demo

Tuesday, June 22, 2021 8:05 PM

Question 1
Incorrect
Mark 0.00 out of 2.00
Flag question

Linear and logistic regression models can be seen as neural networks. Which of the following statement is correct?

Select one:

a. The difference between the two models is the activation function of the output node and nothing else.

b. The difference between the two models is the activation function of the output node and the loss function.

c. The difference between the two models is the activation functions of the hidden layer and the output node, and the loss function. ✖

d. The difference between the two models is the loss function and nothing else.

Question 2
Partially correct
Mark 0.50 out of 2.00
Flag question

Which of the following statement(s) about support vector machines (SVMs) are correct? Select all that apply.

Select one or more:

a. The distance of the observations from the margin is called the hyperplane.

b. Deleting the support vectors will change the position of the hyperplane.

c. SVM can separate the data points that are not linearly separable. ✔

d. In SVM, if the number of input features is 2, then the hyperplane is a plane. ✖

Question 3
Partially correct
Mark 0.67 out of 2.00
Flag question

After training a soft-margin SVM with a linear kernel, you find both training and validation accuracy are low. What will you consider next? Select all that apply.

Select one or more:

a. Increase the cost parameter C ✔

b. Decrease the cost parameter C

c. Include more features

d. Use a nonlinear kernel

There are two possibilities of having low training and validation accuracy. The first one is that the data quality is poor, e.g. insufficient features to explain the response, and noisy features and/or labels. If this is the case, we could consider adding more features (choice c), or cleaning the data (not covered in this course). The other possibility is that the current model is too simple to describe the data, i.e. underfit the data. In this case, we could increase the cost parameter so as to encourage the model to focus on classifying the data correctly rather than seeking a large margin (choice a), or use a nonlinear kernel for generating a nonlinear decision boundary (choice d).

Question 4
Incorrect
Mark 0.00 out of 2.00
Flag question

The figures below show the scatter plot of data from two classes and the decision boundary of 1-nearest neighbour classifier and 9-nearest neighbours classifier, assuming the Euclidean distance is used as the distance measure. Which two of them correspond to the results from 1-nearest neighbour classifier?

