Project 1 - number of classes

Viktor Erbro
Erik Forsberg
Markus Ingvarsson
Devosmita Chatterjee

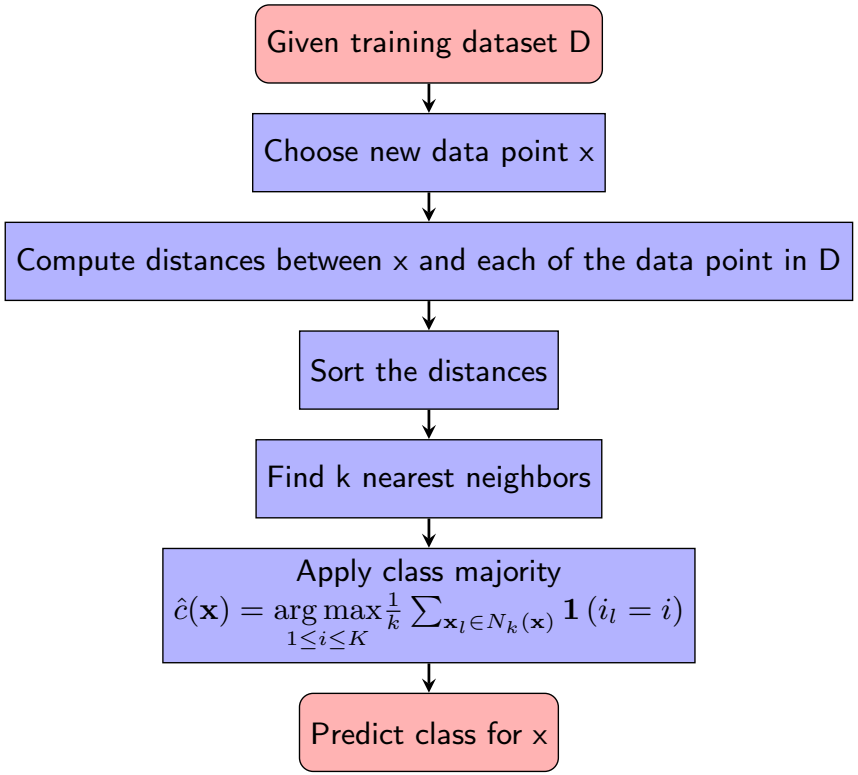
Chalmers University of Technology

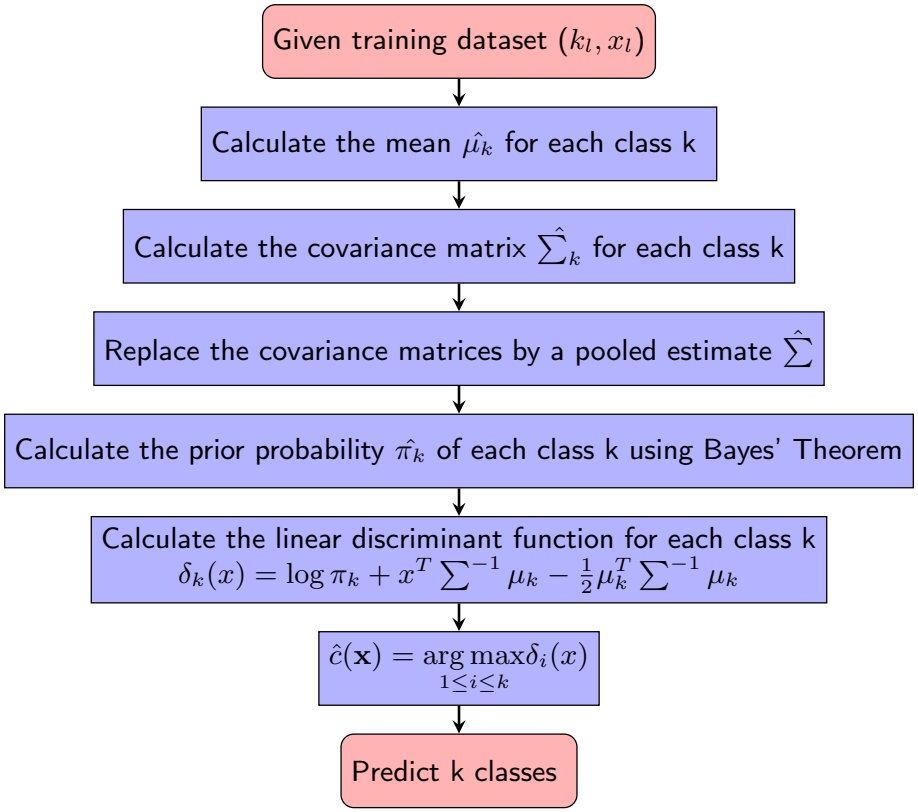
April 12, 2019

Classification methods in statistical learning

Classification Methods	k-Nearest Neighbors(kNN)	Linear Discriminant Analysis(LDA)
Assumption	Similar data have similar labels.	Gaussian data with same variance.
