MSA220: Determining classifier strengths

by Calvin Smith

Gothenburg University

April 2020

Data and classifiers

- Analyzing two different datasets.
- MedGPA.csv contains data on applicants for Medical School and whether they got accepted (A) or denied (D).
- Handwriting.csv is a dataset containing measurements of handwriting and if the person is Male or Female.
- Compairing three different classifiers.
- KNN, Logistic Regression and Linear Discriminant Analysis (LDA).
- Using k-fold cross validation and missclassification rate to determine strength.

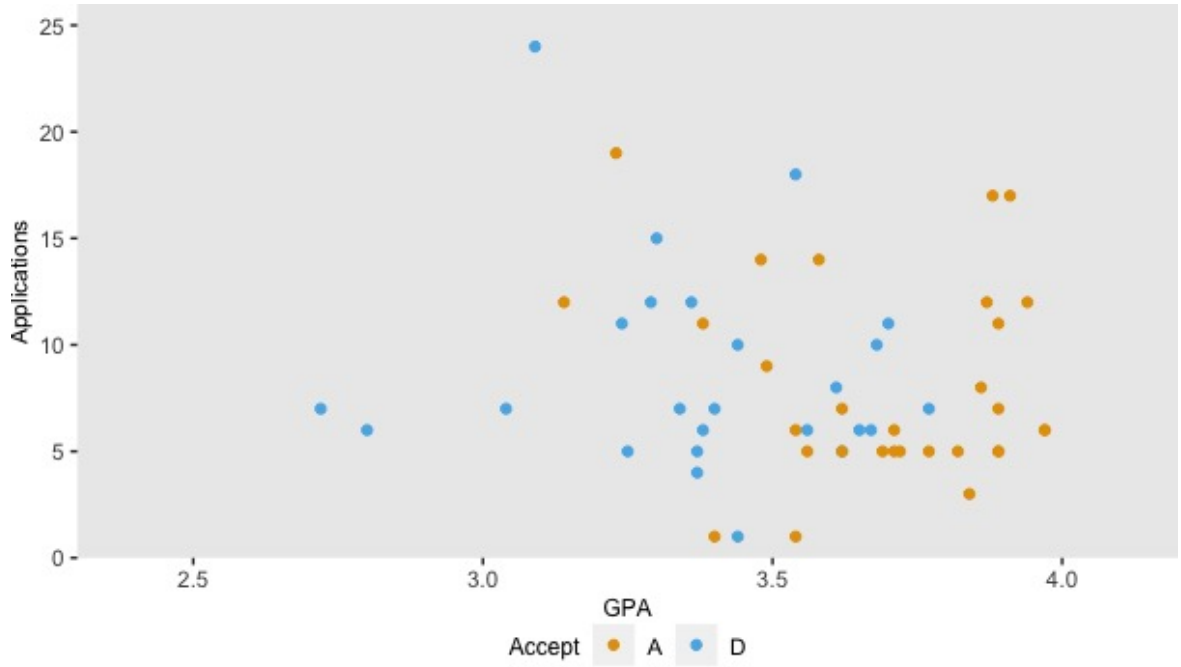
MedGPA.csv

Overview of the data

- Dataset contains 55 observations with 12 variables.
- Using variables GPA score,nr of applications (Apps) and the categorical response Accept.
- Want to predict if a person gets Accepted (A) or denied (D) into Medical School based on GPA and Apps.
- Using k $= 5$ folds.

