

Department of Data Science
IIT Palakkad

DS5006 : Machine Learning

0830-0945

Mid-Semester Examination

Marks : 20

Instructions

1. Write your answers neatly in Blue/ Black ink. Make sure your answers are legible.
2. Doubts and questions will not be answered during the exam. 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. There are total 6 questions.
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. Which of the following regression methods is more appropriate for data with outliers? (An outlier is a data point that differs significantly from other observations¹.) (3)
Explain with an example.
1. least squares
 2. least absolute deviation
2. A rectangular box without a top (a topless box) is to be made from 12 ft² of cardboard. Find the maximum volume of such a box². (4)
3. Which of the following classification approaches will be most and least impacted by the curse of dimensionality? Why? (2)
1. Naive Bayes
 2. k-Nearest Neighbor
4. What will be the effect of varying k in the k-Nearest Neighbor algorithm in terms of bias and variance? (2)
5. Hyperbolic tangent, "tanh" is considered as an alternative to Sigmoid function. It is defined as: (4)

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

Derive the Hessian for the Logistic Regression formulation where the Sigmoid function is replaced by \tanh .

¹<https://en.wikipedia.org/wiki/Outlier>

²<https://tinyurl.com/ywr8awvs>

6. Fitting a naive Bayes spam filter by hand (Source: Daphne Koller and Kevin Murphy). (5)

Consider a Naive Bayes model for spam classification with the vocabulary: $V = \{ \text{"secret"}, \text{"offer"}, \text{"low"}, \text{"price"}, \text{"valued"}, \text{"customer"}, \text{"today"}, \text{"dollar"}, \text{"million"}, \text{"sports"}, \text{"is"}, \text{"for"}, \text{"play"}, \text{"healthy"}, \text{"pizza"} \}$

We have the following example spam messages:

1. million dollar offer
2. secret offer today
3. secret is secret

and normal messages:

1. low price for valued customer
2. play secret sports today
3. sports is healthy
4. healthy low price pizza offer for today

Give the Maximum Likelihood Estimates (MLEs) for the following parameters:

1. $P(\text{spam})$
2. $P(\text{secret}|\text{spam})$
3. $P(\text{secret}|\text{non-spam})$
4. $P(\text{offer}|\text{non-spam})$
5. $P(\text{offer}|\text{spam})$
6. $P(\text{dollar}|\text{spam})$

Let us assume that your experiences with spam emails tell you that the words *offer* and *secret* are key indicators of spam messages. How would you translate this knowledge into the above MLEs? What would be the revised MLEs?