INDIAN INSTITUTE OF TECHNOLOGY (ISM) DHANBAD

End-Term Examination

Machine Learning: MSD527

(Academic Year 2021-22)

Course: Machine Learning **Max Marks: 100**

Date: 29/4/22 **Duration: 3 hours**

Instructions: Scientific Calculators are allowed

Q1. Consider the design matrix

$$\begin{bmatrix} 4 & 6 & 9 & 1 & 7 & 5 \\ 1 & 6 & 5 & 2 & 3 & 4 \end{bmatrix}^T$$

It represents 6 sample points, each with two features f1 and f2. The labels for the data are given below:

$$[101010]^T$$

In this question, we build a decision tree of depth 2 by hand to classify the data.

(a) What is the entropy at the root of the tree? **(2 Marks)**

(b) What is the rule for the first split? Write your answer in a form like f1 ≥ 4 or f2 ≥ 3. Hint: you should be able to eyeball the best split without calculating the entropies. **(4 Marks)**

(c) For each of the two tree nodes after the first split, what is the rule for the second split? **(4 Marks)**

(d) Let's return to the root of the tree, and suppose we're incompetent tree builders. Is there a (not trivial) split at the root that would have given us an information gain of zero? Explain your answer. **(2 Marks)**

Q2. a.) You're solving a binary classification task. You first try a logistic regression. You initialize all weights to 0.5. Is this a good idea? Briefly explain why or why not. **(3 Marks)**

b.) Then, you try a 4-layer neural network. You initialize all weights to 0.5. Is this a good idea? Briefly explain why or why not. **(3 Marks)**

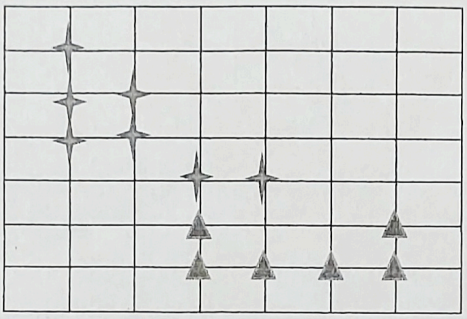
c.) Describe one advantage of using mini-batch gradient descent instead of full-batch gradient descent. **(2 Marks)**

d.) Describe one advantage of using mini-batch gradient descent instead of stochastic gradient descent with batch size 1. **(2 Marks)**

Q3. Answer the following questions with appropriate justification:

a.) Does increasing the number of hidden nodes in a multilayer perceptron improve generalisation? Why (not)? **(4 Marks)**

b.) Fig. 3 shows a data set with two real-valued inputs and one categorical output class. Positive points are shown as rectangles, negative points as triangles. Suppose that you are using a linear SVM with no provision for noise (i.e. a linear SVM that is trying to maximize its margin while ensuring that all data points are on their correct sides of 6 lOMoARcPSD|4424963 the margin). Draw three lines in the diagram showing the classification boundary and the two sides of the margin. Circle the support vectors. **(4 Marks)**



c.) How could you address the problem of recommending movies to a new user? **(4 Marks)**

d.) How could you accomplish this for a new movie? **(3 Marks)**

e.) What is the primary motivation for using the kernel trick in machine learning algorithms? **(3 Marks)**

f.) Overfitting is a central problem in learning. Suppose you have been given an unlabelled data set containing 1000 samples in 20 dimensions. You ran the K-means algorithms on it using 900 centroids, and you achieved a satisfying matching. Now, someone examines your work, and states that your algorithm is overfitting the data. What does she mean? How is this related to overfitting? **(4 Marks)**

Q4. Recall the loss function for k-means clustering with k clusters, sample points $x_1, ..., x_n$, and centres $\mu_1, ..., \mu_k$.

$$L = \sum_{j=1}^{k} \sum_{x_i \in s_j} \|x_i - u_j\|^2$$

where Sj refers to the set of data points that are closer to μj than to any other cluster mean.