Idea	Classify a point by calculating the number of times each label is predicted by its k nearest neighbours.	Classify two or more classes of data where each class is predicted using the linear discriminant function.

Schematic Representation of kNN and LDA





The data set

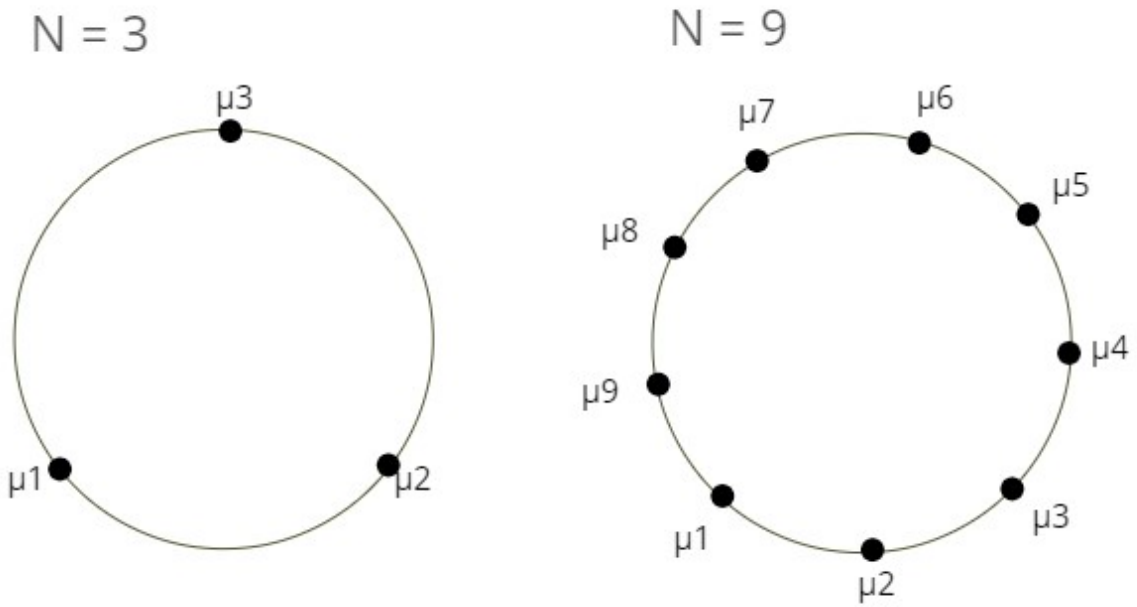


Figure: Placement of the distribution means of the classes.

