3333 Assignment 4

Ravish Kamath: 213893664

01 October, 2022

```
## randomForest 4.7-1
## Type rfNews() to see new features/changes/bug fixes.
```

Question 1

The following figure shows a neural network with two inputs, one hidden layer with two hidden neurons and one output. (For simplicity, we omit the intercept terms here). We initialize the parameters as follows: $w_{11} = 0.1$ $w_{12} = 0.4$, $w_{21} = -0.1$, $w_{22} = -0.1$,$v_{11} = 0.06$, $v_{12} = -0.4$. Given one observation $x_1 = 1$, and $x_2 = 0$, and the observed ouput $t_1 = 0$, update the network parameter w_{11}, using the learning rate $\lambda = 0.01$.

Solution

$$\begin{aligned} net_1 &= 0.1(x_1) + 0.4(x_2) \\ &= 0.1(1) + 0.4(0) \\ &= 0.1 \end{aligned} \qquad \begin{aligned} y_1 = f(net_1) &= \frac{e^{0.1}}{1+e^{0.1}} = 0.52 \\ f'(net_1) &= 0.52(1-0.52) \\ &= 0.25 \end{aligned}$$

$$\begin{aligned} net_1 &= -0.1(x_1) - 0.1(x_2) \\ &= -0.1(1) + (-0.1)(0) \\ &= -0.1 \end{aligned} \qquad \begin{aligned} y_1 = f(net_2) &= \frac{e^{-0.1}}{1+e^{-0.1}} = 0.48 \\ f'(net_2) &= 0.48(1-0.48) \\ &= 0.25 \end{aligned}$$

$$\begin{aligned} net_1^* &= 0.06(y_1) + (-0.4)(y_2) \\ &= 0.06(0.52) + (-0.4)(0.48) \\ &= -0.16 \end{aligned} \qquad \begin{aligned} f'(net_1*) &= 0.46(1-0.46) \\ &= 0.25 \end{aligned}$$

$$\begin{aligned} error = t_1 - z_1 &= 0 - 0.46 \\ &= -0.46 \end{aligned}$$

$$\begin{aligned} \frac{\partial J}{\partial v_{11}} &= -(t_1 - z_1) f'(net_1^*) y_1 \\ &= -(-0.46)(0.25)(0.52) \\ &= 0.0598 \end{aligned} \qquad \begin{aligned} \frac{\partial J}{\partial w_{11}} &= -(t_1 - z_1) f'(net_1^*) v_{11} f'(net_1) x_1 \\ &= -(-0.46)(0.25)(0.06)(0.25)(1) \\ &= 0.001725 \end{aligned}$$

$$\begin{aligned} v_{11}^{(new)} &= v_{11}^{(old)} - \lambda \frac{\partial J}{\partial v_{11}} \\ &= 0.06 - (0.01)(0.0598) \\ &= 0.06 \end{aligned}$$

$$\begin{aligned} w_{11}^{(new)} &= w_{11}^{(old)} - \lambda \frac{\partial J}{\partial w_{11}} \\ &= 0.1 - (0.01)(0.001725) \\ &= \underline{\mathbf{0.09998275}} \end{aligned}$$

Question 2

Principal compnent method can be used to summarize the data in a lower dimension. Suppose each observation in the data set X_i has two features X_{i1}, and X_{i2}. We wish to use the principal compnent method to present the data in one dimension space. We have the following data set.

$$X = \begin{pmatrix} -3 & 6 \\ -6 & 6 \\ -8 & 3.5 \\ -7 & 6 \\ -7 & 5 \\ -9 & 6 \end{pmatrix}$$

Calculate the first principal component for the first observation.

