Jahangirnagar University

Department of Computer Science and Engineering

4th Year 2nd Semester B.Sc. (Hons.) Final Examination -2019

Assignment for Final Examination

Course Title: **Pattern Recognition and Machine Learning**	Course No: **CSE-451**
Submission Deadline: **31/12/2020, 11:59 PM**	Full Marks: **10**

1.(2 Marks) DATA AND CLASSIFIER:

You have a collection of 1000 nature photographs which were taken under many different conditions. All of the images are of size 300 × 300 pixels. You wish to develop a binary classifier that labels a photograph as to whether or not it depicts a sunny day on a beach. The images have been pre-processed in the following manner:

- Each image i ∈{1. . .1000} is partitioned nine regions $R_{i,1. . .}R_{i,9}$. Each region is 100 × 100 pixels. The regions are arranged in a 3 × 3 grid, so that the region R_{i1} is the top-left corner of image i, the region R_{i2} is the top middle portion of the image, and so on.
- For each region $R_{i,j}$, we compute the average hue[1] of pixels within the region $R_{i,j}$. The hue value is quantised into 7 discrete bins: “red”, “orange”, “yellow”, “green”, “blue”, “indigo” and “violet”.

  (a) How would you represent this data in terms of attribute-value pairs?

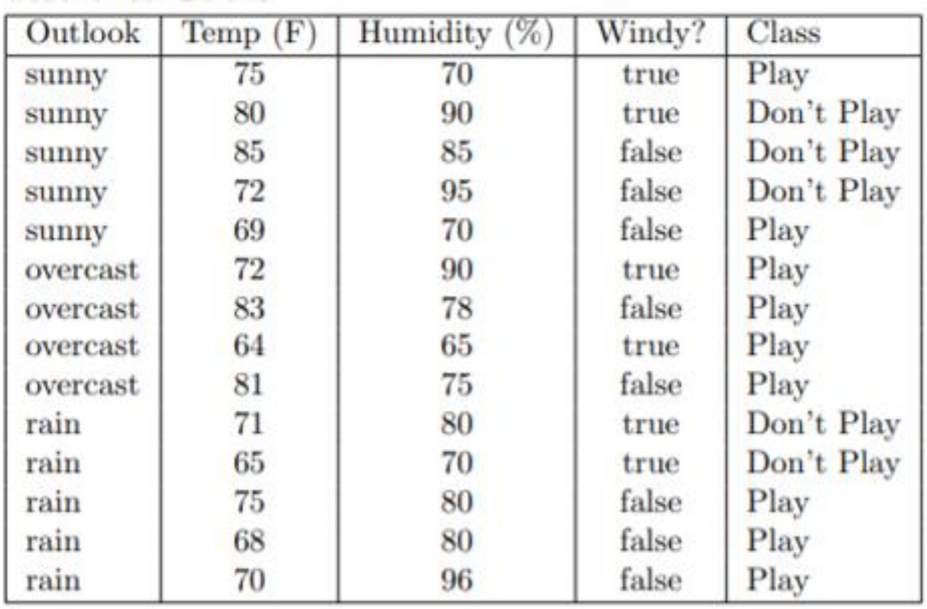
  (b) How many attributes are there? Are they categorical, ordinal or numeric?

  (c) What values can they take on?

[**1] The hue is a scalar representation of color. It ranges from 0◦ to 360◦. For example, colors with hues around 0◦ look red, hues around 120◦ look blue, and hues around 240◦ look green.

2. (2 MARKS) DECISION TREES:

The following table gives a data set for deciding whether to play or cancel a ball game, depending on the weather conditions.

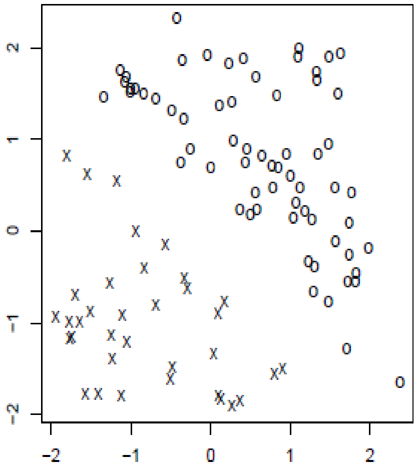


Outlook	Temp (F)	Humidity (%)	Windy?	Class
sunny	75	70	true	Play
sunny	80	90	true	Don't Play
sunny	85	85	false	Don't Play
sunny	72	95	false	Don't Play
sunny	69	70	false	Play
overcast	72	90	true	Play
overcast	83	78	false	Play
overcast	64	65	true	Play
overcast	81	75	false	Play
rain	71	80	true	Don't Play
rain	65	70	true	Don't Play
rain	75	80	false	Play
rain	68	80	false	Play
rain	70	96	false	Play

