K-Means, K-Center, and KNN

John Ensley, Mohammed Mottahedi & Songshan Yang

Penn State University
STAT 557

March 16, 2016

Data Description

- 699 tumor samples
- 458 classified as benign, 241 as malignant
- 9 observations for each: uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, mitoses
- Each predictor is a value from 1 to 10

K-Means

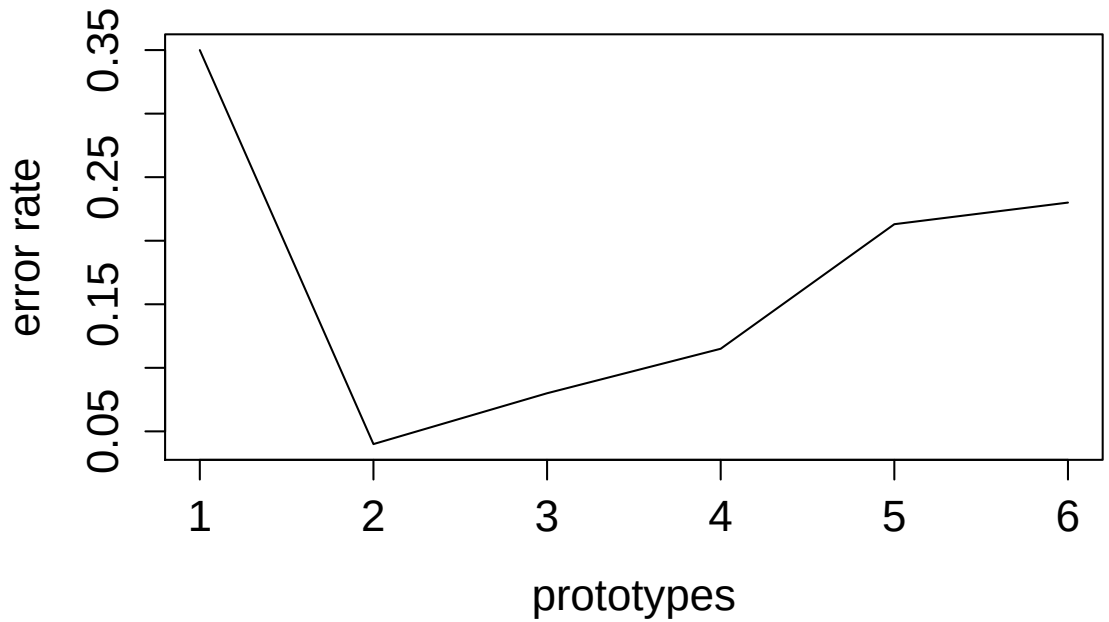
Implemented in two ways:

- Split data set randomly into 50% training, 50% test
- 5-fold cross validation

Compared error rates over varying number of prototypes

Number of prototypes	1	2	3	4	5	6
Error rates, training set	0.35	0.0453	0.086	0.135	0.205	0.234
Error rates, 5-fold CV	0.35	0.0424	0.083	0.116	0.213	0.227

