, and why. Finally, comment on how train and test set performance with the "true" number of clusters (3) compares to more and fewer clusters and why.