Instead of updating μj by computing the mean, let's minimize L with batch gradient descent while holding the sets Sj fixed.

a.) Derive the update formula for μ1 with learning rate (step size)ε. **(6 Marks)**

b.) Derive the update formula for μ1 with stochastic gradient descent on a single sample point xi. Use learning rate, ε. **(4 Marks)**

c.) In this part, we will connect the batch gradient descent update equation with the standard k-means algorithm. Recall that in the update step of the standard algorithm, we assign each cluster center to be the mean (centroid) of the data points closest to that center. It turns out that a particular choice of the learning rate ε (which may be different

for each cluster) makes the two algorithms (batch gradient descent and the standard k-means algorithm) have identical update steps. Let's focus on the update for the first cluster, with center $\mu 1$. Calculate the value of ε so that both algorithms perform the same update for $\mu 1$. **(6 Marks)**

Q5. You are given a design matrix $X = \begin{matrix} 6 & -4 \\ -3 & 5 \\ -2 & 6 \\ 7 & -3 \end{matrix}$ Let's use PCA to reduce the dimension from 2 to 1.

a.) Compute the covariance matrix for the sample points. (Warning: Observe that X is not cantered.) Then compute the unit eigenvectors, and the corresponding eigenvalues, of the covariance matrix. **(6 Marks)**

b.) Suppose we use PCA to project the sample points onto a one-dimensional space. What one-dimensional subspace are we projecting onto? For each of the four sample points in X (not the cantered version of X!), write the coordinate (in principal coordinate space, not in R^2) that the point is projected to. **(6 Marks)**

c.) Given a design matrix X that is taller than it is wide, prove that every right singular vector of X with singular value σ is an eigenvector of the covariance matrix with eigenvalue σ^2. **(6 Marks)**

Q6. Given the following data

Item	X1	X2	Class
A	1	2	Yes = 1
B	2	1	Yes = 1
C	1	1	No = 0
D	1	0	No = 0

a) Are the data linearly separable? State reasons for your answer. **(2 Marks)**

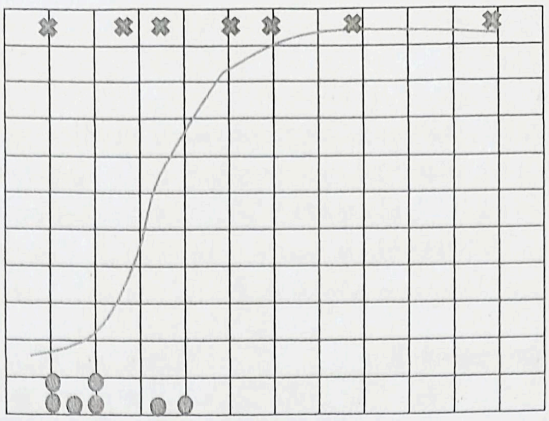
b) We will train a perceptron on the data. We add a bias $x0 = -1$ to each of the data points. Suppose the current weights to be $\boldsymbol{w} = (0, -1, 1)$. Assume a learning rate of 0.1. How should the weights be updated if point A is considered? How would the weights have been updated if the algorithm instead had considered point B? **(6 Marks)**

c) Say, we instead had applied a linear regression classifier. How should the weights have been updated when considering data point, A, again assuming a learning rate of 0.1. And how would they have been updated if we instead considered point B? **(2 Marks)**

Q7. Kim is building a spam filter. She has the hypothesis that counting the occurrences of the letter 'x' in the e-mails will be a good indicator of spam or no-spam. She collects 7 spam messages and 7 no spam messages and counts the number of x-s in each. Here is what she finds.

- Number of 'x'-s in each spam: [0, 3, 4, 8, 9, 13, 21]
- Number of 'x'-s in each no-spam: [0, 0, 1, 2, 2, 5, 6]

She trains a logistic regression classifier on the data and plots the classifier against the data.



Assume the logistic regression model and answer the following questions:

a) Consider an e-mail with no 'x'-s. According to the model, what is roughly the probability of this message being a spam message and what is the probability of it not being a spam. **(3 Marks)**

b) How many x-s must an e-mail contain to guarantee it is a spam mail? **(3 Marks)**

c) How is a logistic regression model normally turned into a binary classifier? If you turn the model into a classifier in this way, what is the accuracy of the classifier on the training data? **(3 Marks)**

d) It is most important that no no-spams are classified as spams. How can this goal be described in terms of precision and recall? How can the logistic regression classifier be modified to try to achieve this goal? **(3 Marks)**