MedGPA.csv

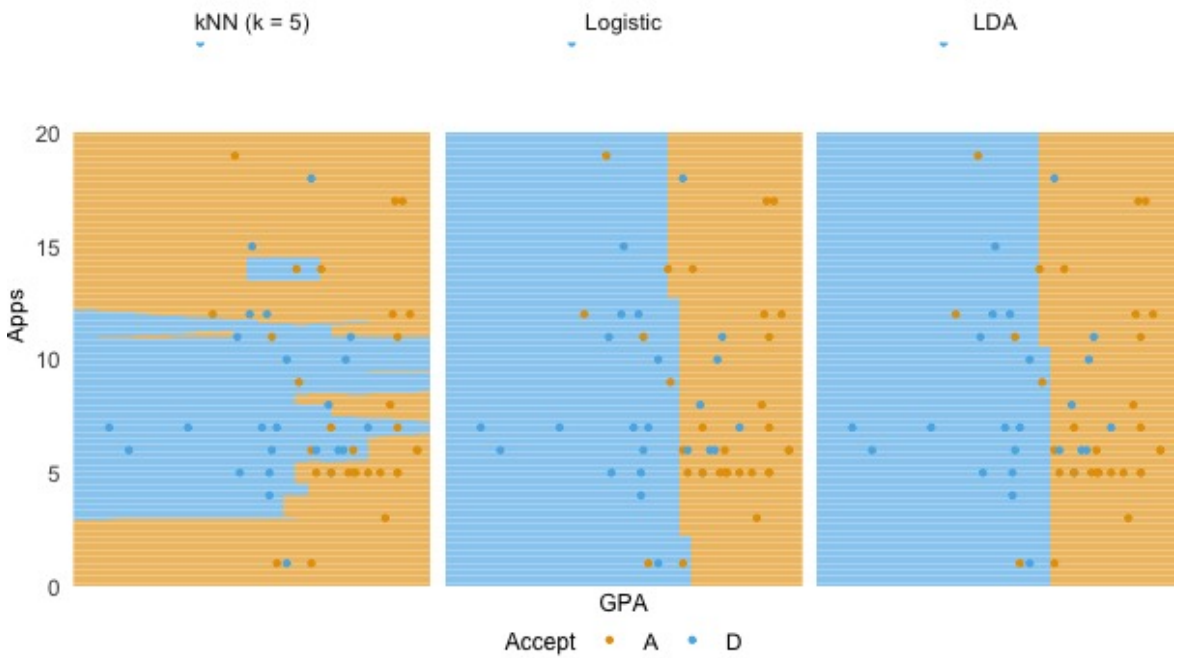
Overview of the data



- Plot shows (A) and (D) for GPA vs Apps.

- Using cross validation with 5 folds, and computing the mean missclassification rate over the folds:

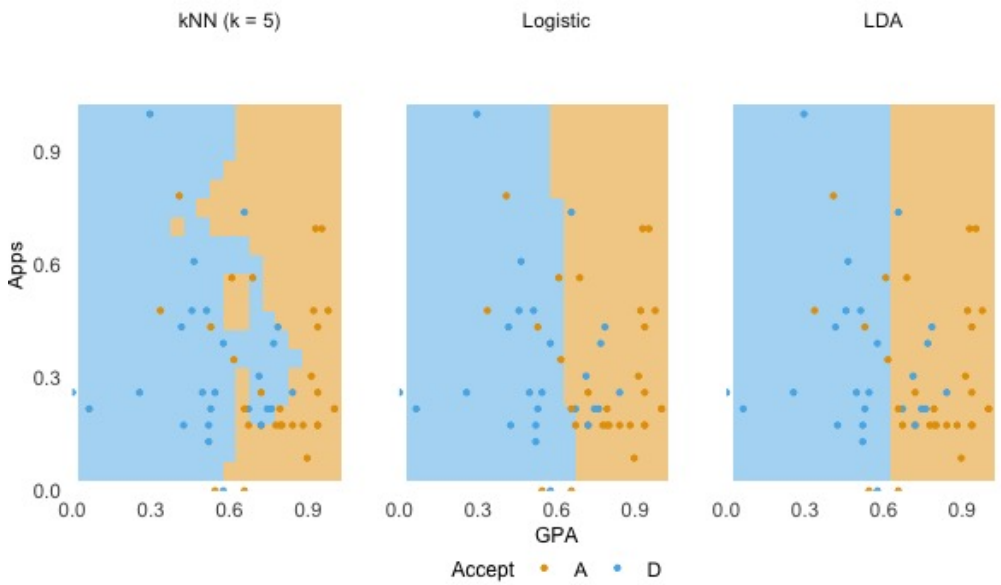
	logistic	knn(k=5)	lda
means	0.31	0.47	0.27
std	0.02	0.03	0.02



