

Department of Data Science
IIT Palakkad

DS5006 : Machine Learning (August 2024)

09:30-12:30

End Semester (21 Nov 2024)

Marks : 45

Instructions

1. Write your answers neatly in Blue/ Black ink. Do not use pencil / Red ink. Make sure your answers are legible.
2. If you have to make any assumption about unspecified things, write the assumption clearly with justification.
3. Write the question number clearly for each answer. Draw a line after the answer.
4. There will be partial markings for the questions, so even if you are not able to solve the entire problem be sincere with the steps.
5. **Be precise.**

- ① What is the "regression toward the mean" effect? (3)
Give an example of the effect.
Given a dataset with one predictor and one response variable, how can you explain the possibility of observing the effects of regression toward the mean?
2. Which of the following words is the best and worst choice as the first word for Wordle? (4)
Briefly explain.
 1. salet
 2. fuzzy
 3. least
 4. lucky
 5. queen
- ③ We want to do a large-scale empirical comparison between supervised classification methods to come up with some guidelines for selecting a classification method for a particular task. How will you design experiments for such empirical comparison? (5)
- ④ (Source: [Han et al., 2011]) Suppose that we would like to select between two prediction models, M1 and M2. We have performed 10 rounds of 10-fold cross-validation on each model, where the same data partitioning in round i is used for both M1 and M2. (4)
The error rates obtained for M1 are 30.5, 32.2, 20.7, 20.6, 31.0, 41.0, 27.7, 26.0, 21.5, 26.0.
The error rates for M2 are 22.4, 14.5, 22.4, 19.6, 20.7, 20.4, 22.1, 19.4, 16.2, 35.0.
Comment on whether one model is significantly better than the other considering a significance level of 1%.
(Refer to the t Table provided at the end of the question paper).

5. (Source: Singh, Jaakkola, Mohammad¹) Consider the data plotted in Figure 1, which consists of two rows of equally spaced points. (3+5)
- (a) Redraw the figure in your answer sheet. If k-means clustering ($k = 2$) is initialized with the two points whose coordinates are (9, 3) and (11, 3), indicate the final clusters obtained (after the algorithm converges).
- (b) Redraw the figure in your answer sheet one more time. Let us assume that there is one more point whose coordinates are (10, 8). What could be a possible k-means clustering ($k = 2$) initialized with the same points with coordinates: (9, 3) and (11, 3)? Briefly explain your answer.