K-Means

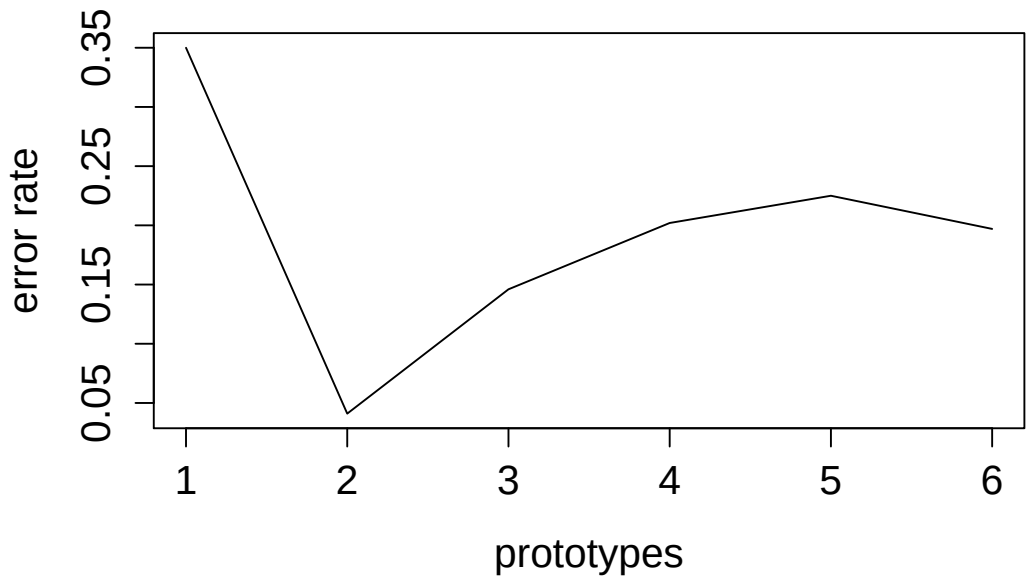


Optimal number of prototypes is 2.

K-Means with Dimension Reduction

Used first two principal components, accounting for 74% of the total variance. Again 2 prototypes optimal. Classification is not improved with this data.

Number of prototypes	1	2	3	4	5	6
Error rates, 2 PCs	0.35	0.041	0.146	0.202	0.2225	0.197