Plots of classification boundaries

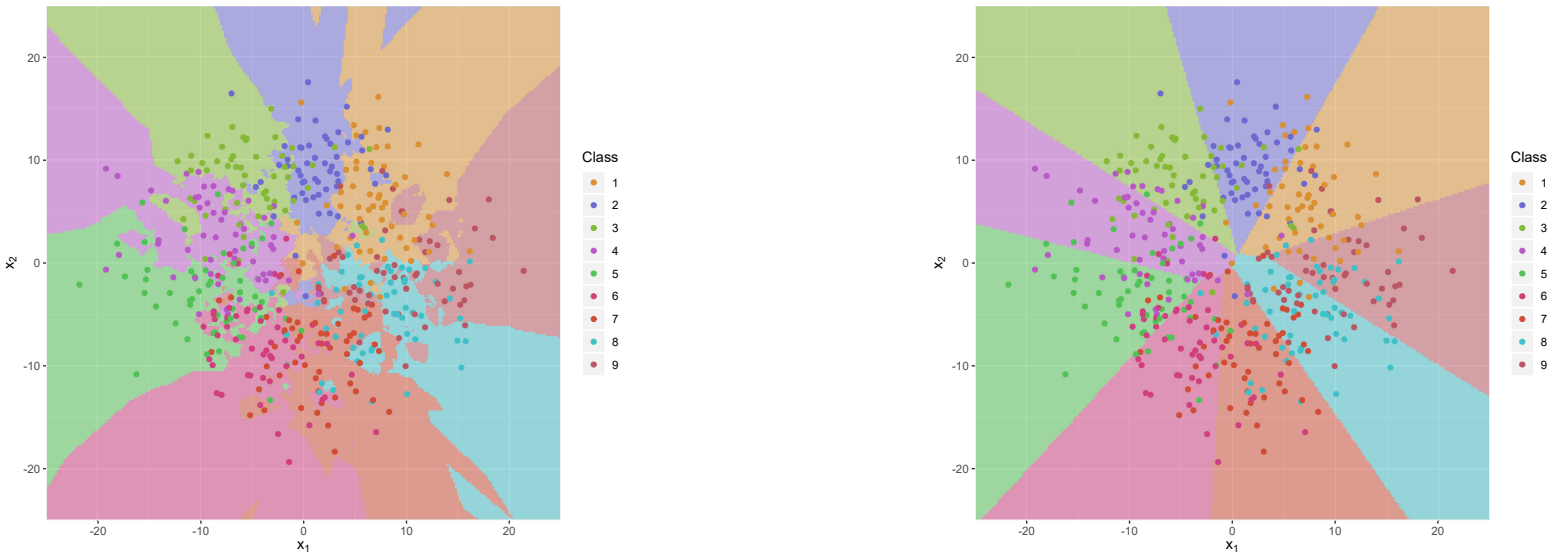
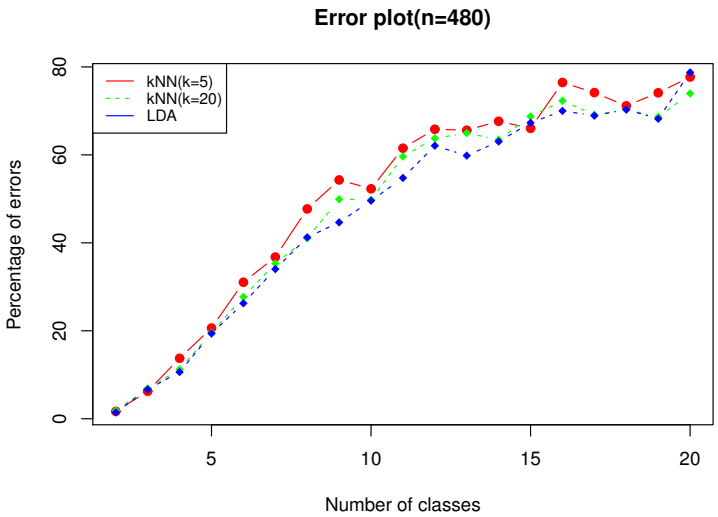
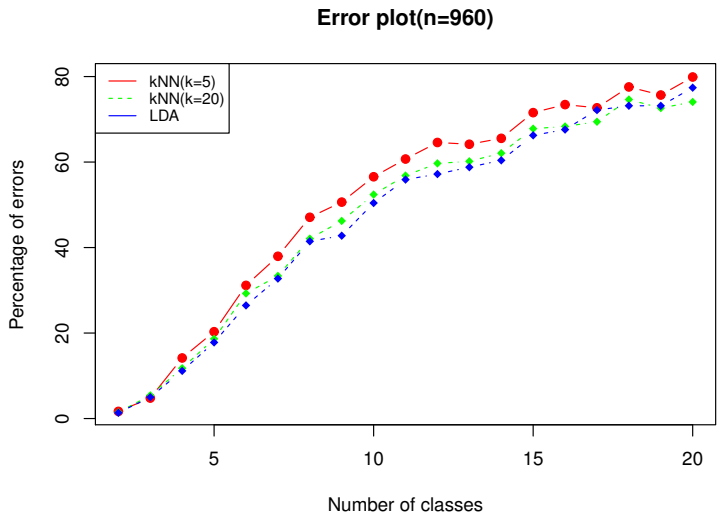