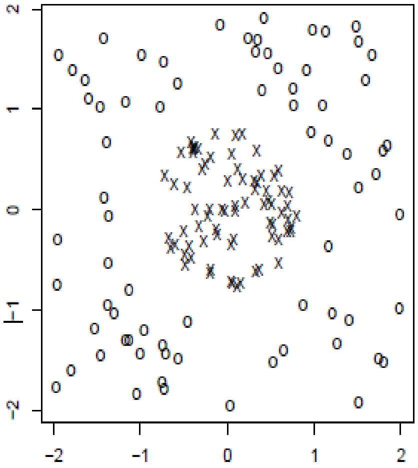
QUESTION: At the root node for a decision tree in this domain, what are the information gains associated with the Outlook and Humidity attributes? (Use a threshold of 75 for humidity (i.e., assume a binary split: humidity ≤ 75 / humidity > 75). Show your computations. Also draw the complete (unpruned) decision tree, showing the information gain at each non-leaf node, and class predictions at the leaves.

3. (2 MARKS) LINEAR CLASSIFICATION:

Consider the following two graphs below:



(I)



(II)

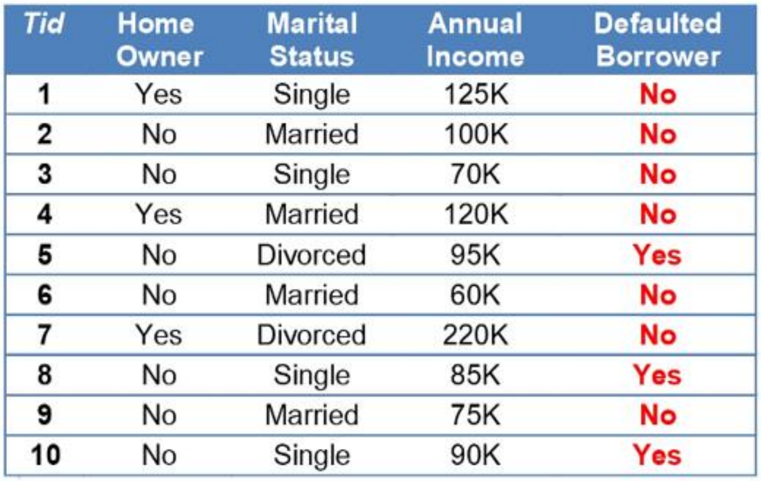
The two graphs each represent a classification problem. In each graph, the two axes x1 and x2 represent feature values, the open circles represent training points in the negative class, and the crosses represent points in the positive class. Consider applying a linear classifier, such as linear regression, to these data sets.

i. For training set (I), does there exist a weight vector such that the resulting classifier separates the data perfectly by class? If so, give an example. If not, explain why not.

ii. For training set (II), does there exist a weight vector such that the resulting classifier separates the data perfectly by class? If so, give an example. If not, explain why not.

4. (2 MARKS) k-NN ALGORITHM:

Consider the training examples shown in the following table for a binary classification. The table shows a training set for a problem of predicting whether a loan applicant will repay his/her loan obligation or defaulting on his/her loan.



Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

QUESTION: Using the kNN approach that we discussed in the class, predict the class label for this test example, X = (Home Owner = No, Marital Status = Married,Income = $120K). Assume that k=3 and distance is L2 norm. Please provide manual, step by step process.

5. (2 MARKS) VALIDATION AND COMPARING PERFORMANCE:

Two students are working on a machine-learning approach to spam detection. Each student has their own set of 100 labeled emails, 90% of which are used for training and 10% for validating the model. Student A runs the Naive Bayes algorithm and reports 80% accuracy on his validation set. Student B experiments with over 100 different learning algorithms, training each one on his training set, and recording the accuracy on the validation set. His best formulation achieves 90% accuracy. Whose algorithm would you pick for protecting a corporate network from spam? Why?

GOOD LUCK