- Logistic regression and LDA perform similarly but KNN is comparably worse.
- The main assumptions of LDA and Logistic holds fairly well for the data. GPA and Apps are not collinear and it seems reasonable to assume that both are approximately normal and the variances of the errors seem pretty much constant (could not fit plots of this).
- KNN is a data-driven algorithm, and the data set is pretty small, which could account for the high error rate.
- I have not accounted for the ranges of GPA and Apps. GPA ranges from 2.4-4 and Apps from 1-25. If the variables are properly scaled (normalized) the KNN errors may improve.

- Same procedure but normalizing data to be between 0 and 1, produces the following results:

	logistic	knn(k=5)	lda
means	0.31	0.36	0.27
std	0.03	0.02	0.03



- Logistic and LDA are not effected, but KNN improves a lot!

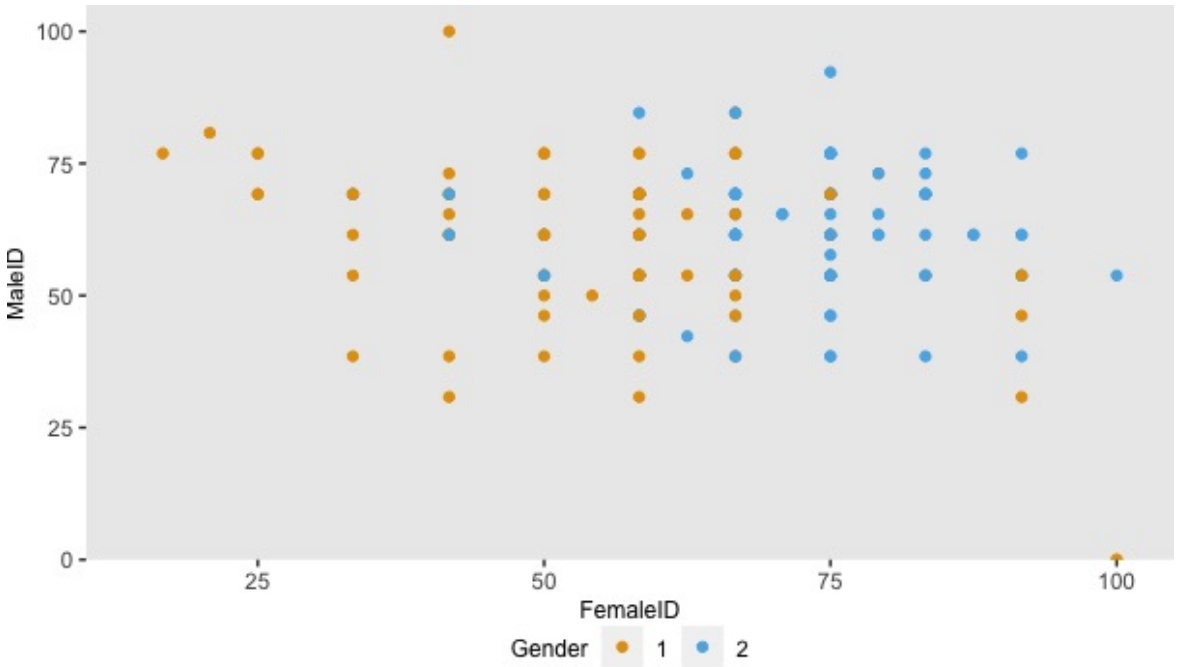
Handwriting.csv

Overview of the data

- 195 observations, 6 variables.
- Want to predict the gender (Male or Female) of a person based on 5 predictors (some sort of "scores" from different handwriting tests.)
- The response is coded as 1 for Male and 2 for Female.
- Once again using 5 folds for cross validation.

