Northeastern University
College of Science

Final Exam for MATH7340

For this final you must answer the first four problems. The fifth problem is extra bonus. Provide the numerical answers and plots as requested, and show your derivations. If one parts just ask about handling data set, provide the R commands to do the task. And please also provide an executable R script file with all the R commands used, and clearly label which question the R commands are for.

Problem 1. (10 points) MLE and bootstrap for Poisson data.

A random sample from the Poisson distribution $poisson(\lambda = e^{\theta})$ is provided in the file "DataPois.txt".

(a) What is the sample size n? What is the sample mean $\bar{Y}$?

(b) Find the value of MLE $\hat{\theta}$ on this data set *using numerical method*. NO credit is given if you used an analytic formula to find MLE.

(c) Test the null hypothesis that $\theta = 1$ at level 0.05, using a bootstrap confidence interval. Do NOT use analytic formula for MLE in your calculation. Use numerical optimization as in part (b). What is your conclusion? Can you find the p-value?

Instructions on inputting the data set: You should download the file, put it in the working directory of your R session. Then load it using command

```
y<-as.numeric(t(read.table(file="DataPois.txt", header=TRUE)))
```

Hint: For $\theta = x$, the probability density of observing Y=y is calculated in R as

```
dpois(y, lambda=exp(x))
```

You can find the negative log-likelihood for the data set using the above expression.

Problem 2. (20 points) ANOVA

We analyze the data set NCI60 data from the ISLR library. (This data set was used in Homework 11).

(a) Delete the cancer types with only one or two cases ("K562A-repro", etc.). Keep only the cancer types with more than 3 cases.

(b) Analyze the expression values of the first gene in the data (first column). Does the first gene express differently in different types of cancers? If so, in which pairs of cancer types does the first gene express differently? (Use FDR adjustment.)

(c) Check the model assumptions for analysis in part (b). Is ANOVA analysis appropriate here?

(d) Apply ANOVA analysis to each of the 6830 genes. At FDR level of 0.05, how many genes express differently among different types of cancer patients?

Problem 3. (10 points) Regression

We consider the regression analysis on the state.x77 data set. In the module 9, we regressed the life expectancy on three variables: the murder rates, percentage of high-school graduates and mean number of frost days.

(a) Make pairwise scatterplots for all variables in the data set. Which variables appears to be linearly correlated with the life expectancy based on the scatterplots?

(b) Conduct a regression analysis different from the example analysis in module 9. We regress the life expectancy on three variables: the per capita income (Income), the illiteracy rate (Illiteracy) and mean number of frost days (Frost). What is your regression equation? In this regression analysis, which of the three variables affect the life expectancy significantly?

(c) Find delete-one-cross-validated mean square errors $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{(-i)})^2$ for this regression model. Here y_i is the i-th response (life expectancy), $\hat{y}_{(-i)}$ is the prediction of the i-th response from the regression fit without using that observation.

Problem 4. (60 points) Predicting B-cell differentiation with gene expression.
We analyze data for the B-cell patients in the ALL data set in the textbook.
(a) Select gene expression data for only the B-cell patients. The analysis in following parts will only use these gene expression data on the B-cell patients.
(b) Select only those genes whose coefficient of variance (i.e., standard deviation divided by the mean) is greater than 0.2. How many genes are selected?
(c) We wish to conduct clustering analysis to study natural groupings of the patients predicted by the gene expression profiles. For this analysis, we first need to reduce the number of genes studied. The filter in (b) is one such choice. Please comment on what filtering methods you would use to choose genes, other than the filter in (b). What would you consider as the best gene filter in this case.

(d) Conduct a hierarchical clustering analysis with filtered genes in (b). (For uniformity in grading, we ask everyone to use the filter in (b). It may not be your best filter in (c).) How do the clusters compare to the B-stages? How does do the clusters compare to the molecule biology types (in variable ALL$mol.biol)? Provide the confusion matrices of the comparisons, with 4 clusters.

(e) Draw two heatmaps for the expression data in (d), one for each comparison. Using colorbars to show the comparison types (B-stages or molecule biology types). The clusters reflect which types better: B-stages or molecule biology types?

(f) We focus on predicting the B-cell differentiation in the following analysis. We merge the last two categories "B3" and "B4", so that we are studying 3 classes: "B1", "B2" and "B34". (Ignore the unknown type "B" in the analysis.) Use linear model (limma library) to select genes that expresses differently among these three classes at FDR of 0.05. How many genes are selected?

(g) Fit SVM and the classification tree on these selected genes in part (f), evaluate their performance with delete-one-cross-validated misclassification rate.

(h) We select the genes passing both filters in (b) and (f). How many genes are selected? Redo part (g) on these genes passing both filters.

(i) Which classifier you will consider best among the classifiers studied in part (g) and part (h)? Why?

Problem 5. Bonus question (Extra 10 points) Poisson regression.

A random sample of $(X_1, Y_1), ..., (X_n, Y_n)$ is provided in the file "DataPoisReg.txt". The Y_i comes from the Poisson distribution $poisson(\lambda = e^{\beta_0 + \beta_1 X_i})$.

(a) Find the value of MLE $(\hat{\beta}_0, \hat{\beta}_1)$ on this data set *using numerical method*. NO credit is given if you used an analytic formula to find MLE.

(b) Is the slope parameter $\beta_1 = 2$? Use appropriate statistical test to answer this question.

Instructions on inputting the data set: You should download the file, put it in the working directory of your R session. Then load it using command

```
my.dat<-read.table(file="DataPoisReg.txt", header=TRUE)
```