Figure: Left figure shows the decision boundaries for kNN while, right figure shows the decision boundaries for LDA.

Result: Classification accuracy with different sample sizes





Different geometry of the data set

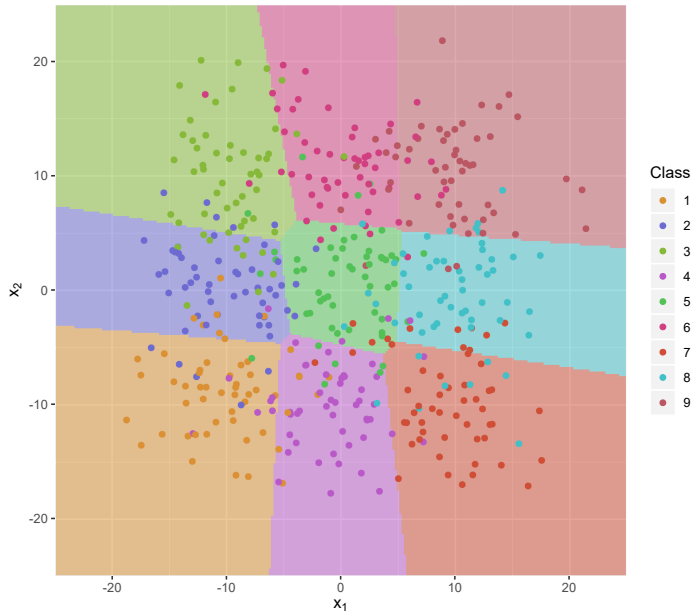


Figure: The means of the data put in a square lattice instead of a circle.

How tweaking parameters affects the result

Sample size	120	240	480	960	1920
KNN-5	58.3%	60.0%	53.1%	56.8%	55.0%
KNN-20	67.5%	59.2%	51.0%	52.8%	51.0%
LDA	65.0%	56.3%	51.7%	49.6%	49.3%

Table: Error rates for 10 classes with different sample sizes

	Original	New geometry	50% higher variance
KNN-5	53.1%	30.6%	69.2%
KNN-20	51.0%	26.2%	64.8%
LDA	51.7%	26.8%	66.0%

Table: Error rates for 10 classes with other parameters tweaked

Conclusions

- Both methods seems to preform way better for a small number of classes (as expected)
- LDA is consistently performing better than kNN20 and kNN5. This is probably an affect of the relatively simple boundaries of the data.
- The number of classes in a data set seems to have a bigger impact than the sample sizes and the method used.

Further possible investigations

- Test with real data to see how methods can handle many classes in a complex data set
- Random sizes of the classes
- More complex geometrical patterns of the classes
- Non-Euclidean distances for the kNN.