Assignment #1

Due: 23:59pm September 20, 2013

Problem 1 (5pts)

X and Y are two independent random variables with distributions $p_X(x)$ and $p_Y(y)$, respectively.
1. Show that the independence of X and Y implies that their covariance is zero.
2. For a scalar constant a, show the following two properties:

$$\mathbb{E}[X + aY] = \mathbb{E}[X] + a\mathbb{E}[Y]$$
$$\text{var}[X + aY] = \text{var}[X] + a^2\text{var}[Y]$$

Problem 2 (Jaynes, 5pts)

Prove that the convolution of two Gaussian distributions is also a Gaussian.

Problem 3 (Murphy, 5pts)

Let $X \in \{0, 1\}$ be a binary random variable with a Bernoulli distribution. Suppose $p(X = 1) = \theta$, so that

$$p(x \mid \theta) = \theta^x(1 - \theta)^{1-x}.$$

Prove that $\mathbb{E}[X] = \theta$ and $\text{var}[X] = \theta(1 - \theta)$.

Problem 4 (10pts)

The beta distribution has support on $(0, 1)$ and has a probability density function given by

$$B(u\,;\, a, b) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} u^{a-1}(1 - u)^{b-1}.$$

For many practical problems, we would like the bounded support of the beta distribution, but on a different interval than $(0, 1)$. One way to achieve this is to linearly transform beta random variates so that they are defined on an interval of our choosing. Write the PDF that results from such a transformation, where the target support is (s, t).

Problem 5 (10pts)

Assume that we have seen N data that are drawn identically and independently from a rescaled beta distribution on (s, t). These data are denoted $\{v_n\}_{n=1}^N$. We would like to find the maximum likelihood estimate of the parameters a and b. The MLE for the beta distribution does not have closed form, so we usually treat it as an optimization problem and use a tool such as BFGS to find the maximum. In this case, the parameters a and b must be positive, however, resulting in a *constrained* optimization problem. To get back to the unconstrained setting, we actually usually work with $\ln a$ and $\ln b$, which take values anywhere on the real line. Derive the gradient of the log likelihood in terms of $\alpha = \ln a$ and $\beta = \ln b$. (Hint: you'll need the digamma function.)

Problem 6 (10pts)

An alternative way to define a distribution with finite support is to use a bijective map from the real line to a bounded interval. Perhaps the most common way to do this is via the logistic function:

$$\sigma(x) = \frac{1}{1 + \exp\{-x\}}.$$

1. Let X be a random variable that has a Gaussian distribution with mean μ and variance ν. If $Y = \sigma(X)$, what is the PDF for Y?
2. If we had data in $(0, 1)$ that were drawn independently and identically from this PDF, how would you quickly compute the MLE of μ and ν? (Hint: The MLE of Gaussian data is easy.)

Jester Collaborative Filtering Data

For the remaining problems, you will need to download the Jester collaborative filtering data. These data are originally from `http://eigentaste.berkeley.edu/dataset/`, however, they have been munged somewhat for this course, so you should download the files directly from the CS281 web page at `http://seas.harvard.edu/courses/cs281#jester`. These problems will only examine the marginal distribution of the ratings themselves. After uncompressing the data, the ratings can be loaded into Matlab with a command such as

```
>> ratings = load('jester_ratings.dat');
```

This will give you a 1761439×3 matrix of doubles. Right now we only care about the ratings, which are the third column. In the following problems, you'll be asked to produce figures. Include these figures in your assignment report.

Problem 7 (10pts)

Generate a set of normalized histograms (histograms which have an area of one) of the ratings and qualitatively describe the empirical distributions that you see. Try several different bin sizes and explain your choices. Are the resulting density estimates uni- or multi-modal? Where do the peaks appear to be? Do these answers change as you vary the number of bins?

Problem 8 (15pts)

Perform a maximum-likelihood fit of a Gaussian distribution to the ratings and report the mean and variance. Overlay the MLE Gaussian fit on top of the normalized histogram. Is it a good fit or a bad fit and why?

Problem 9 (15pt)

Use your results from Problems 5 & 6 to fit a rescaled beta distribution and a (rescaled) logistic normal distribution to the data. Report the parameters you find and plot the MLE densities against your "best" histogram from Problem 7.

Problem 10 (15pts)

Randomly partition the data into ten disjoint sets (called *folds*) of approximately the same size. We will use these partitions to assess the generalization performance of these MLE fits. This is done by creating ten experiments where one fold is taken to be a "test" set and the remaining nine are together considered to be the "training" set. A model is fit on the training data and asked to make predictions of the test set. For a given model, this produces ten log probability numbers that reflect how well the model generalized to the unseen data. If the folds are of different size, the predictive log probabilities can be turned into "averages" by dividing the overall logprob by the number of test cases. Perform this procedure for your histogram density estimators from Problem 7, the Gaussian, rescaled beta, and rescaled logistic Gaussian. That is, fit each of these models ten times on 9/10ths of the data and ask it to make predictions of the remaining 1/10th. To visualize the results, produce a boxplot of the average log probabilities. Which of the three performs best? Why does this appear to be the case?