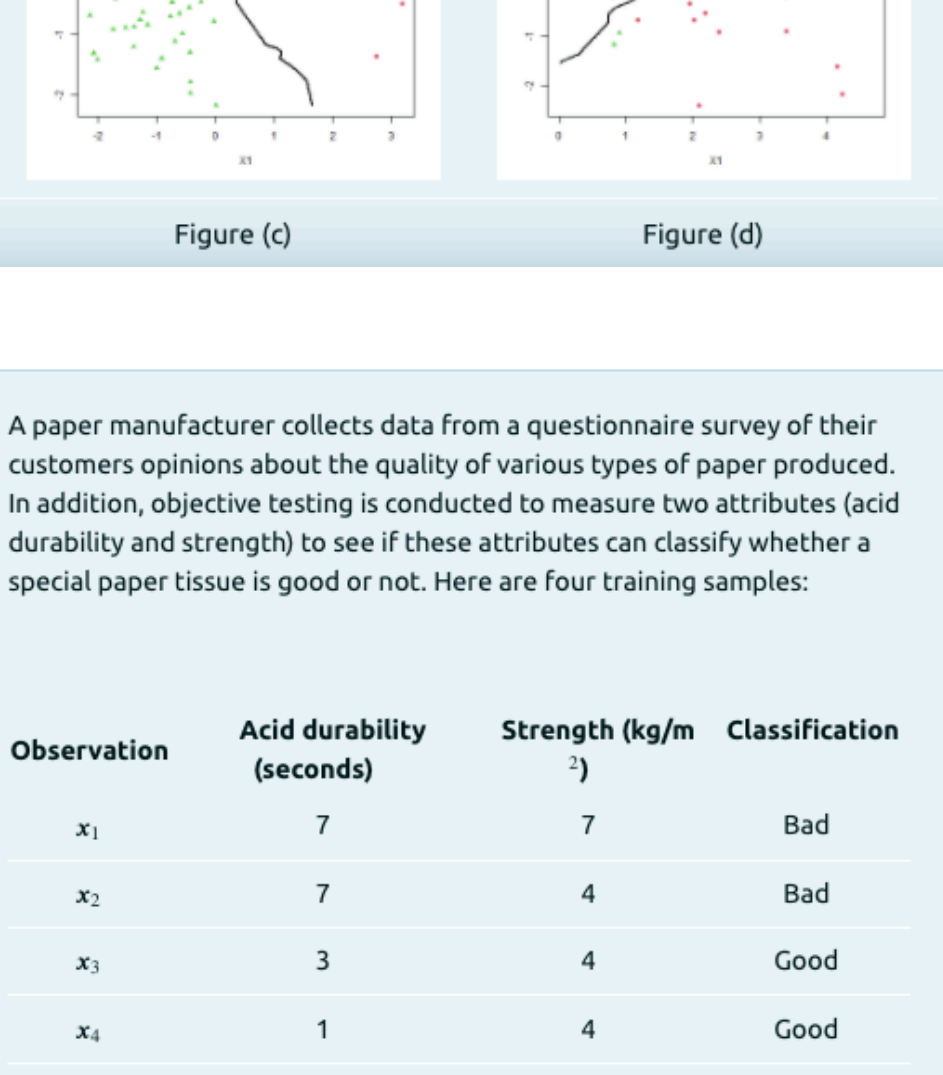

Figure (a) Figure (b)

Figure (c) Figure (d)

a,b

Question 5
Not answered
Marked out of 3.00
Flag question

A paper manufacturer collects data from a questionnaire survey of their customers opinions about the quality of various types of paper produced. In addition, objective testing is conducted to measure two attributes (acid durability and strength) to see if these attributes can classify whether a special paper tissue is good or not. Here are four training samples:

Observation	Acid durability (seconds)	Strength (kg/m^2)	Classification
x_1	7	7	Bad
x_2	7	4	Bad
x_3	3	4	Good
x_4	1	4	Good

Now the factory produces a new paper tissue, x_{test}, that passes laboratory tests with acid durability of 3 seconds and strength of 7 kg/square meter. Predict the class label for this new paper tissue by using the 1-nearest neighbour algorithm with the squared Euclidean distance as the distance measure.

Enter the intermediate and final results as integers.

$d_E^2(x_1, x_{test})$ (the squared Euclidean distance between x_1 and x_{test}): 16

Seconds	kg/m	Distance	Rank	Include in 1-nn	Classification
7	7	16	3	N	
7	4	25	4	N	
3	4	9	1	Y	Good
1	4	13	2	N	

Question 6
Not answered
Marked out of 3.00
Flag question

Consider the two-class dataset below:

Observation	X_1	X_2	Y
1	3	4	Red
2	2	2	Red
3	4	4	Red
4	1	4	Red
5	2	1	Blue
6	4	3	Blue
7	4	1	Blue

Table: a two-dimensional dataset with 7 observations

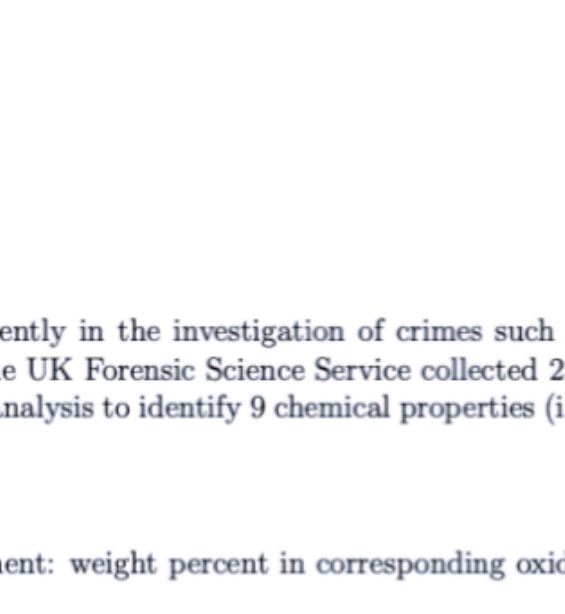

Figure: scatterplot of the data set (the two classes are represented by red circles and blue triangles)

set $w_2=1$, solve w_1, b
$w_1 x + w_2 t b$
$\Rightarrow \begin{cases} w_1 = -1 \\ b = 0.5 \end{cases}$

