ELG5255- Applied Machine Learning
Professor: Dr. Hitham Jleed
Student name :(Solution)

uOttawa

01/11/2023
Midterm Exam (Fall 2023)
Student ID: ( )

This exam is 120 minutes long.
There should be four numbered pages in this exam.
This exam is closed book. It contains 4 questions.
Questions that ask you to "briefly explain" require short (1-3 sentence) explanations.
Calculators are allowed, **but no laptops or phones.**

Question 1. Short Answers

a. Give an example of a loss function for classification problems.

Entropy; hinge loss; number of errors; etc.

b. Explain the trade-off between bias and variance.

Expected Answer (Such as): High bias can cause the model to miss relevant relations between features and target outputs (underfitting), while high variance can cause the model to model the random noise in the training data (overfitting). Ideally, a model should have a balance between bias and variance, ensuring it is neither too simplistic nor too complex.

c. How can increasing the complexity of a model affect its bias and variance?

Expected Answer (Such as): Increasing the complexity of a model typically decreases bias since the model can fit the training data more closely. However, this can lead to an increase in variance as the model becomes more sensitive to fluctuations and noise in the training data, potentially leading to overfitting

d. Suppose you are given an EM algorithm that finds maximum likelihood estimates for a model with latent variables. You are asked to modify the algorithm so that it finds MAP estimates instead. Which step or steps do you need to modify?

1. Expectation 2. Maximization 3. No modification necessary 4. Both

Maximization

Question 2:

a. Below is the probability density function for the Rayleigh distribution.

$$f(x;\theta) = \frac{x}{\theta^2} exp\left(\frac{-x^2}{2\theta^2}\right)$$

Suppose Lord Rayleigh gathered data consisting of $X_1, X_2, \ldots, X_n$, which deems are iid $Rayleigh(\theta)$ random variables

$l(\theta|X) \equiv$ $lik(\theta) = \prod_{i=1}^{n} [\frac{x_i}{\theta^2} exp\left(\frac{-x_i^2}{2\theta^2}\right)]$

Given a set of independent and identically distributed (i.i.d) random variables $X_1, X_2, \ldots, X_n$ from the Rayleigh distribution, the likelihood of observing this data given the parameter θ is:

$\text{lik}(\theta) = \prod_{i=1}^{n} \left[\frac{x_i}{\theta^2} \exp\left(-\frac{x_i^2}{2\theta^2}\right)\right]$

The likelihood function is a product of the individual probabilities of the observed data.

$\mathcal{L}(\theta|X) \equiv$ $log(lik(\theta)) = [\sum_1^n log(x_i)] - 2nlog(\theta) - \frac{1}{\theta^2}\sum_1^n [x_i^2/2]$

Taking the natural logarithm of the likelihood function gives the log-likelihood, which is often easier to work with for estimation purposes:

$\mathcal{L}(\theta|X) = \log(\text{lik}(\theta)) = \sum_{i=1}^{n} \log(x_i) - 2n\log(\theta) - \frac{1}{\theta^2}\sum_{i=1}^{n} \frac{x_i^2}{2}$

The log-likelihood is generally preferred because it turns the product into a sum, making the calculations, especially differentiation, simpler.

$$\hat{\theta}_{MLE} = (\frac{1}{n}\sum_1^n [x_i^2/2])^{1/2}$$

The value of θ that maximizes the log-likelihood function is the maximum likelihood estimator ($\hat{\theta}_{MLE}$), and it is given by the expression:

$\hat{\theta}_{MLE} = \left(\frac{1}{n}\sum_{i=1}^{n} \frac{x_i^2}{2}\right)^{1/2}$

This estimator provides the value of θ that is most likely to result in the observed data according to the Rayleigh model.

b. **Hierarchical Agglomerative Clustering (HAC)**

Given three clusters, *A*, *B* and *C*, containing a total of six points, where each point is defined by an integer value in one dimension, *A* = {0, 2, 6}, *B* = {3, 9} and *C* = {11}, which two clusters will be merged at the *next iteration* of HAC when using Euclidean distance and