Solution

```
X = c(-3, -6, -8, -7, -7, -9, 6, 6, 3.5, 6, 5, 6)
X = matrix(X, ncol = 2)
Xbar = t(colMeans(X))
Xbar
```

```
##           [,1]     [,2]
## [1,] -6.666667 5.416667
```

```
Xstar = apply(X, 2, scale, scale=FALSE, center=TRUE)
Xstar
```

```
##            [,1]       [,2]
## [1,]  3.6666667  0.5833333
## [2,]  0.6666667  0.5833333
## [3,] -1.3333333 -1.9166667
## [4,] -0.3333333  0.5833333
## [5,] -0.3333333 -0.4166667
## [6,] -2.3333333  0.5833333
```

```
covmat = cov(Xstar)
eigen_v = eigen(covmat)
w = eigen_v$vectors
w
```

```
##            [,1]       [,2]
## [1,] -0.9773142  0.2117948
## [2,] -0.2117948 -0.9773142
```

In the first column egienvector, we have the largest eigenvalue to be **4.4255881** vs the 2nd column has the eigenvalue 0.8827452. We will choose the eigenvector witht the largest eigenvalue since it will have the largest variance.

```
y = w[,1]%*%Xstar[1,]
y
```

```
##           [,1]
## [1,] -3.707032
```

As we can see from the above code, we have calculated the first principal component for the first observation to be **-3.707032**.

Question 3

ID3(S,A) is an important algorithm in the construction of decision tree.The set S denote the collection of observations. The set A denote the collection of predictors. In this question, let $A = \{X_1, X_2\}$. Let S be the following data set:

$$S = \begin{pmatrix} Y & X_1 & X_2 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix}$$

We would like to build a classification tree for the response variable Y.

- What is the misclassification error rate if we do a majority vote for Y without splitting X_1 or X_2?
- What is the misclassification error rate if we split the data set based on $X_1 = 1$ versus $X_1 = 0$? What is the misclassification error rate if we split the data set based on $X_2 = 1$ versus $X_2 = 0$?
- Should we split the tree based on the predictor X_1 or X_2 or not split the tree?
- Decision tree is very sensitive to the data set. If there are small changes in the data set, the resulting tree can be very different. Ensemble method can overcome this problem and improve the performance of the decision tree? Use two or three sentences to describe what ensemble method is and name three ensemble methods that can used to improve decision trees.

Solution

- $$\begin{aligned} Error &= C(P_s(y=1)) \\ &= C(\frac{3}{5}) \\ &= \frac{2}{5} \end{aligned}$$
- $$\begin{aligned} Error &= P_{S_1}(X_1=1)C(P_{S_1}(Y=1|X_1=1)) + P_{S_2}(X_1=0)C(P_{S_2}(Y=1|X_1=1)) \\ &= \frac{2}{5}\cdot C(\frac{2}{2}) + \frac{3}{5}\cdot C(\frac{1}{3}) \\ &= \frac{2}{5}\cdot 0 + \frac{3}{5}\cdot\frac{1}{3} \\ &= \frac{1}{5} \end{aligned}$$

  $$\begin{aligned} Eror &= P_{S_1}(X_2=1)C(P_{S_2}(Y=1|X_2=1)) + P_{S_2}(X_2=0)C(P_{S_2}(Y=1|X_2=0)) \\ &= \frac{3}{5}\cdot C(\frac{2}{3}) + \frac{2}{5}\cdot C(\frac{1}{2}) \\ &= \frac{3}{5}\cdot\frac{1}{3} + \frac{2}{5}\cdot\frac{1}{2} \\ &= \frac{2}{5} \end{aligned}$$

- We should split the tree based on the predicto X_1
- The ensemble method is combing different base classifiers togethe using the majority vote. It can utilize the strengths of all the methods and mitigate thei limitations. Each base classifier must be different. Three examples of base ensemble methods are: bagging via bootstrap, boosting and random forest.

