1 Introduction of the data set

In this project, we applied the linear classification methods and QDA to a breast cancer data set. This breast cancer databases was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. There are 699 subjects in the data set and all the subjects are classified to 2 different classes–"Benign" and "Malignant". There are 9 features include clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell Size, bare nuclei, bland chromatin, normal nucleoli, mitoses and all of them range from 1 to 10.

2 K-means

2.1 The effect of the number of prototypes

In the first step, we randomly split the data set into two samples with equal size. We apply the k-means algorithm to the training data set with different number of prototypes. Then we classify the feature X's in the training data set and obtain the error rates of the k-means method when we use different prototypes. The results are summarized in Table 1.

Table 1: Classification error rate versus number of prototypes

Number of prototypes	1	2	3	4	5	6
Error Rates	0.35	0.0453	0.086	0.135	0.205	0.234

We can see that when we only use two prototypes, the classification error rate is the lowest. This is because in the original data set, the points of the two classes are very separated. In addition, the classification error rate does not change too much after we use more than 4 prototypes. The trend can be shown in Figure 1.

Then we use 5-fold cross validation to further examine the relationship between the classification error rate and the number of prototypes. The results are summarized in Table 2: From Table 2, we know that the optimal choice of the number of the prototypes should

Table 2: Classification error rate versus number of prototypes using cross validation

Number of prototypes	1	2	3	4	5	6
Error Rates	0.35	0.0424	0.083	0.116	0.213	0.227

be 2.

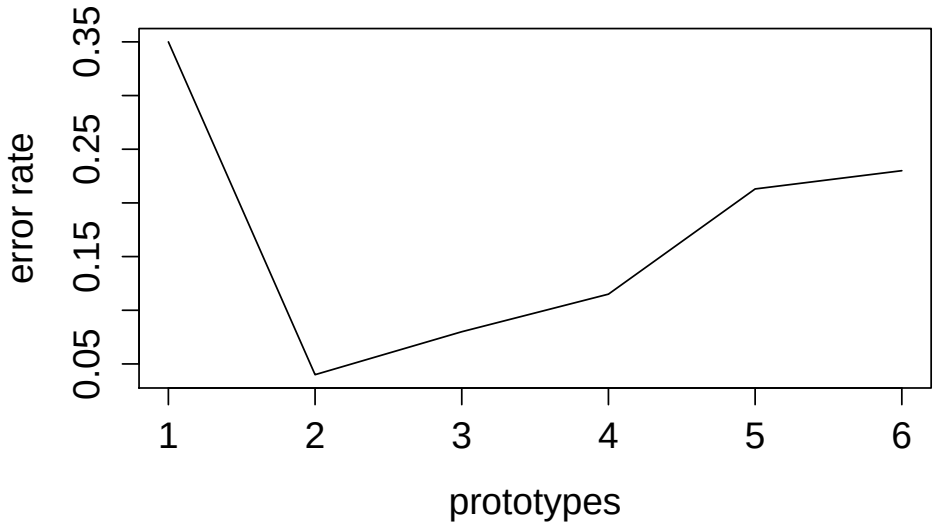

Figure 1: Error rates versus number of prototypes

3 K-means with dimension reduction

In this section, we examined the effect of dimension reduction on classification. We first standardized the design matrix and did eigen decomposition of the covariance matrix of the standardized design matrix. The ratio of each principle component explain the total variance is summarized in Table 3. We decided to use the first two components which account for

Table 3: Contributions of each component

Ratio	0.655	0.086	0.060	0.051	0.042	0.034	0.033	0.029	0.010

74% of the total variance. Then the design matrix only has two columns. By repeating the process described in section 2, the classification error rates are summarized in Table 4. Thus, for the data set we used in the project, dimension reduction does not improve the classification error rate. Maybe it is because the two classes of samples in the breast cancer data set are well separated. The relationship between error rates and number of the prototypes is shown in Figure 2.

