Writing assignment 4 for machine learning

Kyrie

崔颖琦

1730004002

Question 1: Support Vector Machines

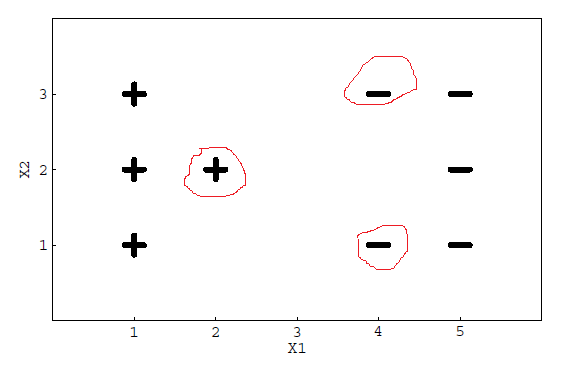
1. Suppose we are using a linear SVM (i.e., no kernel), with some large C value, and are given the following data set. Draw the decision boundary of linear SVM. Give a brief explanation.

**Answer: Find the support vectors of two different classes. (the dotted line in the graph). And then find a plane which is far away from the support vectors.**



1. In the following image, circle the points such that by removing that example from the training set and retraining SVM, we would get a different decision boundary than training on the full sample. You do not need to provide a formal proof, but give a one or two sentence explanation.

**Answer: These points are the support vectors. If we removing one of them, the support vector will be change. So SVM need to fit a new plane which is far away from the support vectors.**



Question 2: K-means

1. Consider the unlabeled two-dimensional data represented in the following figure. Using the two points marqued as squares as initial centroids, draw (on that same figure) the clusters obtained after one iteration of the k-means algorithm (k = 2).

**Answer:**

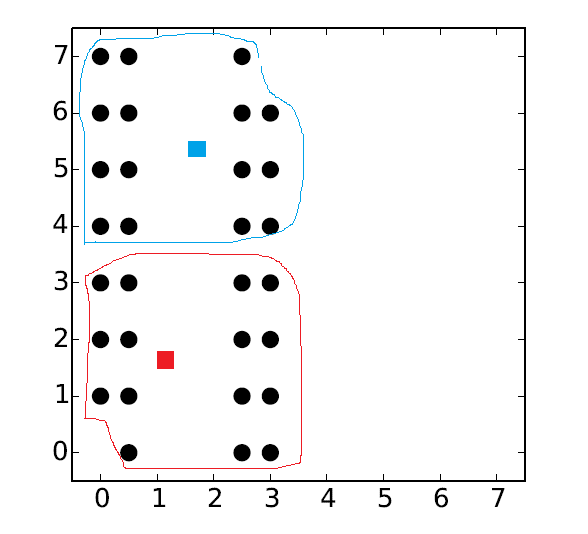
**Start from the first row, from left to right, the point from point1 to point m.**

**c1=1, c2=1, c3=1, c4=1, c5=1, c6=1, c7=1, c8=1, c9=1, c10=1, c11=1, c12=1, c13=1, c14=1, c15=2, c16=1, c17=2, c18=2, c19=2, c20=2, c21=2, c22=2, c23=2, c24=2, c25=2, c26=2, c27=2, c28=2, c29=2, c30=2**

**1.4,1.7)**

**5.3)**

**So the new centroids are (1.4,1.7) and (1.6,5.3).**



1. Does your solution change after another iteration of the k-means algorithm? Why?

**Answer: Yes. Because before the first iteration, the point (3,3) belongs to cluster 2, and point (0,4) belongs to cluster 1. After the first iteration, the cluster centroids are changed. So in the next iteration, the mean of points assigned to cluster k will be change and the centroids will be change again.**

Question 3: k-means

Consider the following clustering method called Leader Clustering. It receives two parameters: an integer k and a real number t. Similar to k-means, it starts by selecting k instances (which will be called leaders) and assigns each training instances to the cluster of the closest leader. During the assignment step, however, if the distance of a training instance to its closest leader is greater than the input threshold t, then this training instance becomes a new leader. During the same assignment step, remaining points can be assigned to these new leaders. After all the training instances have been assigned to a leader’s cluster, new leaders are calculated by averaging each cluster. The process is then repeated until the cluster assignments do not change.

1. Given a dataset and a value k, let t vary from 0 to a very large value. When does Leader Clustering produce more, the same number, or fewer clusters than k-means, assuming that the k initial centers are the same for both? When will the clusterings produced by Leader Clustering and k-means be identical?