Question 4

One of the hierarchical cluster algorithms is agglomerative (bottom up) procedure. The procedure starts with n singleton clusters and form hierarchy by merging most similar clusters until all the data points are merged into one single cluster. Let the distance between two data points be the Euclidean distance $d(x, y) = \sqrt{(x_1 - y_1)^2 + \ldots + (x_d - y_d)^2}$. Let the distance between two clusters A and B be $min_{x \in A, y \in B} d(x, y)$, the minumum distance between the points from the two clusters. THere are 5 observations a, b, c, d and e. Their Euclidean distances are given in the following matrix:

$$\begin{pmatrix} a & b & c & d & e \\ 0 & 4 & 3 & 6 & 11 \\ 4 & 0 & 5 & 7 & 10 \\ 3 & 5 & 0 & 9 & 2 \\ 6 & 7 & 9 & 0 & 13 \\ 11 & 10 & 2 & 13 & 0 \end{pmatrix}$$

For example, based on the matrix above, the distance between a and b is 4. Please derive the four steps in the agglomerative clustering procedure to construct the hierarchical clustering for the dataset. For each step, you need to specify which two clusters are merged and why you choose these two to merge.

Solution

$$\begin{pmatrix} a & b & c & d & e \\ 0 & & & & \\ 4 & 0 & & & \\ 3 & 5 & 0 & & \\ 6 & 7 & 9 & 0 & \\ 11 & 10 & \mathbf{2} & 13 & 0 \end{pmatrix}$$

From the above matrix, we see the smallest number is **2** which is in the minimum distance which would be the (ce).

$$\begin{array}{lll} d(c, a) = 3 & d(e, a) = 1 & d(d(c, e), d(e, a)) = min(3, 11) = 3 \\ d(c, b) = 5 & d(e, b) = 10 & d(d(c, b), d(e, b)) = min(5, 10) = 5 \\ d(c, d) = 9 & d(e, d) = 13 & d(d(c, d), d(e, d)) = min(9, 13) = 9 \end{array}$$

$$\begin{pmatrix} (ce) & a & b & d \\ 0 & & & \\ \mathbf{3} & 0 & & \\ 5 & 4 & 0 & \\ 9 & 6 & 7 & 0 \end{pmatrix}$$

From the above matrix, we see the smallest number is **3** which is in the minimum distance which would be the (ace).

$$\begin{array}{lll} d(ce, b) = 5 & d(a, b) = 4 & d(d(ce, b), d(a, b)) = min(5, 4) = 4 \\ d(ce, d) = 9 & d(a, d) = 6 & d(d(ce, d), d(a, d)) = min(9, 6) = 6 \end{array}$$

$$\begin{pmatrix} (ace) & b & d \\ 0 & & \\ \mathbf{4} & 0 & \\ 6 & 7 & 0 \end{pmatrix}$$

From the above matrix, we see the smallest number is **4** which is in the minimum distance which would be the (aceb).

$$d(ace, d) = 6 \qquad d(b, d) = 7 \qquad d(d(ace, d), d(b, d)) = min(6, 7) = 6$$

$$\begin{pmatrix} (aceb) & d \\ 0 & \\ \mathbf{6} & 0 \end{pmatrix}$$

Question 5

Analyze the German data set from the site: https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data). Apply the support vector machine analysis and the random forest analysis on the dataset. Please randomly select 800 obervations as the training set and use your two models to predict the default status of the remaining 200 loans. Repeat this cross-validation one thousand times and calculate the avergae misclassification errors of the two models.

Solution

```
set.seed(1)
n = nrow(germandata)
nt = 800
rep = 1000
error_SVM = dim(rep)
error_RF = dim(rep)
neval = n - nt
germandata$Default = factor(germandata$Default)

for (i in 1: rep) {
training = sample(1:n, nt)
trainingset = germandata[training,]
testingset = germandata[-training,]

# SVM Analysis
x = subset(trainingset, select = c('duration', 'amount', 'installment', 'age'))
y = trainingset$Default
xPrime = subset(testingset, select = c('duration', 'amount', 'installment', 'age'))
yPrime = testingset$Default

svm_model1 = svm(x,y)
pred_SVM = predict(svm_model1, xPrime)
tableSVM = table(yPrime, pred_SVM)
error_SVM[i] = (neval - sum(diag(tableSVM)))/neval

#Random Forest Analysis
rf_classifier = randomForest(Default ~., data = trainingset, type = classification,
                             ntree = 100, mtry = 2, importance = TRUE)
prediction_RF = predict(rf_classifier, testingset)
table_RF = table(yPrime, prediction_RF)
error_RF[i] = (neval - sum(diag(table_RF)))/neval
}

mean(error_SVM)
mean(error_RF)
```

