CALD 10–702
Statistical Approaches to Learning and Discovery

Assignment 1

January 22, 2003

Due in class on Wednesday, January 29.

Instructions: Please submit written answers to each question. We prefer to receive these in printed (as opposed to handwritten) form. You are encouraged to study in groups; however, all of your written work that you submit individually, including homework, is taken by us as your individual effort. If, nonetheless, your written answers reflect collaborative effort, please indicate that to us in an initial footnote. In addition, you are encouraged to use outside references in preparing your answers; however, as with all of your work at the University, you should give citations for your sources.

This assignment includes three problems involving sufficiency, MLE, and a mixture model. For the third problem you will need to numerically fit models to a data set (a fairly large text corpus) and generate plots that show the models' performance. You do not need to hand in source code for this assignment, but you should include a description of your implementation.

Problem 1. *On Sufficiency and MLEs*

As a further qualification of sufficiency, consider the notion of a *minimal* sufficient statistic.

Definition: Let $t(X) = Y$ be a sufficient statistic of the data X for the parameter θ. Y is a *minimal sufficient statistic* if, for each sufficient statistics $s(X) = Z$ for θ, $Y = h(Z)$ for some function h.

The following theorem, due to Lehmann and Scheffe (1950), gives a helpful characterization of a minimal sufficient statistic:

Let $p(X \mid \theta)$ be the probability distribution for a discrete random variable X, given the parameter θ. (In the case of continuous data, let it be the probability density function given θ.) Then, $t(X)$ is minimally sufficient for θ (with respect to X) if and only if it has the property that: $t(X = x') = t(X = x)$ precisely when the ratio $p(x' \mid \theta)/p(x \mid \theta)$ is constant as a function of θ.

Let $X = (X_1, \ldots, X_n)$ be a sample of n observations, iid given θ according a statistical model. In this problem you'll consider the following three statistical models:

Model 1: $X_i \sim \text{Uniform}(0, \theta]$, with $\theta > 0$.

Model 2: $X_i \sim \text{Uniform}[\theta_1, \theta_2]$, with $\theta_2 > \theta_1$ and $\theta_1, \theta_2 \in \mathbb{R}$.

Model 3: $X_i \sim \text{Uniform}[\theta, \theta + 1]$, with $\theta > 0$.

For each of the above models,

(a) Identify the MLE(s) for the parameter(s).

(b) Establish which of the following three reductions of X are sufficient for θ:

  (a) $Y = t_{\max}$

  (b) $W = (t_{\min}, t_{\max})$

  (c) $Z = (t_{\min}, t_{\text{med}})$

  where $t_{\min}(X) = \min_i X_i$, $t_{\max}(X) = \max_i X_i$, and $t_{\text{med}}(X) = \text{median}(X)$ have the natural definitions.

(c) Show whether any of these three is minimally sufficient.

Problem 2. *On some conjugate and "improper" priors*

For each of the following five, one-parameter statistical models we offer the conjugate family of prior distributions for the parameter of that model.

Statistical model	*Conjugate family of priors*
Bernoulli(θ), $0 \le \theta \le 1$	Beta(α, β), $\alpha, \beta > 0$
NegBinom(m, θ), with fixed $m > 0$ and $0 < \theta < 1$	Beta(α, β), $\alpha, \beta > 0$
Poisson(λ), $\lambda > 0$	Gamma(α, β), $\alpha, \beta > 0$
Normal(μ, k^2), with $k > 0$ fixed and $\mu \in \mathbb{R}$	Normal(θ, τ^2), $\theta \in \mathbb{R}, \tau > 0$
Normal(m, σ^2), with $m \in \mathbb{R}$ fixed and $\sigma > 0$	InvGamma(α, β), $\alpha, \beta > 0$

You can find brief descriptions of the respective distributions either in the appendix to chapter 1 of Tanner's book, or (particularly for the negative Binomial) in the textbooks that we placed on reserve in the E&S Library. For each of the models, assuming that the data are a sample of n iid observations from the model:

(a) Calculate the posterior (conditional pbinbetaprobability) for the statistical parameter of the model and verify the conjugacy of the family.

(b) Explain what happens to the posterior distribution for the parameter as the conjugate prior approaches the boundary or boundaries of the family of prior distributions. Indicate whether or not the limit of the conjugate priors is a "proper" probability distribution, or whether it is "improper."

Problem 3. *Beta-Binomial Models for Word Distributions.*

Binomial and multinomial distributions are often used in information retrieval and text processing applications to model word distributions. In this problem you will explore the use of binomial and beta-binomial models for text.

A simple model to to predict the number of times a specific word w appears in a document is the binomial, $p_{\text{bin}}(k \mid n, \theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$ for observing k occurrences of w in a document of size n. However the parameter θ may vary across documents, even those that are on the same topic. In this case we can use a mixture of binomials, using a beta prior. Given a collection of documents d_i with lengths n_i and observed counts k_i for word w, the *evidence* for this data is given by

$$p_{\text{bin-beta}}(\vec{k} \mid \vec{n}, \alpha_w, \beta_w) \;=\; \prod_i \int_0^1 p_{\text{bin}}(k_i \mid n_i, \theta)\, p_{\text{beta}}(\theta \mid \alpha_w, \beta_w)\, d\theta$$

We've made available a corpus of news documents in the file `/afs/cs/academic/class/10602-s00/data/a094.tkntext.gz`. This is a collection of data from the *Topic Detection and Tracking* (TDT) project, containing news stories from a variety of media. The collection is broken up into "document sets," each containing stories from a particular source. Many of the stories are annotated with topic labels, which you will use for this problem. A more detailed description of the format of the file is given in a README in the same directory.

(a) Estimate and compare binomial and beta-binomial models for specific words. Topic numbers 1, 2, 13 and 15 contain the most documents. Choose three words that are representative of each of these topics. Then, estimate the hyperparameters (α_w, β_w) for each of the words you chose for a topic by maximizing $p_{\text{bin-beta}}(\vec{k} \mid n, \alpha_w, \beta_w)$, using only documents on that topic.

(b) Generate a plot that compares the fit of the beta-binomial model to the maximum likelihood binomial model for each word you estimate parameters for. What do you conclude about the binomial versus the beta-binomial models? What models might give a better fit?

(c) Explain how to use the beta-binomial model in a (Bayesian) classifier for predicting "on topic" stories. That is, suppose you are given a collection of stories (documents) that are on a given topic, and another collection of stories that are "off topic." Show how the beta-binomial model can be used to construct a classifier to predict whether future stories are on or off topic.

Note that for (a) you can use any appropriate numerical procedure, such as Newton's method or conjugate gradient, to estimate the parameters. If you use Matlab, you may find the functions `minimize` and `checkgrad` useful; these can be found at `http://www.gatsby.ucl.ac.uk/~edward/code`

To carry out the estimation, you may find it useful to use simple properties of the *digamma function* $\Psi(x) = \frac{\partial}{\partial x} \log \Gamma(x)$. For example, $\Psi(x+1) = \Psi(x) + \frac{1}{x}$, and thus

$$\Psi(k+\alpha) - \Psi(\alpha) = \frac{1}{k-1+\alpha} + \frac{1}{k-2+\alpha} + \cdots + \frac{1}{1+\alpha} + \frac{1}{\alpha}$$

Reference: S. Lowe, "The beta-binomial mixture model for word frequencies in documents with applications to information retrieval," Proceedings of EuroSpeech, 1999.