Table 4: Classification error rate using dimension reduction

Number of prototypes	1	2	3	4	5	6
Error Rates	0.35	0.041	0.146	0202	0.2225	0.197

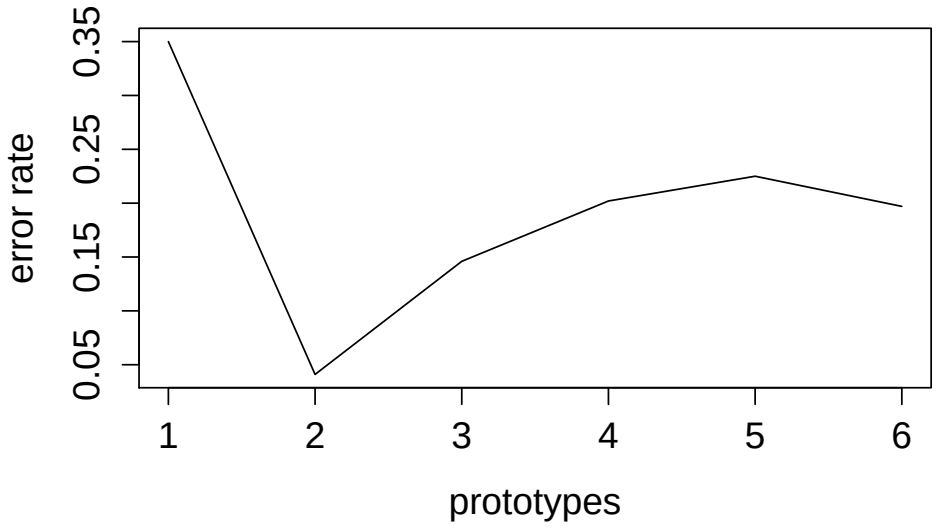

Figure 2: Error rates versus number of prototypes

4 k-nearest neighbor classification

We used the k-nearest neighbor method to classify the data as well. We randomly split the data into equal-sized training and test sets, and we also tried leave-one-out cross-validation as well. Figure 3 compares the error rates for each of these methods for $k = 1, \ldots, 15$ neighbors.

As the figure shows, $k = 3$ performed the best on the split data, and $k = 6$ was best according to cross-validation. It appears that cross-validation performed better over a wider range of values of k; anything from $k = 5$ to 9 had comparably low error rates, and they were all lower than most error rates using the split data set. For both methods, though, the error rates were highest at $k = 1$, then dropped sharply before steadily rising again as k increased.

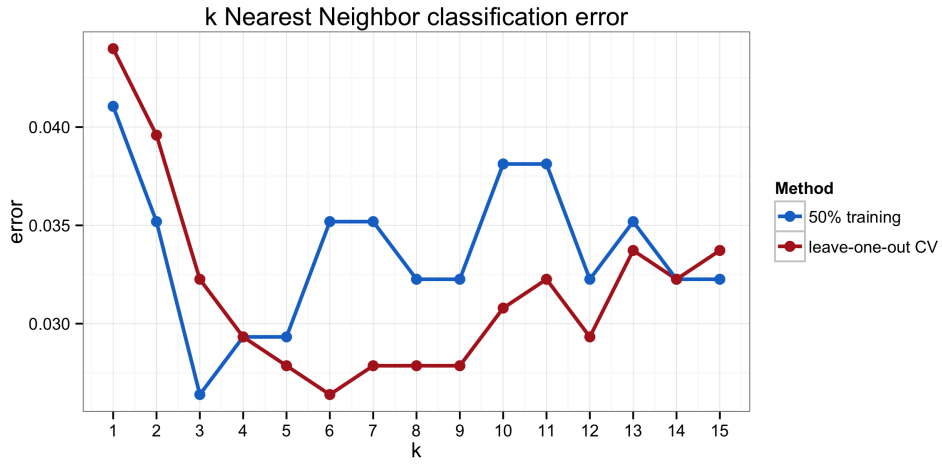

Figure 3: k nearest neighbor classification results.

5 Unsupervised clustering

In this section, we try several different numbers of clusters for the data set. We use 2 to 6 centroids and all the results are shown in Figure 4.

6 K-Center

In k-center clustering we are seeking minimize the maximum inter-cluster distance. In a plane, were seeking to find the smallest radius such that all point in the fix space are contained in k circles with radius r. The k-center problem is NP-hard and the algorithm proposed by Gonzalez guarantees to produce clustering whose cost is at most twice the optimal.

6.1 Unsupervised clustering using K-center

In this section we perform unsupervised clustering by dividing the data into 2 to 6 clusters and the result is shown in Figure 6.

