practice midterm 3

Go Ito

May 13, 2019

Question 1. KNN

(a)

We have the following training data:

$$\mathbf{X} = \begin{bmatrix} 1.4 & 0.2 \\ 1.7 & 0.4 \\ 5.0 & 1.9 \\ 5.4 & 2.3 \\ 5.1 & 1.8 \end{bmatrix}, \mathbf{Y} = \begin{bmatrix} \text{setosa} \\ \text{setosa} \\ \text{virgninica} \\ \text{virgninica} \\ \text{virgninica} \end{bmatrix}$$

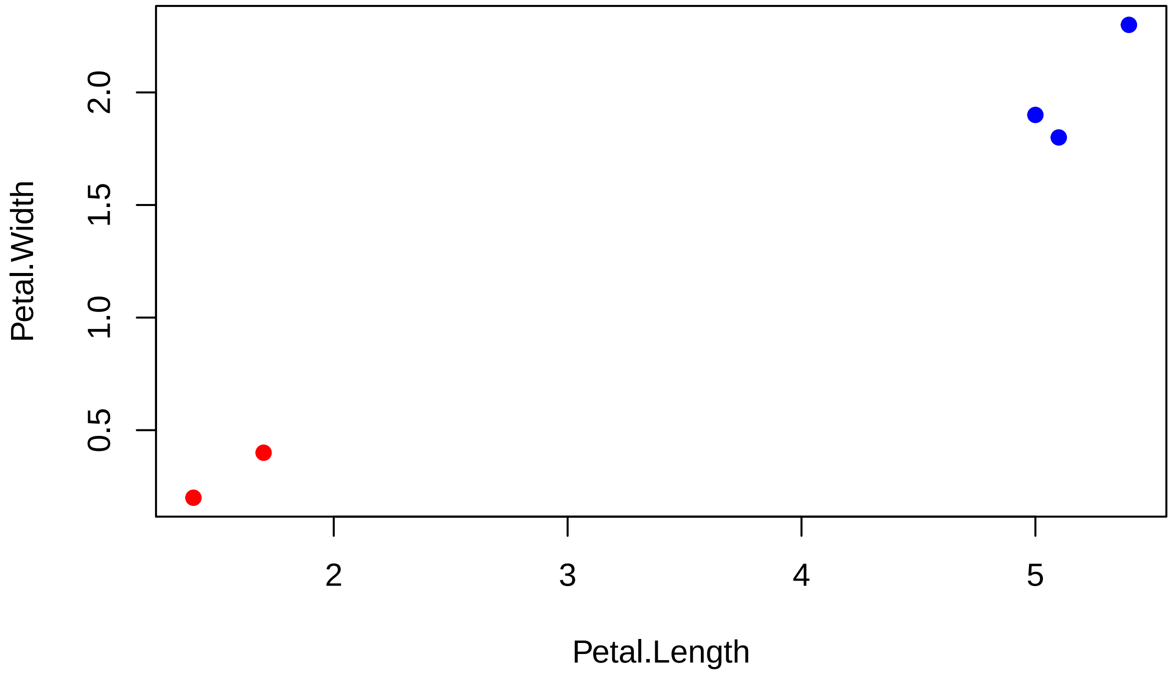
What would the following test data be classified as? Please describe using KNN (k=1 and k=3).

$$\mathbf{X_{new}} = \begin{bmatrix} 5.2 & 2.0 \end{bmatrix}$$

If we choose k=5, what critical problem(s) would happen?

(b)

Using the above data, we have a plot as follows. Draw the equidistant decision boundaries for k=1 and k=5.



Question 2. SVM

Use the above plot, draw radial Kernel decision boundaries with the following four cases:

- gamma = 1, cost = 1
- gamma = 10, cost = 1
- gamma = 0.1, cost = 1
- gamma = 1, cost = 10

Question 3. K-Means

Given the following data with random initial assignment Z:

$$\mathbf{X} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 9 & 9 \\ 9 & 8 \\ 8 & 9 \end{bmatrix}, \mathbf{Z} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

After one iteration of K-Means Clusetering, what is the new centroid and the assignment Z?

After another iteration of K-Means Clustering, what is the new centroid and the assignment Z?

If you take another iteration, what would happen? (no calculation needed.)

Question 4. Kernelized Clustering

Reconsider the dataset used in Question 3.(same training data, same initial assignemt). This time, we are performing Kernelized Clusetering with the following transformation function and Kernel function:

$$\phi(\mathbf{X}) = \begin{bmatrix} X_1^2 \\ X_2^2 \\ \sqrt{2}X_1X_2 \end{bmatrix}, K(\mathbf{X}_1, \mathbf{X}_2) = (\mathbf{X}_1^T\mathbf{X}_2)^2$$

(a). Transforming data

- Obtain the transformed data using ϕ.
- Obtain the matrix of the distances to the centroid.

