MATH 6020 – HW 4

Due back on Sunday, 2nd May [1].

-1- (Discriminant Analysis) Let $\mathrm{e}_1, \ldots, \mathrm{e}_q \in \mathbb{R}^q$ be eigenvectors of $[\mathbf{S}_p^2]^{-1}\,\mathbf{B}$ with eigenvalues $\lambda_1 \geq \ldots \geq \lambda_q \geq 0$. Show that:

$$\max f(\mathrm{a}) = \max \frac{\mathrm{a}_1' \,\mathbf{B}\, \mathrm{a}_1}{\mathrm{a}_1' \,\mathbf{S}_p^2\, \mathrm{a}_1} \text{ is attained at } \mathrm{a}_1 = \mathrm{e}_1\,.$$

Use the fact that

$$\operatorname{argmax} g(\mathrm{a}) = \mathrm{e} \text{ where } g(\mathrm{a}) = f\left([\mathbf{S}_p^2]^{-\frac{1}{2}}\,\mathbf{a}\right).$$

-2- Consider the data set on brands of cereals in `T11-9.DAT`, which is discussed in Exercise 11.34 of Johnson and Wichern. The columns of the data set are

`Brand, Manufacturer, Calories, Protein, Fat, Sodium, Fiber, Carbohydrates, Sugar, Potassium, Group`

The last column group is just a numerical relabeling of the manufacturer variables.

-a- Carry out centroid linkage and complete link agglomerative hierarchical clustering. Form 4 clusters in each case, and show the dendograms. Use the Euclidean distance as your distance function.

-b- Carry out K-means clustering with $K = 4$.

-c- Carry out model-based clustering, forming 4 clusters and assuming a mixture of normal densities.

-d- Are there any noticeable similarities or differences among the results? Are there cereals that are substantially different from the rest? Is there a relationship between clusters and manufacturers?

Note that `R` has most of the built-in methods for this, see for example:

`http://www.sthda.com/english/articles/25-cluster-analysis-in-r-practical-guide/111-types-of-clustering-methods-overview-and-quick-start-r-code/`
`https://cran.r-project.org/web/packages/cluster/cluster.pdf`

For the next part of the homework, you are asked to implement two types of random forests in `R`: one for regression on data with a continuous output, and another for classification on different data that has a categorical output. Online you can find many, many tutorials for how to implement both types of random forests in `R`. You don't have to use `R` either, you can also use Python or some other package.

-3- First is the regression tree. We will use the `Ames Housing` dataset, whose documentation can be found here:

[1] HW Version: 2021-04-28 at 20:49

`http://jse.amstat.org/v19n3/decock.pdf`

The data set is 2390 records of house sales in Ames, Iowa. The goal is to use random forests to build a regression model of the sales price on some of the other variables. Note that there are 80 predictor variables for each home, which is quite large, so we will only use a smaller number to build the model.

-a- Use `install.packages("AmesHousing")` to install the dataset. See

`https://cran.r-project.org/web/packages/AmesHousing/AmesHousing.pdf`

for a description of the predictor variables.

-b- Use the function `make_ordinal_ames` to get a clean version of the data (see the documentation above). We will use the following 30 variables as predictors:

`Lot Frontage, Lot Area, Bldg Type, House Style, Overall Qual, Overall Cond, Year Built, Roof Style, Roof Matl, Exterior 1, Exter Qual, Foundation, Bsmt Cond, Bsmt Unf SF, Total Bsmt SF, Heating, Central Air, Electrical, 1st Flr SF, 2nd Flr SF, Full Bath, Half Bath, Bedroom, Kitchen, KitchenQual, Fireplaces, Garage Cars, Garage Area, Fence, Misc Val`

Omit any rows that contain NA values.

-c- Split the data into a testing and training set, with 70% of the data going to the training set and the rest to the testing set.

-d- Build a random forest out of the training set, predicting *Sale Price* on all 30 predictor variables listed above. What MSE do you get?

-e- Return and plot the importance of each predictor variable in the random forest.

-f- Predict the sales prices for the houses in the testing set, based on the forest built from the training set, and compute the MSE.

-4- For 97 different students the following data is reported in `Grades.csv` (available on Canvas):

- `Quizzes:` cumulative score (out of 100) on semester quizzes
- `Mid1:` cumulative score (out of 100) on first midterm
- `Mid2:` cumulative score (out of 100) on second midterm
- `Final:` cumulative score (out of 100) on final exam
- `Grade:` final letter grade in course, possible values are `A, A-, B+, B, B-, C+, C, C-, D, F`

-a- Run a classification tree of the grade against the other four variables. Which of the numerical scores is most important in determining the letter grade?

-b- Plot and label the classification tree.

-c- Split the 97 students into a training set and a testing set, with 70% of the data in the training set and the rest in the test. Build the random forest off of the training set.

-d- Of the four measurements, what are their relative importances for predicting the letter grade (as per the random forest)?

-e- What is the overall accuracy of the random forest on predicting the letter grades for the test set?

-f- Since the initial dataset is imbalanced (for example A is the most common grade by far), if you are ambitious you can play with the `classwt` parameter of the `randomForest` function to see if that might improve the accuracy.