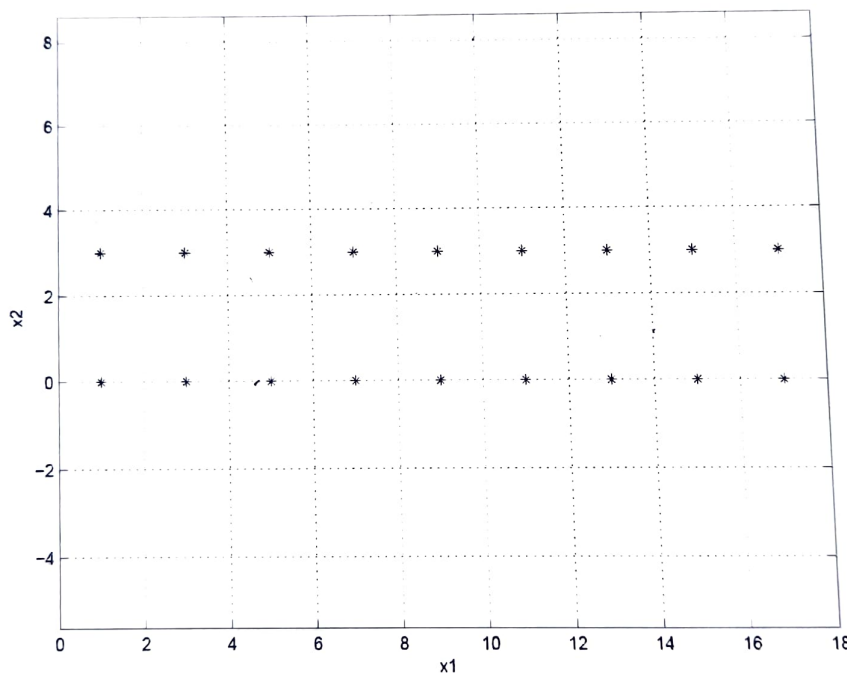


Figure 1: Data for K-means Clustering

6. (Source: [Tan et al., 2018]). Consider the training examples shown in Table 1 for a binary classification problem. (2+4+2)
- (a) Compute the Gini index for the overall collection of training examples.
- (b) Compute the Gini index for Customer ID, Gender, Car Type, and Shirt size attributes.
- (c) Considering the Gini index for Customer ID, Gender, Car Type, and Shirt size attributes, which attribute is better as a root while constructing a decision tree?

¹https://ocw.mit.edu/courses/6-867-machine-learning-fall-2006/resources/final_f04/

Customer ID	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

Table 1: Binary Classification Dataset

7. Compare the advantages and disadvantages of model (or prototype)-based and instance (or exemplar)-based classification methods. (3)
8. (Source: [Tan et al., 2018]) We take a sample of adults and measure their heights. If we record the gender of each person, we can calculate the average height and the variance of the height, separately, for men and women. Suppose, however, that this information was not recorded. Would it be possible to still obtain this information? Explain. (4)
9. Consider the XOR problem where there are four training points: (3)

x_1	x_2	Class label
-1	-1	-1
-1	+1	+1
+1	-1	+1
+1	+1	-1

Table 2: XOR Dataset

How can we transform the points into a feature space so that they are linearly separable?

10. PCA and Kernel methods both aim to find a “better” representation of data. However, their approaches are opposite: PCA projects data into a lower-dimensional space while Kernel methods project data into a higher-dimensional space. (3)

We did not discuss Kernel PCA in class, but its core idea is as follows: In Kernel PCA, the Kernel trick is used to project the data in a higher (or even infinite) dimensional space— F , and then PCA is performed in the F space to get a lower-dimensional representation. It has been observed that Kernel PCA finds a better lower-dimensional space than simple PCA in certain scenarios.

If the “curse of dimensionality” is a major obstacle in machine learning, then going to higher-dimensional space is counterintuitive.

What is the missing link in the above argument? What are the possible scenarios in which Kernel PCA is likely to find a better lower-dimensional space than simple PCA?

END

t Distribution: Critical Values of t

Degrees of freedom	Two-tailed test: One-tailed test:	Significance level					
		10%	5%	2%	1%	0.2%	0.1%
		5%	2.5%	1%	0.5%	0.1%	0.05%
1		6.314	12.706	31.821	63.657	318.309	636.619
2		2.920	4.303	6.965	9.925	22.327	31.599
3		2.353	3.182	4.541	5.841	10.215	12.924
4		2.132	2.776	3.747	4.604	7.173	8.610
5		2.015	2.571	3.365	4.032	5.893	6.869
6		1.943	2.447	3.143	3.707	5.208	5.959
7		1.894	2.365	2.998	3.499	4.785	5.408
8		1.860	2.306	2.896	3.355	4.501	5.041
9		1.833	2.262	2.821	3.250	4.297	4.781
10		1.812	2.228	2.764	3.169	4.144	4.587
11		1.796	2.201	2.718	3.106	4.025	4.437
12		1.782	2.179	2.681	3.055	3.930	4.318

References

- [Han et al., 2011] Han, J., Kamber, M., and Pei, J. (2011). *Data Mining: Concepts and Techniques*. The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann.
- [Tan et al., 2018] Tan, P.-N., Steinbach, M., Karpatne, A., and Kumar, V. (2018). *Introduction to Data Mining (2nd Edition)*. Pearson, 2nd edition.