MIS 382N: Advanced Predictive Modelling

Assignment #1

Due: Wed, Sep 17, 2014, 3.30 pm; Total points: 40

Your homework should be written using a word-processor. 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. **(1 × 3=3 pts) Data Mining Applications**

   Suppose you are working as a data mining consultant for an Internet Search Engine Company. Briefly describe (1 short paragraph each) 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. **(2 × 3=6 pts) Sampling**

   A recent survey estimated that 30% of all Europeans aged 20 to 22 have driven under the influence of drugs or alcohol, based on a simple "Yes or No" question. A similar survey is being planned for Americans. The survey designers want the 90% confidence interval to have a margin of error of at most ± 0.09.

   (a) Find the necessary sample size needed to conduct this survey assuming that the expected percentage of "yes" answers will be very close to that obtained from the European survey?

   (b) Suppose the tolerance level was kept the same but the confidence level needs to increase to 95%. What is the required sample size for this new specification?

   (c) If one does not know where the true "p" may lie, one can conservatively conduct a survey assuming the worst case (in terms of required minimum sample size) scenario of $p = 0.5$. Redo part (b) for this "worst case" scenario.

3. **(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 an exponential distribution with (unknown) parameter, λ. Derive the maximum likelihood estimate for λ from first principles. (i.e. do no just write down the answer).

4. **(6 pts) Ilustration of "curse of dimensionality"** Consider the set of 2^{10000} binary vectors of length 10000 (i.e. $00\ldots0$ through $111\ldots1$). The Hamming distance between any two vectors (number of bits they differ by) from this will be a number between 0 and 10000. What is the probability that the Hamming distance of an arbitrary chosen pair of vectors is between 4950 and 5050? (Hint: distribution of distances is binomial; can be approximated by a Normal distribution as $n \gg 30$. You can get a table for a Standard Normal distribution from many sources, e.g., http://www.stat.ucla.edu/ỹwu/teaching/normal.pdf)

   This problem shows that if data is randomly distributed at the corners of a very high-dimensional hypercube, then for any point, most of the other points are bunched at about the same distance, so there is little scope of focusing on points with known "y" in a nearby neighborhood to do any prediction.

5. **(3+4+3=6 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 the *contour* function in R to get the plots.

(b) View 3-D plots for the two distributions from atleast two different viewing perspectives each. The function `Persp` allows you to specify viewing perspective by setting "theta" and "phi".

(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.

6. **(2×5 =10 pts) Exploratory Data Analysis using R**

The "student" data set, part of the LearnBayes package, records properties of 657 students. For a description of the data, see http://cran.r-project.org/web/packages/LearnBayes/LearnBayes.pdf.

(a) Construct a histogram of the variable `Shoes`. (Set the value of the parameter `breaks` to 20)

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

(c) Summarize the variable `Haircut` using the `summary` 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.