Homework 3 - 131/231

Due on Sunday May 20, 2018 at 11:59 pm

For this homework you will need use the following packages.

```
library(tidyverse)
library(ROCR)
library(tree)
library(maptree)
library(class)
library(lattice)
```

Analyzing drug use

The first half of this homework involves the analysis of drug use. The data set includes a total of 1885 observations on 32 variables. A detailed description of the data set can be found here. For each observation, 12 attributes are known:

- ID: number of record in original database. Used for reference only.
- Age: Age of the participant
- Gender: Gender of the participant (M/F)
- Education: Level of education of the participant
- Country: Country of current residence of the participant
- Ethnicity: Ethnicity of the participant

Many of the covariates have been transformed: some ordinal or categorical variables have been given numeric codes. Part of this problem will involve appropriately re-transforming these variables. The data also contains the following personality measurements:

- Nscore: NEO- FFI- R Neuroticism (Ranging from 12 to 60)
- Escore: NEO- FFI- R Extraversion (Ranging from 16 to 59)
- Oscore: NEO- FFI- R Openness (Ranging from 24 to 60)
- Ascore: NEO- FFI- R Agreeableness (Ranging from 12 to 60)
- Cscore: NEO- FFI- R Conscientiousness (Ranging from 17 to 59)
- Impulsive: Impulsiveness measured by BIS- 11
- SS: Sensation Seeking measured by ImpSS

Finally, participants were questioned concerning their use of 18 legal and illegal drugs (alcohol, amphetamines, amyl nitrite, benzodiazepine, cannabis, chocolate, cocaine, caffeine, crack, ecstasy, heroin, ketamine, legal highs, LSD, methadone, mushrooms, nicotine and volatile substance abuse) and one fictitious drug (Semeron) which was introduced to identify over-claimers. All of the drugs use the class system of CL0-CL6: CL0 = "Never Used", CL1 = "Used over a decade ago", CL2 = "Used in last decade", CL3 = "Used in last year", CL4 = "Used in last month", CL5 = "Used in last week", CL6 = "Used in last day".

```
drug_use <- read_csv('drug.csv',
                     col_names = c('ID','Age','Gender','Education','Country','Ethnicity',
                                   'Nscore','Escore','Oscore','Ascore','Cscore','Impulsive',
                                  'SS','Alcohol','Amphet','Amyl','Benzos','Caff','Cannabis',
                                  'Choc','Coke','Crack','Ecstasy','Heroin','Ketamine',
                                 'Legalh','LSD','Meth','Mushrooms','Nicotine','Semer','VSA'))
```

1. Logistic regression for drug use prediction

This problem has 4 parts for 131 students and 5 parts for 231 students. As mentioned, the data uses some strange encodings for variables. For instance, you may notice that the gender variable has type `double`. Here the value -0.48246 means male and 0.48246 means female. Age was recorded at a set of categories but rescaled to a mean 0 numeric variable (we will leave that variable as is). Similarly education is a scaled numeric quantity (we will also leave this variable as is). We will however, start by transforming gender, ethnicity, and country to factors, and the drug response variables as ordered factors:

```
drug_use <- drug_use %>% mutate_at(as.ordered, .vars=vars(Alcohol:VSA))
drug_use <- drug_use %>%
  mutate(Gender = factor(Gender, labels=c("Male", "Female"))) %>%
  mutate(Ethnicity = factor(Ethnicity, labels=c("Black", "Asian", "White",
                                                "Mixed:White/Black", "Other",
                                                "Mixed:White/Asian",
                                                "Mixed:Black/Asian"))) %>%
  mutate(Country = factor(Country, labels=c("Australia", "Canada", "New Zealand",
                                            "Other", "Ireland", "UK", "USA")))
```

(a). Define a new factor response variable `recent_cannabis_use` which is "Yes" if a person has used cannabis within a year, and "No" otherwise. This can be done by checking if the `Cannabis` variable is greater than *or equal* to `CL3`. Hint: use `mutate` with the `ifelse` command. When creating the new factor set `levels` argument to `levels=c("No", "Yes")` (in that order).