The decision boundary of a linear SVM can be written in the form of $X_2 = aX_1 + b$. Find the decision boundary for this dataset and enter the corresponding values of a and b.

a: 1

b: -0.5

2. Glass is a material which figures prominently in the investigation of crimes such as burglary. To study the type of glasses, the UK Forensic Science Service collected 214 glass samples and carried out a chemical analysis to identify 9 chemical properties (i.e. 9 variables) for each sample:

- RI (X_1): refractive index
- Na (X_2): Sodium (unit of measurement: weight percent in corresponding oxide; same applies to variables X_3-X_9)
- Mg (X_3): Magnesium
- AI (X_4): Aluminium
- Si (X_5): Silicon
- K (X_6): Potassium
- Ca (X_7): Calcium
- Ba (X_8): Barium
- Fe (X_9): Iron

A sample observation looks as follows:

RI	Na	Mg	Al	Si	K	Ca	Ba	Fe
1.5	14	4.5	1.1	72	0.06	8.8	0	0

(a) To reduce the dimension of this data set, a researcher has applied principal component analysis (PCA) based on the correlation matrix. Suggest why PCA might have been run on the correlation matrix instead of the covariance matrix? **[2 MARKS]**

The measurement of unit of reflective index is different from that of chemical components. Variance may be very different across variables

(b) Partial output from the principal component analysis is given below.

```
> glass.pca <- princomp(Glass,cor=TRUE)
> glass.pca

Standard deviations:
Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
  1.58   1.43   1.19   1.08   0.96   0.73   0.61   0.25   0.04
```

```
> glass.pca$loadings
Loadings:
   Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
RI  0.545  0.286         0.147         0.115         0.752
Na -0.258  0.270 -0.385  0.491 -0.154 -0.558  0.149  0.128 -0.312
Mg  0.111 -0.594         0.379 -0.124  0.308 -0.206        -0.577
Al -0.429  0.295  0.329 -0.138               -0.699  0.274 -0.192
Si -0.229 -0.155 -0.459 -0.653                0.216  0.380 -0.298
K  -0.219 -0.154  0.663         0.307 -0.244  0.504  0.110 -0.261
Ca  0.492  0.345        -0.276  0.188 -0.149        -0.399 -0.579
Ba -0.250  0.485         0.133 -0.251  0.657  0.352 -0.145 -0.198
Fe  0.186         0.284 -0.230 -0.873 -0.243
```

<0.1

Comment on the loadings of the first principal component. Comment on the role of the variable Fe in the principal component analysis. **[4 MARKS]**

SOLUTION: *(moderate)*
The first principal component can be interpreted as the difference between the weighted average of the variables RI, Mg, Ca and Fe, and the weighted average of Na, Al, Si, K and Ba [2 marks]. The contribution of variable Fe is relatively small as it only appears in the loadings of five of the nine components. In particular, its contribution to the first principal component is the second smallest among all variables and the contribution to the second principal component is small (smaller than the cutoff value of 0.1) [2 marks].

(c) The researcher chose to use the Proportion of Variation approach to determine the number of principal components to be retained. Based on the previous R output, decide how many components should be kept in order to explain 85% of the variability of the data set. **[4 MARKS]**

$var(Y_j) = \lambda_j$

	PC1	PC2	PC3	PC4	PC5
λ_j	$1.58^2 = 2.4964$	2.0449	1.4161	1.1664	0.9216
prop. var.	0.2774	0.2272	0.1573	0.1296	0.1024
cum. prop. var	0.2774	0.5046	0.6619	0.7915	0.8939

When the Proportion of Variation approach is used, we should select the first q principal components such that $\sum_{j=1}^{q} \lambda_j / \sum_{j=1}^{9} \lambda_j \geq 0.85$.

The first row of the table, λ_j, calculates the jth eigenvalue, which equals to the square of the standard deviation of the corresponding component, i.e. `sdev` in the R output.

The second row, proportion of variance (abbreviated to prop. variance), is defined as $\lambda_j / \sum_{j=1}^{9} \lambda_j$. Since PCA is computed based on the correlation matrix, $\sum_{j=1}^{9} \lambda_j = \sum_{j=1}^{9} \text{Var}(X_j) = 9$.

The third row, cumulative proportion of variance at Comp i, is defined as $\sum_{j=1}^{i} \lambda_j / \sum_{j=1}^{9} \lambda_j$.