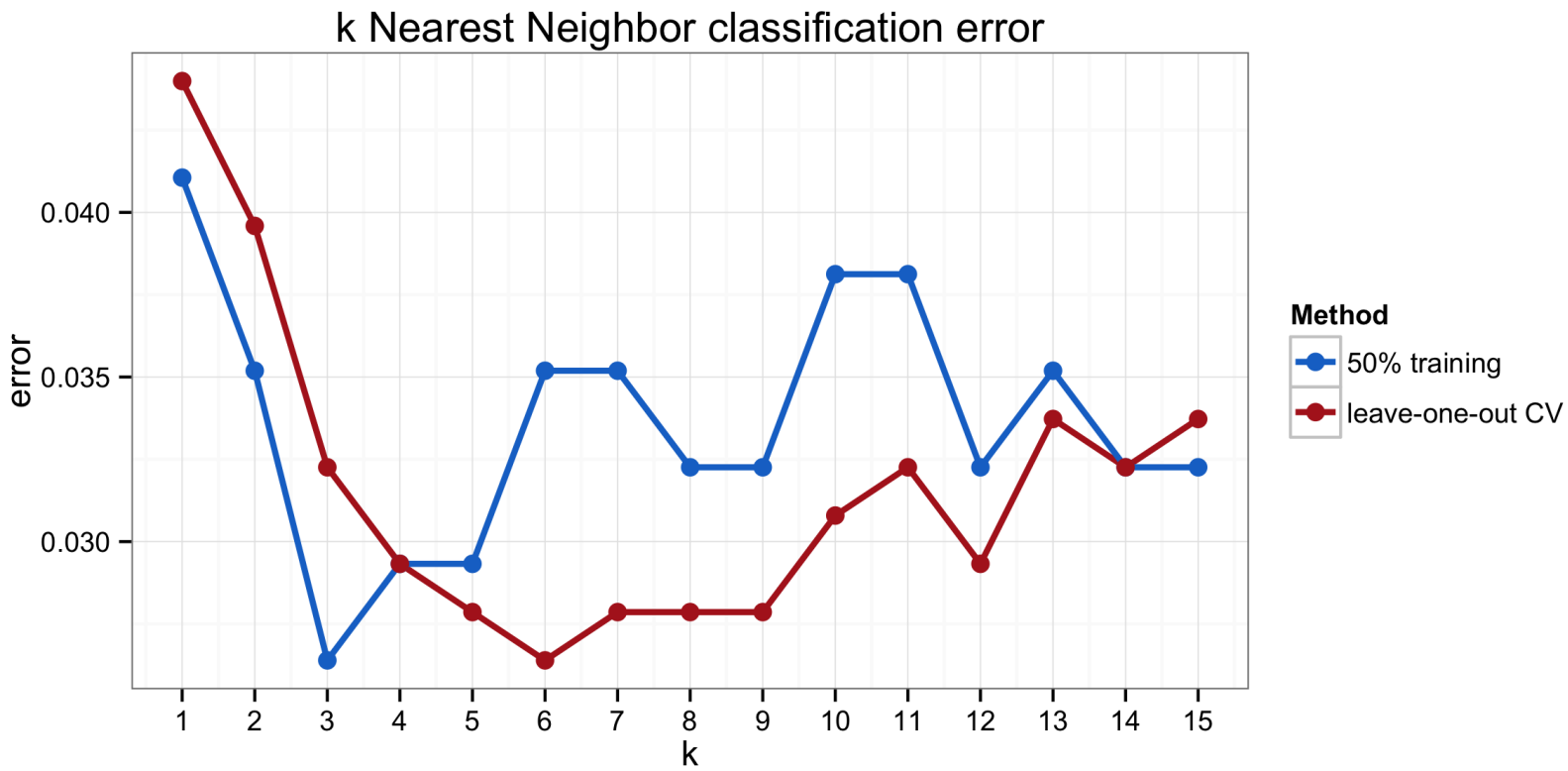


K-Nearest Neighbor

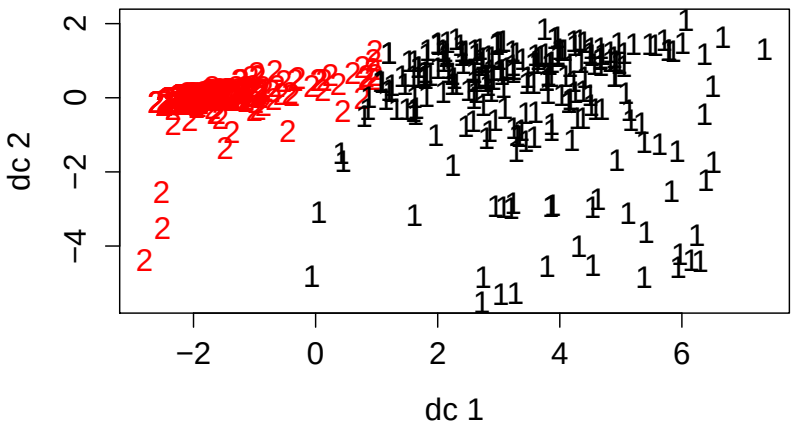
Implemented in two ways:

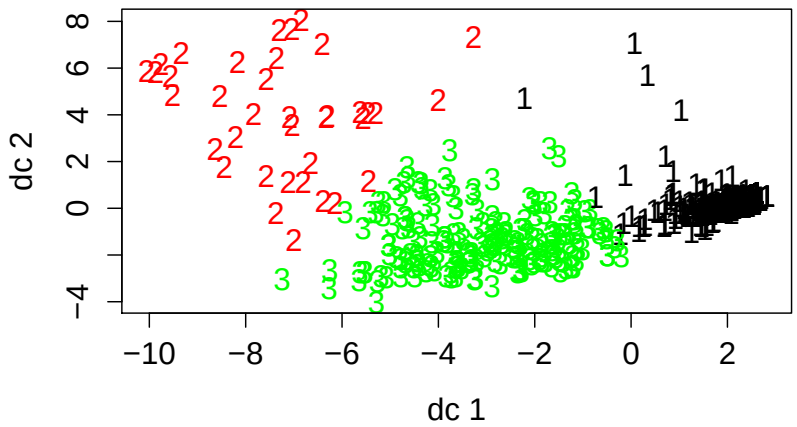
- Split data set randomly into 50% training, 50% test
- Leave-one-out cross validation

