Homework Assignment 1

Due in class, Tuesday September 20th

SDS 383C Statistical Modeling I

1. (7 pts) **Convergence of random variables:** Assume $X_1, \dots X_n \sim f(\mu, \sigma, \mu_3, \mu_4)$. The sample variance $S_n = \frac{\sum_i (X_i - \bar{X}_n)^2}{n-1}$. Prove the following statements. You will need to know the continuous mapping theorem, which states that, for a continuous function g, $X_n \xrightarrow{p} X$, then $g(X_n) \xrightarrow{p} g(X)$ and if $X_n \xrightarrow{d} X$, then $g(X_n) \xrightarrow{d} g(X)$.

   (a) (3 pts) $S_n^2 \xrightarrow{p} \sigma^2$

   (b) (1 pt) $S_n \xrightarrow{p} \sigma$

   (c) (1 pt) $\frac{\bar{X}_n}{S_n} \xrightarrow{p} \frac{\mu}{\sigma}$

   (d) (2 pt) $\sqrt{n}\frac{\bar{X}_n - \mu}{S_n} \xrightarrow{d} N(0,1)$

2. (15 pts) **Maximum Likelihood Estimates:** While the MLE enjoys interesting theoretical properties, its not necessary that it exists or is unique. We illustrate this with the two distributions specified below. For these questions, you must provide a proof.

   (a) (1+1/2+1/2 pts) Let $X_1, \dots, X_n \sim Uniform([\theta, \theta + 1])$. What is the MLE? Is it unique? Does it exist?

   (b) (2 + 3 + 2 pts) $X_1, \dots, X_n \sim Uniform([\theta, 1])$.

      i. What is the MLE? Is it unique? Does it exist? Prove your answer.

      ii. Show that $n(\hat{\theta} - \theta)$ converges in distribution to an exponential distribution. What is the parameter of this exponential distribution?

      iii. In the last question, is the MLE behaving the way it should? Explain your answer.

   (c) (3+3 pts) Let $X_1, \dots, X_n \sim N(\mu, 1)$. Let $\theta := e^{\mu}$. Create a data set with $\mu = 5$ with a hundred observations.

      i. Use the delta method to get the variance of the estimator and a 95% confidence interval of θ.

      ii. Compare the above with the corresponding quantities estimates using the parametric bootstrap.

3. (13 pts) We will use a iterative algorithm to calculate the MLE of the parameters of a Dirichlet distribution. The conjugate prior to the multinomial is the Dirichlet distribution on the k-simplex. You can read more about conjugate priors at here. The density is given by:

$$f(x|\alpha) = \frac{\Gamma(\sum_{i=1}^{k+1} \alpha_i)}{\prod_{i=1}^{k+1} \Gamma(\alpha_i)} \prod_{i=1}^{k+1} x_i^{\alpha_i - 1}$$

where $x_i > 0$, $\sum_i^{k+1} x_i = 1$ and $\alpha_i \geq 0$.

(a) (Extra credit: 5 pts) Prove that $E[\log x_i] = \Psi(\alpha_i) - \Psi(\sum_{i=1}^{k+1} \alpha_i)$, where $\Psi(\alpha) = d \log \Gamma(\alpha)/d\alpha$ is the digamma function. *Hint: you can use the fact that if $X \sim Beta(\alpha, \beta)$, then $E[\log X] = \Psi(\alpha) - \Psi(\alpha + \beta)$.*

(b) (3 pts) For n data-points $\{x^{(i)}, i = 1, \ldots, n\}$ generated from $f(x|\alpha)$, show that the MLE $\hat{\alpha}$ satisfies

$$\overline{\log x_i} = \Psi(\hat{\alpha}_i) - \Psi(\sum_{i=1}^{k+1} \hat{\alpha}_i),$$

where $\overline{\log x_i} = \sum_j \log x_i^{(j)}/n$ is the average computed from data. Since the MLE cannot be computed in closed form, we will use numerical methods to find the MLE. Simple algorithm for doing that is

$$\Psi(\alpha_i^{new}) = \overline{\log x_i} + \Psi(\sum_{i=1}^{k+1} \alpha_i^{old})$$

This will require inverting the digamma function. For a simple method of doing so look at `http://research.microsoft.com/en-us/um/people/minka/papers/dirichlet/minka-dirichlet.pdf` (appendix C).

(c) Use the dataset dir1.txt for a Dirichlet 2-dimensional simplex. It can be found next to the homework link on your instructor's course web page. Now do the following:

i. (1 pt) Give a scatter plot of the data.

ii. (6 pts) Compute the MLE $\hat{\alpha}$, and plot the log-likelihood as a function of iteration. Briefly give a description of the algorithm you use. You can use the built in R/matlab code for digamma, trigamma, gamma or related functions.

iii. (3 pts) Give a scatter plot of the data together with a contour plot of the Dirichlet distribution with parameters $\hat{\alpha}$ that you have computed.