(b). We will create a new tibble that includes a subset of the original variables. We will focus on all variables between `age` and `SS` as well as the new factor related to recent cannabis use. Create `drug_use_subset` with the command:

```
drug_use_subset <- drug_use %>% select(Age:SS, recent_cannabis_use)
```

Split `drug_use_subset` into a training data set and a test data set called `drug_use_train` and `drug_use_test`. The training data should include 1500 randomly sampled observation and the test data should include the remaining observations in `drug_use_subset`. Verify that the data sets are of the right size by printing `dim(drug_use_train)` and `dim(drug_use_test)`.

(c). Fit a logistic regression to model `recent_cannabis_use` as a function of all other predictors in `drug_use_train`. Fit this regression using the training data only. Display the results by calling the `summary` function on the logistic regression object.

(d). (**231** only). Generalized linear models for binary data involve a link function, g which relates a linear function of the predictors to a function of the probability p: $g(p) = \beta_0 + \beta_1 X_1 + ...\beta_p X_p$. g is a function which maps $p \in [0, 1]$ to $\mathbb{R}$. Logistic regression is based on the *logit* link function, $g(p) = log(p/(1-p))$. In class we mentioned another link function, called the probit: $g(p) = \Phi^{-1}(p)$ where Φ is the cumulative density function of the normal distribution. Another often used link function is the "c-log-log" link: $g(p) = log(-log(1-p))$.

Plot the fitted values for logistic regression fit of the training data on the x-axis and the fitted values for the probit regression on y-axis. In the plot command (assuming you use the base plotting package, not ggplot) set `pch=19` and `cex=0.2` (this makes the points smaller and more legible). Include the line y=x with the command `abline(a=0, b=1, col="red")`. Make another identical plot, this time replacing the y-axis with the predicted values from a cloglog regression.

Comment on the differences between the estimated probabilities in each plot. Things you should comment on include: 1) which link function (probit or cloglog) leads to predictions that are most similar to the logistic regression predictions? 2) for what range of probabilities are the probit and cloglog predictions values more or less extreme than the logit values? 3) Does either probit or cloglog regression seem to estimate systematically smaller or larger probabilities than the logistic regression for a certain range of probabilities?

Hint: in logistic regression we set `family=binomial(link="logit"")`. To fit probit and cloglog regressions change the value of the `link` argument appropriately.

2. Decision tree models of drug use

This problem has 3 parts for all students (25 points total)

Construct a decision tree to predict `recent_cannabis_use` using all other predictors in `drug_use_train`. Set the value of the argument `control = tree_parameters` where `tree_parameters` are:

```
tree_parameters = tree.control(nobs=nrow(drug_use_train), minsize=10, mindev=1e-3)
```

This sets the smallest number of allowed observations in each leaf node to 10 and requires a deviance of at least 1e-3 to split a node.

(a). Use 10-fold CV to select the a tree which minimizes the cross-validation misclassification rate. Use the function `cv.tree`, and set the argument `FUN=prune.misclass`. Note: you do not need to use a `do.chunk` function since the `tree` package will do cross validation for you. Find the size of the tree which minimizes the cross validation error. If multiple trees have the same minimum cross validated misclassification rate, set `best_size` to the smallest tree size with that minimum rate.

(b). Prune the tree to the size found in the previous part and plot the tree using the `draw.tree` function from the `maptree` package. Set `nodeinfo=TRUE`. Which variable is split first in this decision tree?

(c). Compute and print the confusion matrix for the *test* data using the function `table(truth, predictions)` where `truth` and `predictions` are the true classes and the predicted classes from the tree model respectively. Note: when generated the predicted classes for the test data, set `type="class"` in the `predict` function. Calculate the true positive rate (TPR) and false positive rate (FPR) for the confusion matrix. Show how you arrived at your answer.

3. Model Comparison

This problem has 2 parts for all students worth 15 points total.

(a). Plot the ROC curves for both the logistic regression fit and the decision tree on the same plot. Use `drug_use_test` to compute the ROC curves for both the logistic regression model and the best pruned tree model.

(b). Compute the AUC for both models and print them. Which model has larger AUC?