Handwriting.csv

Overview of the data



- Figure shows FemaleID vs MaleID with Male = 1 and Female = 2 for illustration purposes. However I am using 5 different predictors when fitting the models.

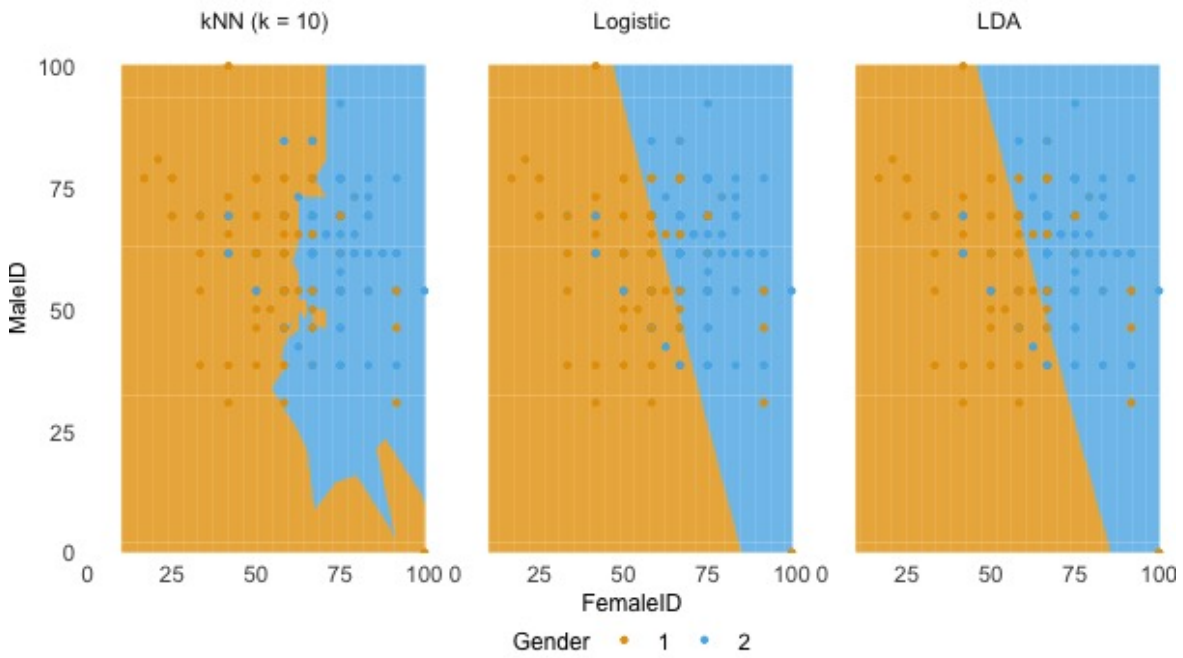
Handwriting.csv

Results

- Mean missclassifation rate over the 5 folds:

	logistic	knn(k=10)	lda
means	0.272	0.246	0.282
std	0.016	0.009	0.015

- Decision boundary plot:



Handwriting.csv

Results

- KNN algorithm performs best this time. However the difference is not huge. Once again, LDA and Logistic are similar.
- Why is KNN "best"?
- The variables in the dataset are in the same range and comparable with each other.
- If you look at the previous plot (FemaleID vs MaleID) there is a pretty clear distinction where classification goes from Male to Female, this could be benefitial to the KNN model, and from the decision boundary plot we can see that KNN is "straighter" in the middle compaired to LDA and Logistic which are not.
- However, KNN is worse at handling outliers and predicting "outside" of the data as we can see from the bottom right corner of the decision boundary plot. This can also serve as an example of the KNN algorithm having a high variance compaired to LDA and Logistic which seem more stable, perhaps with a higher bias.