(i) Single linkage?

i. Merge *A* and *B*
ii. Merge *A* and *C*
iii. Merge *B* and *C*

```
(i) AB = 1, AC = 5 and BC = 2, so merge A and B
```

(ii) Complete linkage?

i. Merge *A* and *B*
ii. Merge *A* and *C*
iii. Merge *B* and *C*

```
(iii) AB = 9, AC = 11 and BC = 8, so merge B and C
```

Question 3:

a. Two litters of a particular rodent species have been born, one with two brown-haired and one gray-haired (litter 1), and the other with three brown-haired and two gray haired (litter 2). We select a litter at random and then select an offspring at random from the selected litter. Find.

- The probability that the animal chosen is brown-haired $\equiv P(Brown\ hair)$
- Given that a brown-haired offspring was selected, the probability that the sampling was from litter $1 \equiv P(Little1|Brown\ hair)$

$$\begin{aligned} &P(\text{Brown Hair}) \\ &= P(\text{Brown Hair}|\text{Litter 1})P(\text{Litter 1}) + P(\text{Brown Hair}|\text{Litter 2})P(\text{Litter 2}) \\ &= \left(\frac{2}{3}\right)\left(\frac{1}{2}\right) + \left(\frac{3}{5}\right)\left(\frac{1}{2}\right) = \frac{19}{30}. \end{aligned}$$

$$P(\text{Litter 1}|\text{Brown Hair}) = \frac{P(BH|L1)P(L1)}{P(BH|L1)P(L1) + P(BH|L2)P(L2} = \frac{\left(\frac{2}{3}\right)\left(\frac{1}{2}\right)}{\frac{19}{30}} = \frac{10}{19}.$$

(b) **K-Means Clustering**

☆ True or False: *K-Means Clustering* is guaranteed to converge (i.e., terminate).

True

☆ True or False: A good way to pick the number of clusters, *k*, used for *k*-Means clustering is to try multiple values of *k* and choose the value that minimizes the distortion measure.

False

☆ (True/False) Running k-means with a larger value of k always enables a lower possible final objective value than running k-means with smaller k.

True

Question 4:

a. Our goal is to construct a decision tree classifier for predicting flight delays. We have collected data for a few months and a summary of the data is provided in the following Table.

Feature	Value = yes	Value = no
Rain	Delayed - 30, not Delayed - 10	Delayed - 10, not Delayed - 30
Wind	Delayed - 25, not Delayed - 15	Delayed - 15, not Delayed - 25
Summer	Delayed - 5, not Delayed - 35	Delayed - 35, not Delayed - 5
Winter	Delayed - 20, not Delayed - 10	Delayed - 20, not Delayed - 30
Day	Delayed - 20, not Delayed - 20	Delayed - 20, not Delayed - 20
Night	Delayed - 15, not Delayed - 10	Delayed - 25, not Delayed - 30

i. Based on the table, which feature should be at the root of the decision tree (briefly explain, no need to provide exact values for information gain)?

Solution: The root would be Summer. It is easy to see that using Summer would lead to the fewest mistakes (10 overall) and so would lead to the highest information gain.

ii. Based on the table, which feature should be on the second level (the level just beneath the root) of the decision tree (briefly explain, no need to provide exact values for information gain)?

Solution: It is impossible to tell. To determine the second level feature we need to know the breakdown of delays for the other features given the value of Summer. Since we only have summaries in this table it is not enough information for determining the feature that would lead to the highest information gain AFTER we used Summer.

b. Given the following data of transactions at a shop, calculate the support and confidence values of milk → bananas, bananas → milk, milk → chocolate, and chocolate → milk.

Transaction	Items in basket
1	milk, bananas, chocolate
2	milk, chocolate
3	milk, bananas
4	chocolate
5	chocolate
6	milk, chocolate

milk → bananas : Support = 2/6, Confidence = 2/4

bananas → milk : Support = 2/6, Confidence = 2/2

milk → chocolate : Support = 3/6, Confidence = 3/4

chocolate → milk : Support = 3/6, Confidence = 3/5