CALD 10–602
Statistical Approaches to Learning and Discovery

Assignment 1

January 31, 2000

Due 12:00 (noon), February 1.

This assignment includes five problems. For one of the problems you will need to write a program that implements a numerical calculation. Your program must be written in the C or C++ programming language, and should be "stand alone" so that no other programs or non-standard libraries are needed. You should pass you program together with a "Makefile" that will compile it in a standard Unix environment using gnu facilities. (Contact John Lafferty if you have any questions about this.)

Your program and Makefile should be copied to the directory `/afs/andrew/course/10/-602/studentir/<id>/ass1` where `<id>` is your andrew or cs userid. Your written solutions can be handwritten and passed in on paper, or they can be formatted and copied to the above directory as the postscript file `ass1.ps`.

1. *Sufficient statistics.* A statistic $T(X_1, \ldots, X_n)$ is a *sufficient statistic* if $\theta \to T(\vec{X}) \to \vec{X}$ is a Markov chain for any distribution on the parameter θ, and a *minimal sufficient statistic* if $\theta \to T(\vec{X}) \to U(\vec{X}) \to \vec{X}$ is a Markov chain for every other sufficient statistic $U(\vec{X})$. Consider the following three statistics defined on real-valued random variables:

$$
\begin{aligned}
T_{\max}(X_1, \ldots, X_n) &= \max(X_1, \ldots, X_n) \in \mathbb{R} \\
T_{\min,\max}(X_1, \ldots, X_n) &= (\min(X_1, \ldots, X_n), \max(X_1, \ldots, X_n)) \in \mathbb{R}^2 \\
T_{\min,\mathrm{med}}(X_1, \ldots, X_n) &= (\min(X_1, \ldots, X_n), \mathrm{median}(X_1, \ldots, X_n)) \in \mathbb{R}^2
\end{aligned}
$$

For each of the following models $\{f(x \mid \theta)\}_{\theta \in \mathbb{R}}$, which of the statistics $T_{\max}$,$T_{\min,\max}$, $T_{\min,\mathrm{med}}$ are sufficient? Which are minimal? Prove your answers are correct.

(a) $f(\cdot \mid \theta) = \mathrm{Uniform}(0,\, \theta)$ with $\theta > 0$

(b) $f(\cdot \mid \theta) = \mathrm{Uniform}(\theta_1,\, \theta_2)$ with $(\theta_1, \theta_2) \in \mathbb{R}^2$ and $\theta_2 > \theta_1$.

(c) $f(\cdot \mid \theta) = \mathrm{Uniform}(\theta,\, \theta + 1)$ with $\theta > 0$.

2. *Numerical computation of the MLE.* (Tanner, problem 2.9.) 197 animals are distributed into four categories: $Y = (y_1,\, y_2,\, y_3,\, y_4)$ according to the *genetic linkage model* $((2+\theta)/4,\ (1-\theta)/4,\ (1-\theta)/4,\ \theta/4)$.

(a) What is the likelihood for the data $Y = (125, 18, 20, 34)$?

(b) What is the likelihood for the data $Y = (14, 0, 1, 5)$?

(c) Use the Newton-Raphson algorithm to obtain the MLE $\widehat{\theta}$ of θ for $Y = (125, 18, 20, 34)$. Try starting your algorithm at $\theta = 0.1, 0.2, 0.3, 0.4, 0.6$ and 0.8. How did you assess the convergence of the algorithm?

(d) Repeat (c) for $Y = (14, 0, 1, 5)$. Did the algorithm converge for all of the starting points?

(e) Plot the normalized likelihood and the associated normal approximation in the same figure for the data $Y = (125, 18, 20, 34)$. Discuss the adequacy of the normal approximation.

(f) Repeat (e) for $Y = (14, 0, 1, 5)$.

3. *Biased estimates may be better.* Consider the problem of estimating μ and σ^2 from n samples of data drawn i.i.d. from a $\mathcal{N}(\mu, \sigma^2)$ distribution.

(a) Show that $\overline{X}$ is an unbiased estimator of μ.

(b) Show that the estimator

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

is biased and the estimator

$$S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

is unbiased.

(c) Show that S_n^2 has a lower mean squared error than S_{n-1}^2. This illustrates the idea that a biased estimator may be "better" than an unbiased estimator.

4. *Examples of Fisher information.* The Fisher information $J(\theta)$ for the family $f_\theta(x), \theta \in \mathbb{R}$ is defined by

$$J(\theta) = E_\theta\left(\frac{\partial f_\theta(X)/\partial\theta}{f_\theta(X)}\right)^2 = \int \frac{(\frac{\partial f_\theta(x)}{\partial \theta})^2}{f_\theta(x)}\,dx$$

Find the Fisher information for the following families:

(a) $f_\theta(x) = \mathcal{N}(0, \theta) = \frac{1}{\sqrt{2\pi\theta}}e^{-x^2/2\theta}$.

(b) $f_\theta(x) = \theta e^{-\theta x}$, $x \geq 0$

(c) What is the Cramér-Rao lower bound on $E_\theta(\hat{\theta}(X) - \theta)^2$, where $\hat{\theta}(X)$ is an unbiased estimator of θ for (a) and (b)?

5. *Cramér-Rao inequality with a bias term.* Let $X \sim f(x \mid \theta)$ and let $T(X)$ be an estimator for θ. Let $b_T(\theta) = E_\theta(T) - \theta$ be the bias of the estimator. Show that

$$E_\theta(T - \theta)^2 \geq \frac{(1 + b_T'(\theta))^2}{J(\theta)} + b_T^2(\theta)$$