Compared error rates across different values of k

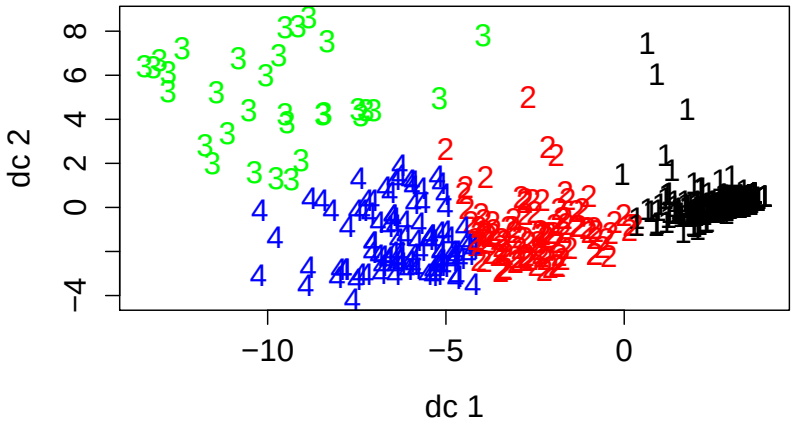


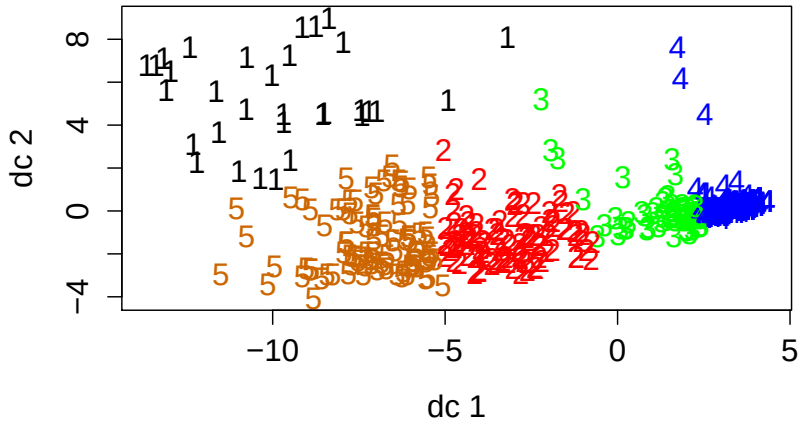
Without considering class labels we try several different numbers of clusters $(2-6)$ for the data set.

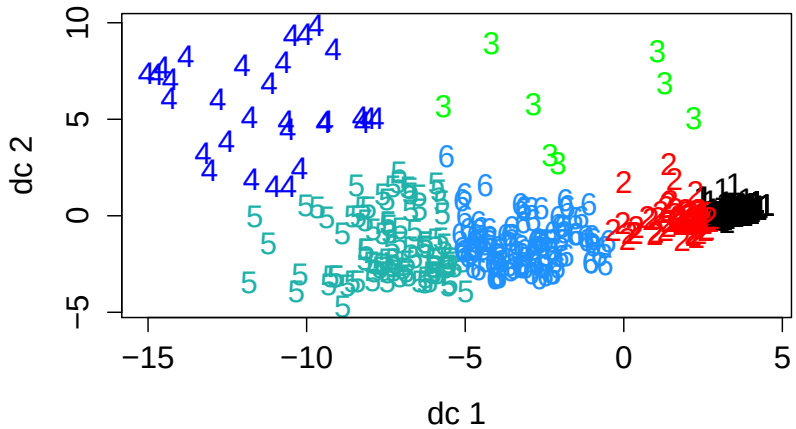




Unsupervised Clustering
K-means







We try to minimize the maximum inter-cluster distance. Ignoring class labels:

