Assignment 3 - CSC/DSC 265/465 - Spring 2020 - Due April 29

Q1: Suppose a logistic regression model is fit with response $Y =$ `Minor injury during the past year` and one predictor $X =$ `age`. The sample size is $n = 100$ and the range of X in years is $[3.2, 18.1]$. The data was fit using the `R glm()` function, and produced the following output:

```
> fit = glm(y ~ x, family='binomial')
> summary(fit)

Call:
glm(formula = y ~ x, family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0094  -0.6021  -0.4758  -0.3770   2.3251

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.89379    1.09305  -3.562 0.000368 ***
x            0.19859    0.09147   2.171 0.029927 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 84.542  on 99  degrees of freedom
Residual deviance: 79.500  on 98  degrees of freedom
AIC: 83.5

Number of Fisher Scoring iterations: 5
```

What is the estimated odds ratio for probability of `Minor injury during the past year` between 15 year old and 5 year old subjects? Give an approximate 95% confidence interval.

Q2: For this question, use the `NCI60` data set from the `ISLR` package. From the `help` page:

```
NCI microarray data. The data contains expression levels on 6830 genes
from 64 cancer cell lines. Cancer type is also recorded.
```

The object `NCI60$data` is a 64×6830 matrix. Each row represents a cancer cell line, and contains 6830 gene expression measurements. Assume that the measurements are already standardized. The object `NCI60$labs` is a vector of length 64 containing text identifying the type of cancer (`OVARIAN`, `MELANOMIA`, and so on).

(a) Use the `hclust()` function (from library `stats`) to create a hierarchical clustering of the cancer cell lines. Use the option `method = 'average'` to specify the average distance agglomeration method. Plot the dendogram using labels from the `NCI60$labs` object.

(b) From the dendogram, it can be seen that some cancer types cluster more definititively than others. The tendency of a cancer type to cluster can be quantified by comparing the maximum cophenetic distance between samples of this type, and the minimum cophenetic distance between a sample of this type and a sample not of this type. Determine these quantities for the `MELANOMA`, `RENAL` and `COLON` cancer types, and comment briefly on what you find.

Q3: For this question, use the `mammals` dataset from the `MASS` library. This data frame contains average body and brain weights for 62 species of land mammals (in kilograms and grams respectively). The names of the mammals can be accessed using `row.names(mammals)`.

(a) First, log-transform the data. Then for each $K = 1, \ldots, 10$ calculate a K-means cluster solution based on the two log-transformed features. Use option `nstart=100`. For each solution calculate $R^2 = 1 - SS_{within}/SS_{total}$, and plot these values against K. Identify the smallest value of K, say K^*, for which $R^2 \geq 0.8$.

(b) Draws a scatterplot of the features, and superimpose the centers output with the clustering solution for K^*. Distinguish the observations by using separate colors for the clusters identified by the solution. These clusters can be identified as follows:

```
fit = kmeans(x,centers=nc,nstart=100)
fit$cluster
```

Create separate lists of the species names for each cluster. Does the clustering make sense? Comment briefly.

Q4: For this question, use the `Khan` data set from the `ISLR` package. From the `help` page:

```
The data consists of a number of tissue samples corresponding to four distinct types
of small round blue cell tumors. For each tissue sample, 2308 gene expression
measurements are available.

...

Format

The format is a list containing four components: xtrain, xtest, ytrain, and ytest.
xtrain contains the 2308 gene expression values for 63 subjects and ytrain records
the corresponding tumor type. ytrain and ytest contain the corresponding testing
sample information for a further 20 subjects.
```

Consolidate the training and test data into a single dataset:

```
> library(ISLR)
> library(class)
>
> x = rbind(Khan$xtrain,Khan$xtest)
> y = c(Khan$ytrain,Khan$ytest)
> dim(x)
[1]   83 2308
> length(y)
[1] 83
```

There is now a single data set. The object `x` is an 83×2308 table, with 2308 gene expression measurements for each of 83 tissue samples. Then `y` is a vector of length 83 containing the tumor type, labeled 1 to 4.

(a) Use the `prcomp()` function (from library `stats`) to create a matrix of principal components, using the gene expressions as a feature set. Use centering, but not scaling.

(b) Build a KNN classifier for tumor type based on (a) the entire set of gene expressions; and (b) the first 10 principal components. Use the method of **Q4** of Assignment 2, using `k.list = seq(1,50,1)`. For each analysis plot CE against K (the neighborhood size). Report only CE. Which classifier is preferable (give several reasons for your answer)?

Q5: [**For Graduate Students**] Explain why scaling makes a difference for K-means clustering but not linear discriminant analysis.