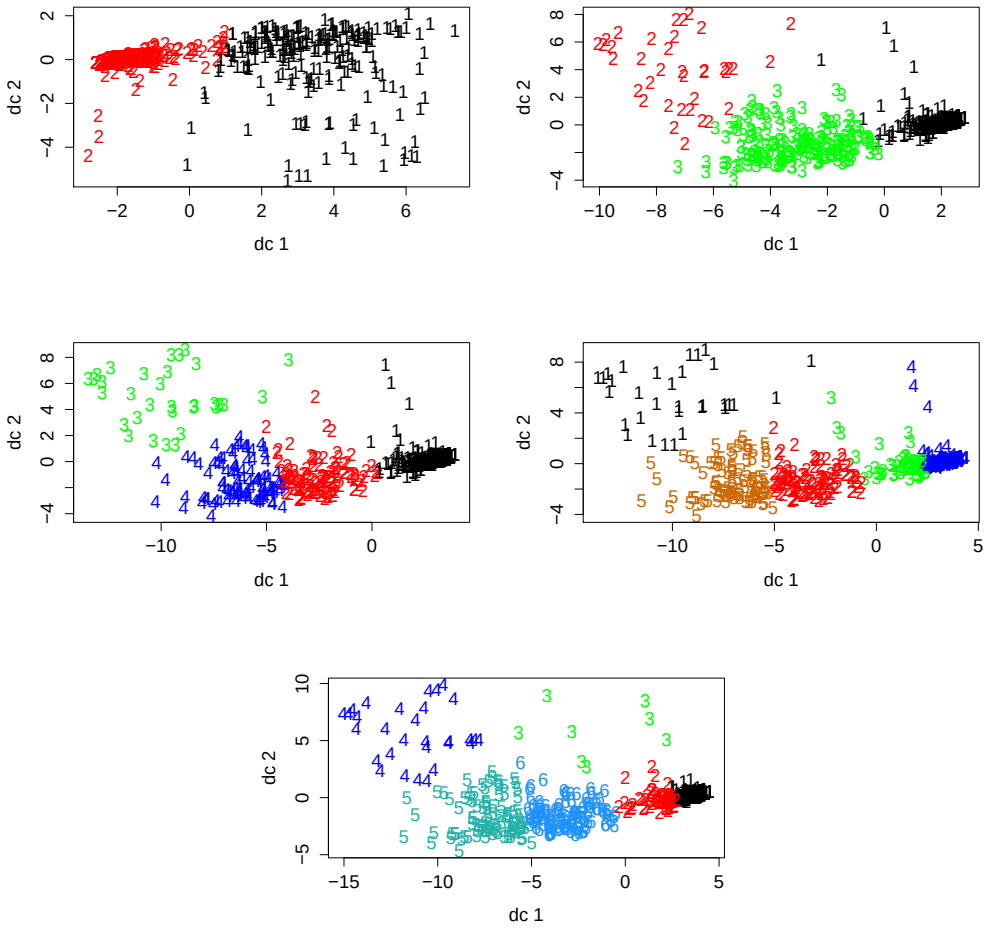

Figure 4: Unsupervised clustering results

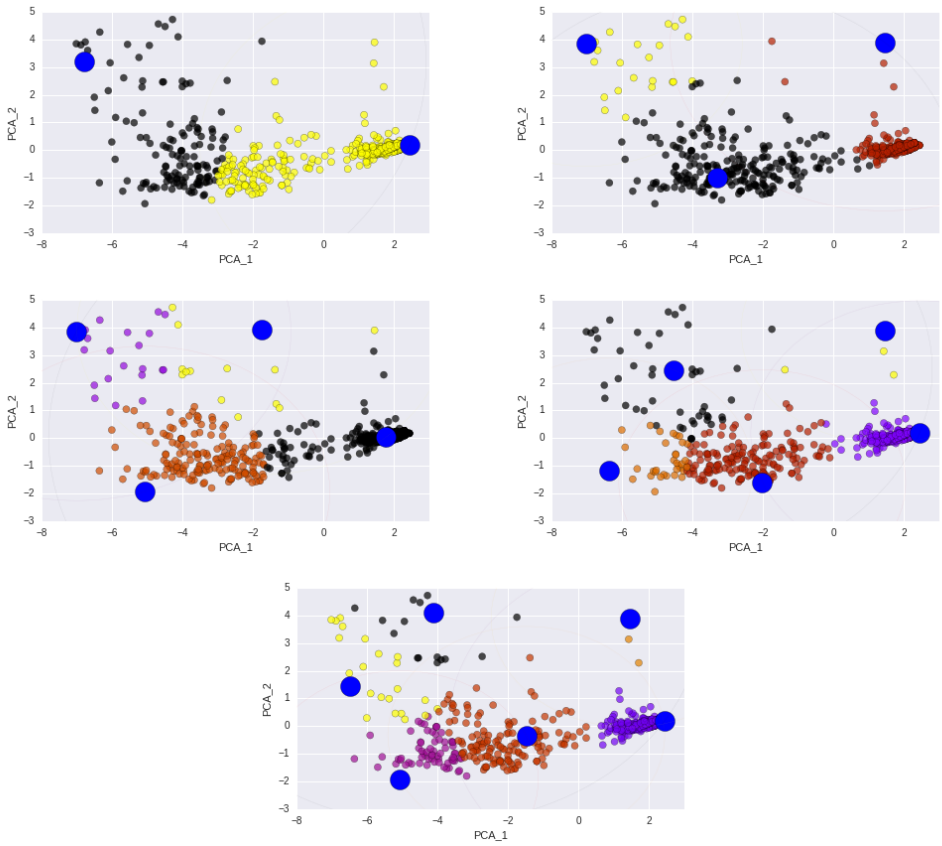

Figure 5: K-center unsupervised clustering results