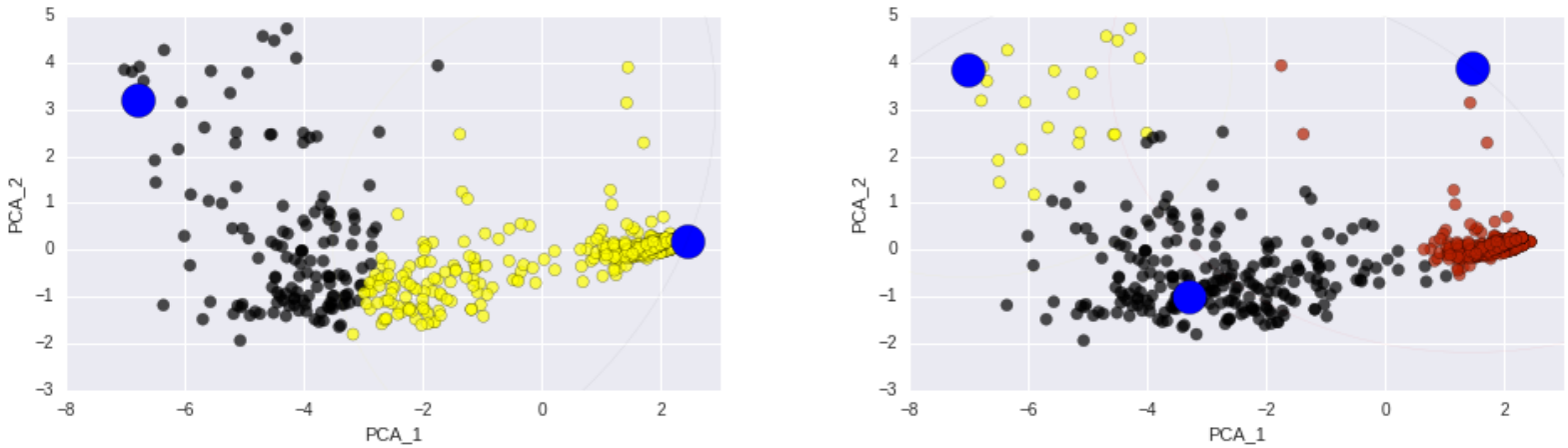


Figure : K-center unsupervised clustering results

Unsupervised Clustering
K-centers

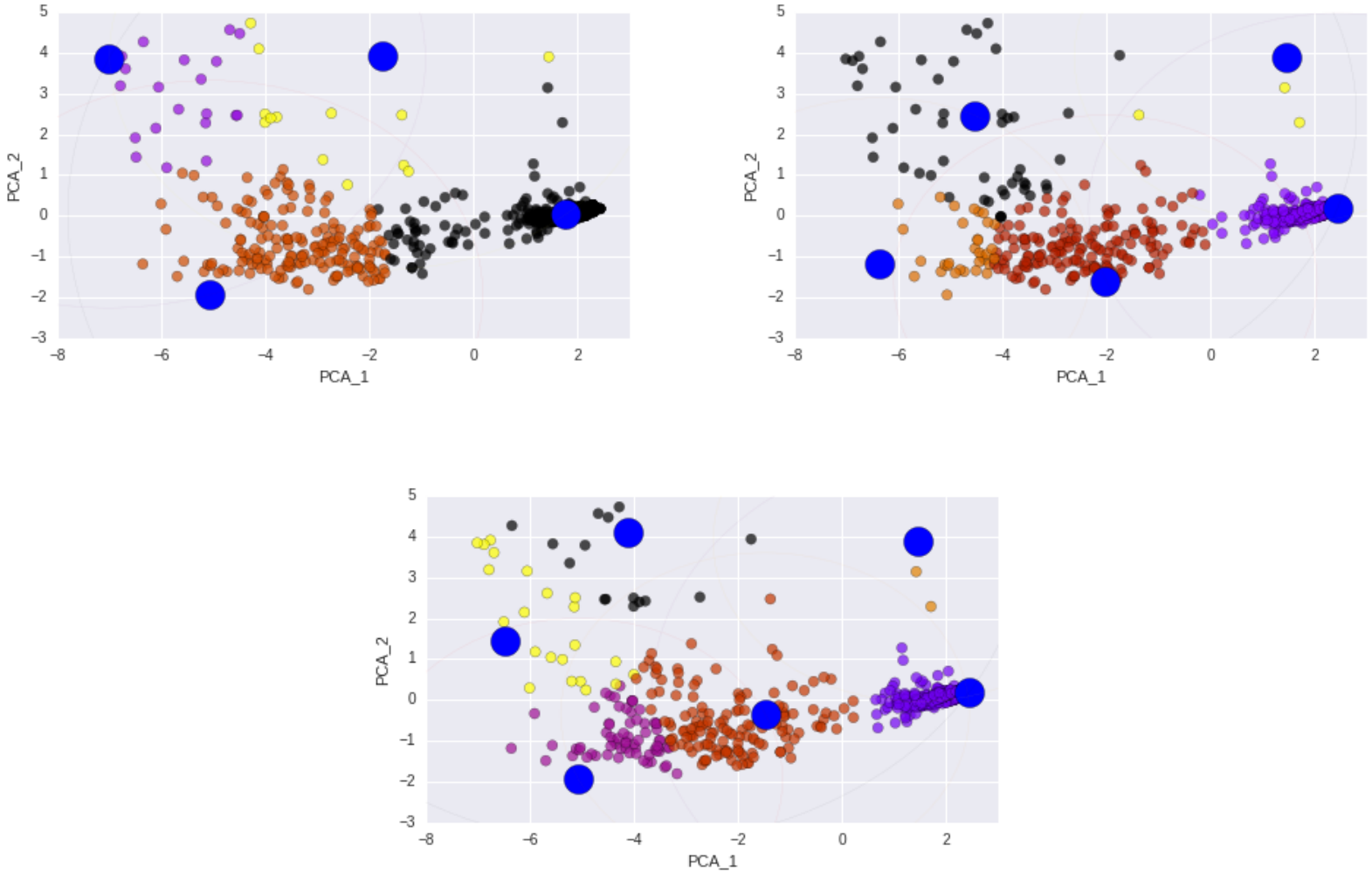
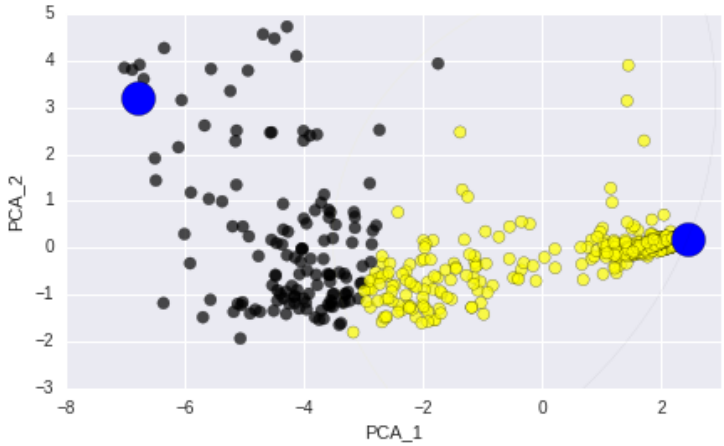