To explain 85% of original variability, 5 principal components should be retained.

(d) An important task in forensic science is to identify if the glass sample is a window glass (class 1) or non-window glass (class −1). For this purpose, a support vector machine with a linear kernel is applied. Partial output from the fitted model is given below.

```
> C.val <- c(0.1,0.5,1,2,5,10)
> glass.cv <- tune.svm(Class~., data=Glass.train,
                type="C-classification", cost=C.val,
                kernel="radial", gamma=10)
                #Glass.train: training data set of Glass
> C.opt <- glass.cv$best.parameters$cost
> glass.svm <- svm(Class~., data=Glass.train,
             type="C-classification", cost=C.opt,
             kernel="radial", gamma=10)
> summary(glass.svm)
Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  2

Number of Support Vectors:  156
 ( 118 38 )


Number of Classes:  2

Levels:
 -1 1
> table(Glass.train$Class,predict(glass.svm,Glass.train))
     glass.pred
       -1   1
  -1   38   0
  1     0 122
> table(Glass.test$Class,predict(glass.svm,Glass.test))
       #Glass.test: test data set of Glass
     glass.pred
      -1  1
  -1   1 12
  1    0 41
```

soft margin
non-linear SVM

Comment on the training and test performance of the fitted support vector machine. **[2 MARKS]**

Classification (lecture 3)

SOLUTION: *(easy)*
The support vector machine can correctly classify all training samples.
For the test samples, the correct classification rate is only 77.8%, which indicates the possibility of overfitting to the training data. Moreover, the method can correctly classify all samples from the window class. However, for the non-window class, 12 out of 13 (92.3%) samples are misclassified. This may be a consequence of imbalanced class distribution.

(e) Suggest one way to improve the test performance of the previous support vector machine. Explain your suggestion. **[2 MARKS]**

link para in non-linear SVM with accuracy

SOLUTION: *(moderate)*
As the method tends to overfit, one potential solution is to decrease the width parameter (γ) used in the radial basis function kernel. Decreasing γ will make use of training samples that are farther away from the test sample. [2 marks]
An alternative answer: As the method misclassifies samples from the non-window class (minority class), one potential solution is to set different cost parameters when training samples violate the margin constraint. Higher penalty should be applied when non-window samples violate the margin or is misclassified.

(f) Write down a piece of R code to evaluate the accuracy of the following support vector machine using leave-one-out cross-validation on the training data set (`Glass.train`): the support vector machine uses a polynomial kernel of degree 2 and the cost parameter for violating the margin constraint is set to 1. Note that you CANNOT use any built-in function, such as `svm.tune()` and `tune()`. You can either handwrite the code or append the typed code to your script.

```
# name of the data set: Glass
# type of SVM: kernel - polynomial kernel of degree 2
#              cost parameter - 1

# Perform leave-one-out cross-validation (LOOCV)
library(e1071)
set.seed(1)
### write your R code here

# Return the final result, i.e. LOOCV accuracy
LOOCV.accuracy <-
```

[4 MARKS]

SOLUTION: *(hard)*

```
# Perform leave-one-out cross-validation (LOOCV)
library(e1071)
set.seed(1)
n <- nrow(Glass)
LOOCV <- numeric(n) #a vector to store the accuracy at each fold
for (i in 1:n){
    Glass.train <- Glass[-i,]
    Glass.test <- Glass[i,] #[1 mark]
    glass.svm <- svm(Class~., data=Glass.train, type="C-classification",
                  kernel="polynomial", degree=2, cost=1) #[1 mark]
    glass.pred <- predict(glass.svm, Glass.test) #[1 mark]
    LOOCV[i] <- mean(glass.pred==Glass.test$Class)
    }

# Return the LOOCV accuracy
LOOCV.accuracy <- mean(LOOCV) #[1 mark]
```

(g) State one advantage and one disadvantage of leave-one-out cross-validation compared to 10-fold cross-validation. **[2 MARKS]**

SOLUTION: *(easy)*
Leave-one-out cross-validation is approximately unbiased for the true (expected) prediction error [1 mark] but can have high variance. It is also computationally

CONTINUED OVERLEAF/

intensive since it involves fitting the model as many times as the number of data points. [1 mark]