**Answer: When t is small, Leader Cluster will produce more cluster than k-means. When t is large, Leader Cluster will produce the same number with the k-means.**

1. Which of the two methods, k-means or Leader Clustering, will be best at dealing with outliers (data instances that are ”far away” or very different to the other instances in the dataset)? Explain.

**Answer: Leader clustering is best at dealing with outliers. For example, if k=2. If there is a point very far away from other points. In k-means, the outlier point will be assigned to the cluster which its centroid is closer than the other. And it will cause the centroid moves to the outlier point. That will make the distance of the points which is close to it become larger. And this points probably are assigned the other cluster because its centroid moves to the outlier point. It will make mistake. But in leader clustering, the distance between the point which is far away from others and the centroid will larger then threshold t, so the outlier point will be the new leader and it will not influence other clusters. That’s the reason why leader clustering is best at dealing with outliers.**

Question 4:

Check all the binary classifiers that are able to correctly separate the training data (circles vs. triangles) given in the following figure. Explain why.

* Logistic regression

**Answer: No. logistic regression is linear model. It can’t solve the XOR problem.**

* SVM with linear kernel

**Answer: No, because linear model cannot separate XOR problem. We can’t use a straight line to separate the points in the graph.**

* SVM with RBF kernel

**Answer: Yes. SVM with RBF kernel are not linear model. It calculate the distance between the point and the kernel to classify.**

* Two layer neural network

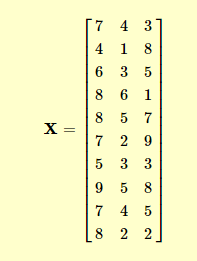
**Answer: No. Two layer neural network is like logistic regression. It’s linear model. It can’t solve the XOR problem.**

* Three layer neural network

**Answer: Yes, XOR = (x1 or x2) AND Not (x1 AND x2). It can be done by three layer neural networ**k.



Question 5:  
Let us analyze the following 3-variate dataset with 10 observations.



1. Do the mean normalization.

**Answer:**

**Because in the matrix X, the number of row represents the number of example. The column means the features.**

**The mean of each column means=**

**Subtract the means**

**X’-means’**

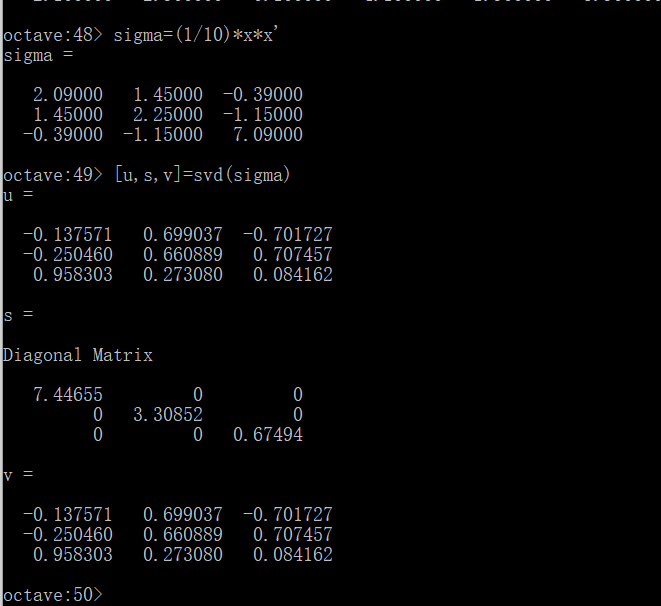
**=**

1. Compute the Covariance matrix of the dataset.

**Answer: Sigma=**

1. Do the SVD composition and compute the S, U, V.

**Answer: [u, s, v]=svd(Sigma)**



1. Compute the new features for vector [3, 4, 5] if we only want to keep the first two principal components.

**Answer: V= [3,4,5], V’=V-means=[-3.9 0.5 -0.1]**

**Because keep the first two principal components. U=[u(:,1),u(:,2)]=**

**The new features for vector [3, 4, 5]**

**W=**

Question 6:

Describe the steps when we use PCA to help SVM to do classification.

**Answer:**

1. **Get training data**
2. **Run PCA to reduce to dimension to get**
3. **And then get the new training set**
4. **Training SVM on**
5. **Test on the test set: Map . Run the test set on the fitting model.**