10-602: Statistical Learning

Problem Set 1
Date: January 25, 2001 **Due: start of class February 1, 2001**

---

1. This problem reviews the concept of independence. For each case below, are x and y statistically independent? Why or why not?

   (a) The point (x, y) is uniformly distributed in the unit square.

   (b) The point (x, y) is uniformly distributed in a circle of diameter 1 (a subset of the unit square).

   (c) Let a and b be independent and both uniform on $[0, 1]$. Let $x = a + b$ and $y = a - b$.

   (d) Let a and b be independent and both Gaussian with zero mean and unit variance. Let $x = a + b$ and $y = a - b$.

2. This problem reviews joint distributions and marginalization. Let x and y be random variables with density functions

$$p(x = \hat{x}|y = \hat{y}) = \begin{cases} e^{-(\hat{x}-\hat{y})} & \hat{y} \leq \hat{x} < \infty \\ 0 & \hat{x} < \hat{y} \end{cases}$$

   and

$$p(y = \hat{y}) = \begin{cases} 1 & 0 \leq \hat{y} \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

   (a) What is $p(x = \hat{x}, y = \hat{y})$? Specify the region where the joint density is nonzero.

   (b) Using your favorite software package, generate a 3D plot of $p(x = \hat{x}, y = \hat{y})$.

   (c) Find $p(x = \hat{x})$ and plot that as well.

3. Suppose an urn consists of N balls, Y of which are yellow. The true proportion of yellow balls is thus $p = Y/N$. Out of the urn we take a sample of size n balls, y of which turn out to yellow. (There are $N - n$ balls still left in the urn.) The observed proportion of yellow balls is $\bar{p} = y/n$.

   (a) Find the distribution of y, the number of yellow balls in the sample, given n, N, and Y. Check that your answer is right when $y = Y = 1$ (there was one yellow ball and you found it).

   (b) Is y sufficient for p? Explain.

   (c) Show that $\bar{p}$ is an unbiased estimate of p, and compute the mean square error of $\bar{p}$. It may be helpful to think of y as a sum of indicator variables and to recall the formula for the mean and variance of such a sum.

(d) Suppose we ignore the first ball drawn from the urn and act as if it were yellow. Show that our revised proportion of yellow balls, p^*, is a biased estimate of p and find its mean square error.

(e) Applying the Rao-Blackwell Theorem to the estimator in (d) gives an estimator $E[p^*|y]$. It does not depend on p. Compare the mean square error of this estimator to that of $\bar{p}$ and p^*.

4. Let $Y_1, Y_2, \ldots, Y_n$ be i.i.d. random variables from $\mathrm{U}(\theta_1, \theta_2)$.

(a) Find the MLEs of θ_1 and θ_2. Are they minimal sufficient statistics? Explain.

(b) Suppose that $\theta_2 = \theta_1 + a$ where a is known. What is the minimal sufficient statistic(s) for θ_1? Is the MLE for θ_1 unique? Explain.

5. The two-parameter Gamma distribution has the form (Tanner, p9):

$$p(y|a,b) = \frac{b^a}{\Gamma(a)} y^{a-1} \exp(-by) \qquad y \geq 0, a > 0, b > 0$$

$$E[y] = \frac{a}{b} \qquad \mathrm{var}(y) = \frac{a}{b^2}$$

$$\begin{aligned} E[\log(y)] &= \Psi(a) - \log(b) \\ \text{where } \Psi(a) &= \frac{\Gamma'(a)}{\Gamma(a)} \end{aligned} \tag{1}$$

(a) Put the Gamma distribution into exponential family form and by inspection write down the maximum-likelihood equations for i.i.d. observations $y_1, ..., y_n$. You won't necessarily be able to solve explicitly for a and b.

(b) Prove the result in equation (1). Hint: from the fact that the Gamma distribution is normalized, we know that

$$\log \int_0^\infty y^{a-1} \exp(-by) dy = \log \frac{\Gamma(a)}{b^a}$$

Take the derivative of both sides with respect to a.