4. Clustering and dimension reduction for gene expression data

This problem involves the analysis of gene expression data from 327 subjects from Yeoh *et al* (2002). The data set includes abundance levels for 3141 genes and a class label indicating one of 7 leukemia subtypes the patient was diagnosed with. The paper describing their analysis of this data can be found here. Read in the csv data in `leukemia_data.csv`. It is posted on Piazza in the resources tab with the homework:

```
leukemia_data <- read_csv("leukemia_data.csv")
```

This problem has 4 parts for 131 students for a total of 35 points and 7 parts for 231 students for a total of 50 points.

(a). The class of the first column of `leukemia_data`, `Type`, is set to `character` by default. Convert the `Type` column to a factor using the `mutate` function. Use the `table` command to print the number of patients with each leukemia subtype. Which leukemia subtype occurs the least in this data?

(b). Run PCA on the leukemia data using `prcomp` function with `scale=TRUE` and `center=TRUE` (this scales each gene to have mean 0 and variance 1). Make sure you exclude the `Type` column when you run the PCA function (we are only interested in reducing the dimension of the gene expression values and PCA doesn't work with categorical data anyway). Plot the proportion of variance explained by each principal component (PVE) and the cumulative PVE side-by-side.

```
pve <- ## Fill this in
cumulative_pve <- ## fill this in

## This will put the next two plots side by side
par(mfrow=c(1, 2))

## Plot proportion of variance explained
plot(pve, type="l", lwd=3)
plot(cumulative_pve, type="l", lwd=3)
```

(c). Use the results of PCA to project the data into the first two principal component dimensions. `prcomp` returns this dimension reduced data in the first columns of `x`. Plot the data as a scatter plot using `plot` function with `col=plot_colors` where `plot_colors` is defined

```
rainbow_colors <- rainbow(7)
plot_colors <- rainbow_colors[leukemia_data$Type]
```

This will color the points according to the leukemia subtype. Add the leukemia type labels to the plot using `text` with `labels` argument set to the leukemia type and the col to `plot_colors` (it may help legibility to make the points on the plot very small by setting `cex` to a small number). Which group is most clearly separated from the others along the PC1 axis? Which genes have the highest absolute loadings for PC1 (the genes that have the largest weights in the weighted average used to create the new variable PC1)? You can find these by taking the absolute values of the first principal component loadings and sorting them. Print the first 6 genes in this sorted vector using the `head` function.

(d). (**231 Only**) PCA orders the principal components according to the amount of total variation in the data that they explain. This does not mean, however, that the principal components are sorted in terms of how useful they are at capturing variation between the leukemia groups. For example, if gene expression varied significantly with age and gender (independent of leukemia status), the first principal components could reflect genetic variation due to age and gender, but not to leukemia. The first scatter plot shows that the second PC is not a good discriminator of leukemia type. See if the 3rd PC is better at discriminating between leukemia types by plotting the data projected onto the *first* and *third* principal components (not the second).

(e.) (**231 Only**) For this part we will be using the `ggridges` library. Create a new tibble where the first column (call it `z1`) is the projection of the data onto the first principal component and the second column is the leukemia subtype (`Type`). Use `ggplot` with `geom_density_ridges` to create multiple stacked density plots of the projected gene expression data. Set the ggplot aesthetics to `aes(x = z1, y = Type, fill = Type)`. Make another identical plot, except replace `z1` with `z3`, the projection of the data onto the third principal component. Identify two leukemia subtypes that are nearly indistinguishable when the gene expression data is projected onto the first PC direction, but easily distinguishable when projected onto the third PC direction.

(f.) Use the `filter` command to create a new tibble `leukemia_subset` by subsetting to include only rows for which `Type` is either T-ALL, TEL-AML1, or Hyperdip50. Compute a euclidean distance matrix between the subjects using the `dist` function and then run hierarchical clustering using complete linkage. Plot the resulting dendrogram.

(g). (**231 only**). Use `levelplot` to plot the distance matrix from the part above. Order the rows and columns by the hierarchical clustering you obtained in the previous part. If the plot seems faint, you can adjust the color bar by setting the `at` argument to `at=pretty(c(min, max), n=10)` where you will choose the value of `min` and `max` as smallest and largest distances in the color range. You should see a matrix with a *block diagonal* structure. The labels (corresponding to leukemia types) will be hard to read on the plot. Print them out by looking at `leukemia_subset$Type` ordered by clustering order. Based on this plot an the orderings which two leukemia types (of the three in the subset) seem more similar to one another?