

Assignment 3

CSG220, Spring 2007

Due: Thursday, Feb. 22

1. (a) Consider trying to estimate the parameters of a multinomial random variable that can take any of K values. These parameters are the underlying probabilities that the variable takes on each of its possible values. Show that applying the m -estimate method (equation 6.22 on p. 179), for a fixed m , is equivalent to computing Bayes estimates of these parameters based on a Dirichlet prior. In particular, express the parameters m and p appearing in that definition (actually one p_i for each possible value of the variable) in terms of the hyperparameters $\alpha_1, \alpha_2, \dots, \alpha_K$ of the Dirichlet distribution.

(b) Show that a simple practical way to implement the use of a Dirichlet prior (and hence the m -estimate) in Bayesian estimation of the parameters of a multinomial random variable is to initialize each of the K counters used to gather relevant statistics to something other than zero when processing the observed data. In particular, specify what counter initialization should be used when the Dirichlet prior with hyperparameters $\alpha_1, \alpha_2, \dots, \alpha_K$ is used. As a special case, how should these counters be initialized if the uniform Dirichlet prior is to be used?

2. Write a program that takes as input any subset of the 14 *PlayTennis* examples from the textbook and implements the corresponding naive Bayes classifier based on this training data. (The discussion on pp. 178-180 should be helpful in clarifying the approach for you on this data.) However, design it so that it estimates each of the conditional probabilities like $P(\text{Outlook} = \text{sunny} \mid \text{PlayTennis} = \text{yes})$ using m -estimates, where m is an input parameter to your program. Implement your program so that the single value of m supplied as input is used to estimate all the relevant conditional probabilities, with the prior probability estimates uniform over all possible values. For example, since there are three possible values for Outlook, the prior estimate of the conditional probability $P(\text{Outlook} = v \mid \text{PlayTennis} = \text{yes})$ should be $1/3$ for each value of v , while the prior estimate of $P(\text{Wind} = u \mid \text{PlayTennis} = \text{yes})$ should be $1/2$ for each of the two possible values of u . Your program should be written to compute all of the necessary quantities for classifying any test instance.

Note: When different attributes have different numbers of possible values, as is the case with the *PlayTennis* data, using the same value of m for all m -estimates is *not* the same as assuming the same Dirichlet prior for all the relevant conditional probability estimates, but this does not matter to us here.

Turn in printout of your source code for this program. It is acceptable to obtain code from elsewhere, but it is preferable that you write this yourself. (This algorithm is particularly straightforward to write.) If you do get some code for this elsewhere, be sure to indicate where.

3. (a) For each value of m from 1 through 10, use your program to produce a corresponding leave-one-out cross-validation (LOOCV) estimate of the probability of misclassification of the naive Bayes classifier (using m -estimates) on the *PlayTennis* examples. Plot a graph showing the number of errors (or proportion of errors) as a function of m . (If your program assigns probability $1/2$ to $P(\text{PlayTennis} = \text{yes})$, count this as $1/2$ an error, on the grounds that an actual classification could then be obtained by applying a coin flip.) Which value(s) of m in $[1, 10]$ give the minimum LOOCV misclassification estimate?

(b) Explain why the result for the best-performing value(s) of m found in part (a) does not yield a very sophisticated classifier for this data. Suggestion: You may want to perform additional experiments to see

why this is the case. Furthermore, it will help if you derive mathematically a simple characterization of the behavior of a naive Bayes classifier using m -estimates (with uniform prior conditional estimates) as $m \rightarrow \infty$.

For this problem turn in: (a) output data obtained during (or upon completion of) the running of the program; (b) any input data or script files used in the running of your program (Lisp dribble files are very helpful here to capture both input and output); and (c) explicit answers to the questions asked. Be sure to annotate any printout you turn in where appropriate.