Homework 5: Normal Distributions & Regression

UC Irvine CS177: Applications of Probability in Computer Science

Due on November 16, 2017 at 11:59pm

Question 1: (20 points)

a) Your friend's new puppy gets loose at Crystal Cove, and begins wandering aimlessly. Every minute, he travels south 5 meters with probability 1/2, or north 5 meters with probability 1/2. The directions of travel at successive minutes are independent. Using the central limit theorem, what is the approximate probability distribution of the puppy's location after 1 hour? Where is he most likely to be?

b) Suppose that immediately after getting loose, the puppy smells a barbecue to the south. Every minute, he travels south 5 meters with probability 2/3, or north 5 meters with probability 1/3. The directions of travel at successive minutes are independent. Using the central limit theorem, what is the approximate probability distribution of the puppy's location after 1 hour? Where is he most likely to be?

c) In order to find the lost puppy, you and your friend decide to identify an interval of locations that contain the puppy with 95% probability. What is the length of the smallest possible interval (in meters) for the motion model in part (a)? What is the length of the smallest possible interval for the motion model in part (b)?

Question 2: (20 points)

A computer job must pass through two queues before it is processed. Suppose the waiting time in the first queue has an exponential distribution with mean $1/\alpha$, and the waiting time in the second queue has an exponential distribution with mean $1/\beta$, independent of the first.

a) Find the expected total waiting time in terms of α and β.

b) Find the standard deviation of the total waiting time in terms of α and β.

c) Assume that $\alpha \neq \beta$, and find the probability density function of the total time the job spends waiting in the two queues. Plot the probability density function, for a grid of total waiting times varying between 0 and 10, in the case where $\alpha = 1$ and $\beta = 2$.

Question 3: (25 points)

Xavier and Yelena have decided to track their personal energy consumption. Assume their monthly energy usages are independent, continuous random numbers that are uniformly distributed between 500 and 1000 kWh. Let X be the energy used by Xavier, and Y be the energy used by Yelena. Recall that the correlation coefficient between any two random variables A and B equals

$$\rho(A, B) = \frac{\text{Cov}(A, B)}{\sqrt{\text{Var}(A)\text{Var}(B)}}.$$

Below we use the correlation coefficient to determine the strength of dependence between the energy use of Xavier and Yelena, and various summary statistics.

a) Determine $\rho(X, Y)$, the correlation between the energy X used by Xavier and the energy Y used by Yelena.

b) Let $Z = X + Y$, the total energy used. Determine $\rho(X, Z)$, the correlation between Xavier's energy use X and the total energy use Z.

c) Let $Z = X + Y$ and $W = X - Y$, the sum and difference of their energy usage. Determine $\rho(Z, W)$, the correlation between this sum and difference.

Question 4: (35 points)

In this problem, your goal is to predict the horsepower Y of a car based on observation of some other feature X of that car. The input feature X and output response Y are both real numbers, so we will use regression models based on bivariate normal distributions. The training data, and corresponding bivariate normal approximations, are plotted in Figure 1.

Suppose that $E[X] = \mu_x$, $E[Y] = \mu_y$, $\text{Var}(X) = \sigma_x^2$, $\text{Var}(Y) = \sigma_y^2$, and $\rho = \frac{\text{Cov}(X,Y)}{\sigma_x \sigma_y}$ is the correlation coefficient. We have provided code that estimates these means, variances, and covariances based on the empirical distribution of the training data. For some test data where $X = x$, you will then predict Y via its conditional mean:

$$\hat{y} = E[Y \mid X = x] = \mu_y + \frac{\rho \sigma_y}{\sigma_x}(x - \mu_x). \tag{1}$$

Given a test dataset of M observations (x_i, y_i), and a model that predicts $\hat{y}_i = E[Y \mid X = x_i]$ for true response y_i, we measure the *root mean squared error* in our predictions as follows:

$$L(y, \hat{y}) = \sqrt{\frac{1}{M} \sum_{i=1}^{M} (y_i - \hat{y}_i)^2} \tag{2}$$

Our goal is to create regression models that make this error as small as possible. *For full credit, the specified error values and plots must be included in your solution pdf.*

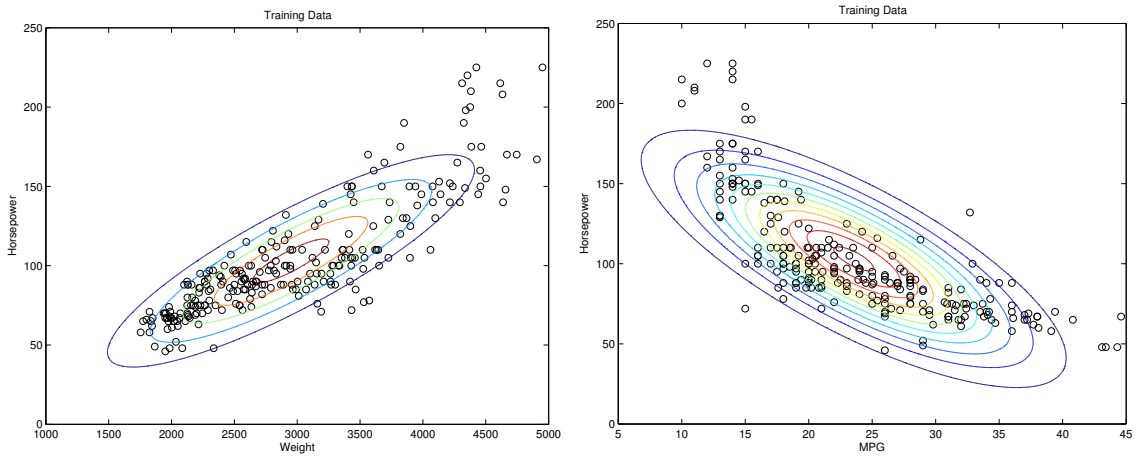

Figure 1: Scatter plots of Y = horsepower versus X = weight (left), and Y = horsepower versus X = miles per gallon (MPG, right), for the automobile training data. For each dataset, we plot contours of constant probability for a bivariate normal distribution fit to the data.

a) We have provided a function `pred_linear` *that predicts horsepower Y as the mean of the training data, $\hat{y} = \mu_y$, ignoring the input features x. Improve this function so that it predicts Y using the conditional mean of Equation* (1).

b) Let the input feature X be vehicle weight, and apply your `pred_linear` *method to predict the test vehicles' horsepower. What is the root mean squared error of your predictions? Plot the true and predicted horsepower values for each test example.*

c) Let the input feature X be vehicle MPG, and apply your `pred_linear` *method to predict the test vehicles' horsepower. What is the root mean squared error of your predictions? Plot the true and predicted horsepower values for each test example.*

From visual inspection, the distribution of the MPG data X does not seem to fit our Gaussian assumption. Fixing a *threshold* of $t = 20$ MPG, we will explore whether we can build a more accurate model by splitting the data into two groups. The first group contains all the examples where $X \leq t$, and the second the examples where $X > t$. Using the training data, we estimate separate bivariate normal distributions for each of the two groups. For a test input x_i, we check whether $x_i \leq t$ or $x_i > t$, and predict $\hat{y}_i$ using the corresponding normal.

d) Implement the threshold-based regression method described above. Let the input feature X be vehicle MPG, the threshold $t = 20$ MPG, and apply your method to predict the test vehicles' horsepower. What is the mean squared error of your predictions? Plot the true and predicted horsepower values for each test example. Compare the accuracy of this approach to the regression methods in parts (b) and (c).