(b). Using Kernel directly

- Obtain the expression of the kerelized inner product (use the untranformed data)!
- Obtain the matrix of the distance to the centroid.
- Confirm that two distance matricies are identical.

Question 5. EM Algorithm

Using the following two univariate normal distributions (i.e. this is the correct distribution):

$$\begin{aligned} X_1 &\sim N(2, 1) \\ X_2 &\sim N(9.5, 1.5) \end{aligned}$$

6 random points are generated:

$$\mathbf{X} = (1.8, 2.5, 2.7, 6.9, 7.8, 9.5)^T$$

Given the following arbiturary initial information:

$$\begin{aligned} \alpha &= (0.7, 0.3)^T \\ \mu_1 &= 2.2 \\ \sigma_1 &= 0.5 \\ \mu_2 &= 10 \\ \sigma_2 &= 2 \end{aligned}$$

Do two iterations of EM-Algorithm. What's the resulting final weights(6×2), means(2×1) and standard deviations(2×1)? Also, what's the resulting assignments(6×2)?

Question 6. PCA

Given the data:

$$\mathbf{X} = \begin{bmatrix} 1 & 1 \\ 2 & 3 \\ -1 & -2 \\ -2 & -4 \end{bmatrix}$$

Do the following:

(a.)

- Center the data.
- Find the covariance matrix (Note: $\mathbf{\Sigma} = \frac{1}{N}\mathbf{X}_c^T\mathbf{X}_c$)
- Get the eigen-values of the covariance matrix and the corresponding eigen-vectors.
- Plot the resulting projected data.

(b.) Challenge yourself

- Show that the projected data are uncorrelated.

Question 7.

- In what occasion, we can possibly have a "tie" among the KNN? If that's the case, what should we do to classify the test cases?
- Regarding KNN, what's the issue of having too small K? Too big K?
- What is the object of hard-margin SVM? This method tends to overfit to the training data. How so? How would you improve this method to alleviate this issue? Please explain two ways to do so.
- What is supervised learning? Unsupervised learning? How is the existance of the target values useful for supervised learning?
- How do we choose the value for K for K-means clustering?
- When is Kernelized clustering useful compare to K-means?
- How come it is impossible to visualize the centroid when we use Gaussian Kernel? Give a breif mathmatical explanation.
- In what occasions, EM-algorithm is more useful than K-means or Kernelized clustering?
- Mr. Avash claims the following statement: "When we create statistical models, we often aim to minimize the model SSE (Loss). So, with the similar logic, we should aim to minimize the variance of the projected data when using PCA!" What is is the critical mistake in his statement? How so? (Shame on you Avash :P)
- What is model?
- When the penalty parameter λ for LASSO is too big, what would happen to Lasso estimated coefficients? What is the advantage of using Lasso regresison?
- What's the purpose of gradient descent? What is the effet of having too big "step size" parameter γ?
- What is the purpose of Cross Validation? Why is K-fold CV more preferred than LOOCV? (two reasons)
- Suppose you have a dataset with 10 observations and want to conduct K-fold CV. When we set K=4, how many observations do we have for each block? What value of K do we need to do LOOCV?
- What is over-fitting? How is it related to bias-variance trade off?
- What is Bias? What is variance?
- Consider Ridge regression. We have discussed that Ridge coefficients tends to become "similar" in value, then move toward the origin (zero). Suppose your initical coefficient estimate is $\hat{\mathbf{w}} = (0.2, -0.7)^T$. Which one of the following lines would this coefficients "move" toward first?: $X_1 = X_2$ or $X_2 = -X_1$?

- What is the difference between probability density and likelhood?
- What is the difference between probablistic classifier and non-probablistic classifier? List 2 examples for each type of classifiers.
- What is Naive Bayes Classifier? Pros and cons?
- How is Hessian useful in general? How is it useful in terms of log-likelihood?
- Recall the marvel example from the class (6-1 page 19). We have 50% red marvels from factory B. However, when the test case (10/20 are red) were concerned, the Bayes Classifier suggested that this test case bag is more likely from factory A, where only 40% of the marvels are red. Why did this happen?
- What are the pros and cons of Neural Network?
- What are the four main components of Baye's classifier? Explain each.
- What is the score? How is it used in Fisher Information? Explain why the expected value of the derivative of the score is equivalent to variance?
- What does it mean when the value of a Hessian is overall high? When would this happen?