Because we are repeating the cross validations a 1000 times, I decided to leave the final number out as it takes a while to run it. However running it through Rstudio, we got the average misclassification error for SVM to be **0.282915** and for random forest to be **0.24224**.

Question 6

The idea of support vector machine (SVM) is to maximize the distance of the separating plane to the closest observation which are referred as the support vectors. Let $g(x) = w_0 + w_1x_1 + w_2x_2 = 0$ be the separating line. For a given sample $x = (x_1, x_2)$, the distance of x to the straight line $g(x) = 0$, is

$$\frac{|w_0 + w_1x_1 + w_2x_2|}{\sqrt{w_1^2 + w_2^2}}$$

- Let the separating line be $x_1 + 2x_2 - 3 = 0$, and the given observation is $x = (1.5, 1.5)$. Calculate the distance of the observation to the separating line.

- In the linear SVM, the dot product $x_i^T x_j$ is an important operation which facilitates the calculation of the Euclidean distance. Let the nonlinear mapping of the sample from the original space to the projected space by ϕ. In nonlinear SVM, the dot product between the images of the mapping $\phi(x_i)$ and $\phi(x_j)$ are calculated by the kernel function $K(x_i, x_j) = \phi(x_i)^T\phi(x_j)$. Suppose in the original space $x_i = (x_{i1}, x_{i2})$ and $x_j = (x_{j1}, x_{j2})$. The nonlinear mappings are $\phi(x_i) = (x_{i1}^2, x_{i2}^2, \sqrt{2}x_{i1}x_{i2})$ and $\phi(x_j) = (x_{j1}^2, x_{j2}^2, \sqrt{2}x_{j1}, x_{j2})$. Calculate the kernel function $K(x_i, x_j)$. If it is a polynomial kernel function, determine the degrees of the polynomial kernel function.

Solution

- $$\begin{aligned}\frac{|w_0 + w_1 + w_2x_2|}{\sqrt{w_1^2 + w_2^2}} &= \frac{|-3 + 1.5 + 2(1.5)|}{\sqrt{(1)^2 + (2)^2}} \\ &= \frac{3\sqrt{5}}{10}\end{aligned}$$

- $$\begin{aligned}K(x_i, x_j) &= \phi(x_i)^T\phi(x_j) \\ &= (x_{i1}^2, x_{i2}^2), \sqrt{2}x_{i1}x_{i2})^T(x_{j1}^2, x_{j2}^2, \sqrt{2}x_{j1}x_{j2}) \\ &= x_{i1}^2x_{j1}^2 + x_{ii2}^2x_{j2}^2 + 2x_{i1}x_{i2}x_{j1}x_{j2} \\ &= (x_{i1}x_{j1} + x_{i2}x_{j2})^2 \\ &= (x_i^T x_j)^2\end{aligned}$$

Question 7

You don't need to submit this question on Crowdmark. This question is only for your practice. In the following table we have the playlist of 10 Spotify users. There are 5 artists A, B, C, D and E. If th user chooses the artist, the corresponding entry will be 1, otherwise, it will be zero.

obs	A	B	C	D	E
1	1	1	0	1	1
2	1	0	1	1	0
3	0	1	1	1	0
4	0	1	1	0	0
5	0	1	1	0	1
6	1	0	0	0	1
7	1	1	1	1	1
8	0	1	1	1	0
9	0	0	1	1	1
10	1	0	1	1	1

- Suppose A is the antecendent and B is the consequent. Calculate the confidence of B and the lift of A on B. Based on the lift value, do you recommend B to the user after the user has played artist A? Why?

Solution