EE 361M: Introduction to Data Mining

Assignment #1

Due: Thur, Feb 4, 2016, 2:00 pm; Total points: 40

Your homework should be written using a word-processor in groups of 3. You may however insert equations by hand if you wish. Homeworks are due at the beginning of class on the due date, and should be submitted through Canvas . If a particular programming language is specified in the problem, please don't use any other language.

1. **(Not for grade: Discuss in group) Data Mining Applications**

   Suppose you are working as a data mining consultant for an Internet Search Engine Company. Briefly discuss how data mining can help the company by giving one example each of an application for which the techniques for (i) regression (ii) classification, (iii) anomaly detection can be used.

2. **(5 pts) Maximum Likelihood Estimation**

   Wishing to estimate the average time it takes to load Canvas on her tablet, Alice does the following study: She records the time x_i that Canvas takes to load (in milliseconds), e $i = 1, ..., N$, at N randomly selected time-points during one day. Suppose that the time it takes to load the webpage can be well represented by $\lambda e^{-\lambda x}$ with (unknown) parameter, λ. Derive the maximum likelihood estimate for λ from first principles. (i.e. do no just write down the answer).

3. **(3+4+3=10 pts). Bivariate Visualization and Mathematical Form**

   Suppose X and Y are two random variables whose joint distribution is Normal (Gaussian), centered at (0,0) and with correlation ρ. (See "Bivariate Case" in the Wikipedia entry for "Multivariate Normal Distribution" for the equation, or use just use the vector form given in the class notes, with $\sigma_{12} = \sigma_{21} = \rho\sigma_x\sigma_y$). Consider 2 cases
   i) $\sigma_x^2 = 4; \sigma_y^2 = 9; \rho = 0$
   ii)$\sigma_x^2 = 4; \sigma_y^2 = 9; \rho = 0.5$

   (a) Obtain contour plots for each of the two distributions using Python (`http://stanford.edu/~mwaskom/software/seaborn/generated/seaborn.kdeplot.html`).

   (b) View 3-D plots for the two distributions from at least two different viewing perspectives each (`http://matplotlib.org/examples/mplot3d/rotate_axes3d_demo.html`).

   (c) Consider the bivariate Normal Distribution given in part (ii). Reading the "Bivariate Case" under "Conditional distributions" in the Wikipedia entry will help you answer this problem; alternatively you can consult any undergraduate text on probability/statistics. What is the mathematical form of the conditional distribution that is obtained when (a) x is set to 1, and (b) when y is set to 1? (no need to actually derive the formulae from first principles; rather just obtain the result by substition in the formula for a bivariate Gaussian.

4. **(2×5 =10 pts) Exploratory Data Analysis using Python**

   The "student" data set (found on Canvas) records properties of 657 students. For a description of the data, see http://cran.r-project.org/web/packages/LearnBayes/LearnBayes.pdf. Python packages that will be useful: Pandas, Matplotlib, and Seaborn.

   (a) Construct a histogram of the variable `Shoes`. (Use 20 bins)

(b) Use data visualization to check if the variable `Dvds` (approximately) follows a log-normal distribution.

(c) Summarize the variable `Haircut` using the Panda's describe command. Also, report the 2.5th and 97.5th percentiles.

(d) Construct a barplot of the individual values of `Drink` that were observed. Also, highlight the distribution of the variable `Drink` between the two genders on the same barplot.

(e) Construct a scatter plot of the variables `ToSleep` and `WakeUp`. Do you observe a positive correlation between the two variables?

Note: Omit missing values if any are present.

5. **$(2+2+5+6$ =15 pts) Regression with scikit-learn**

For this problem, we will be using the scikit-learn package in Python to predict housing prices using regression. The housing data can be found here: `https://archive.ics.uci.edu/ml/machine-learning-databases/housing/`.

(a) Fit a simple linear regression of medv on lstat. What is the coefficient on lstat and what is its interpretation?

(b) Plot medv and and lstat in a scatter plot with the regression line. Comment on the fit.

(c) Split the data randomly into a training and test set (1/3 for test). Train a multiple linear regression on all the variables using the training data and evaluate the trained model on the test data using root mean squared error. Discuss the results.

(d) Let us try to fit an MLR to this dataset, with MEDV as the dependent variable. MEDV has a somewhat longish tail and is not so Gaussian-like, so we will take a log transform, and then predict LMDEV instead. (You should convince yourself that this is a better idea by looking at the histograms and quantile plots to assess normality; however no need to submit such plots). Keep the first 350 records as a training set (call it Bostrain) which you will use to fit the model; the remaining 156 will be used as a test set (Bostest). Use only the following variables in your model: $LMEDV \sim LSTAT + RM + CRIM + ZN + CHAS$.

i. Report the MSE obtained on Bostrain. How much does this increase when you score your model on Bostest?

ii. Report the coefficients obtained by learning the regression model.

iii. Do you think your MLR model is reasonable for this problem? You may look at the distribution of residuals to provide an informed answer.