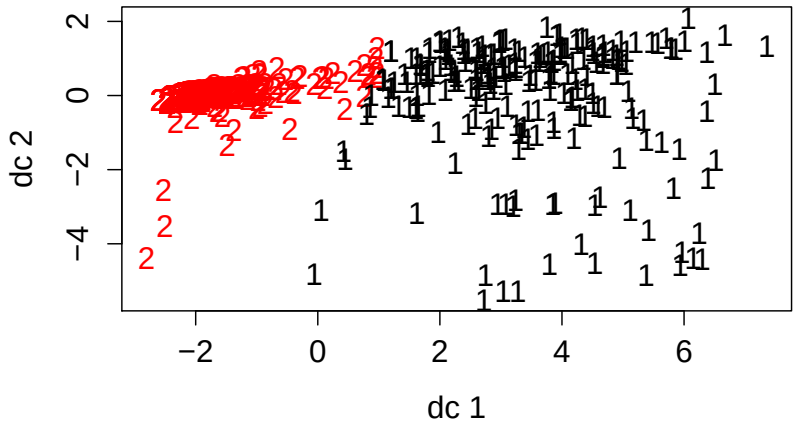


Figure : K-center unsupervised clustering results

Unsupervised Clustering
k-means vs k-centers

K-center focuses on worst case scenario and k-means focuses on average distance.





Questions?