BBM406 – Fall 2022

WRİTTEN ASSIGNMENT

Q1) Consider the following sequence of positive and negative training examples describing the concept "pairs of people who live in the same house." Each training example describes an ordered pair of people, with each person described by their sex, hair color (black, brown, or blonde), height (tall, medium, or short), and nationality (US, French, German, Irish, Indian, Japanese, or Portuguese).

\+ ⟨⟨male brown tall US⟩ ⟨female black short US⟩⟩

\+ ⟨⟨male brown short French⟩ ⟨female black short US⟩⟩

\- ⟨⟨female brown tall German⟩ ⟨female black short Indian⟩⟩

\+ ⟨⟨male brown tall Irish⟩ ⟨female brown short Irish⟩⟩

Consider a hypothesis space defined over these instances, in which each hypothesis is represented by a pair of 4-tuples, and where each attribute constraint may be a specific value, "?," or "Ø," just as in the EnjoySport hypothesis representation. For example, the hypothesis below :

⟨⟨male ? tall ?⟩ ⟨female ? ? Japanese⟩⟩

represents the set of all pairs of people where the first is a tall male (of any nationality and hair color), and the second is a Japanese female (of any hair color and height).

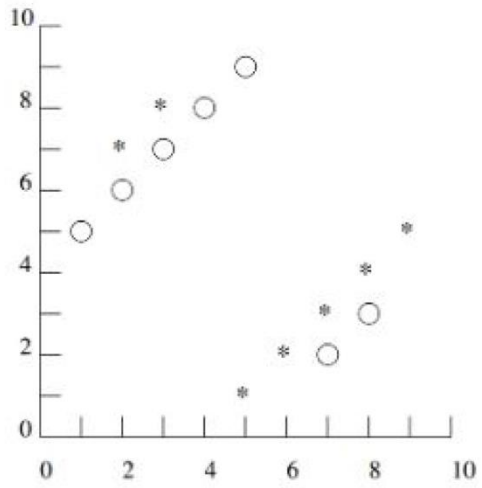
a) Provide a hand trace of the FIND-S algorithm learning from the above training examples and hypothesis language.
b) Provide a hand trace of the CANDIDATE-ELIMINATION algorithm learning from the above training examples and hypothesis language. In particular, show the specific and general boundaries of the version space after it has processed the first training example, then the second training example, etc.
c) How many distinct hypotheses of CANDIDATE-ELIMINATION algorithm are consistent with the following single positive training example?

\+ ⟨⟨male black short Portuguese⟩ ⟨female blonde tall Indian⟩⟩

Q2) Let k-NN(S) denote the k-Nearest Neighbor classifier on a sample set S, containing samples from 2 classes (positive, negative).

a) Show that if in both 1-NN(S_1) and 1-NN(S_2) the label of point x is positive, then in 1-NN($S_1 \cup S_2$) the label of x is positive.
b) Show an example such that in both 3-NN(S_1) and 3-NN(S_2) the label of x is positive, and in 3-NN($S_1 \cup S_2$) the label of x is negative.

Q3) One of the problems with k-nearest neighbor learning is how to select a value for k. Say you are given the following data set. This is a binary classification task in which the instances are described by two real-valued attributes ("*" and "O" denote positive and negative classes, respectively).



a) What value of k minimizes the training set error for this data set, and what is the resulting training set error? Why is training set error not a reasonable estimate of test set error, especially given this value of k?
b) What value of k minimizes the leave-one-out cross-validation error for this data set, and what is the resulting error? Why is cross-validation a better measure of test set performance?
c) Why might using too large values k be bad in this dataset? Why might too small values of k also be bad?
d) Sketch the 1-nearest neighbor decision boundary for this dataset.

Q4) Suppose that a dataset contains 1000 instances of patients with 950 instances of healthy patients, 50 instances of patients having diabetes. You are developing a machine learning model to classify these patients for diabetes disease.

a) Give an example showing that observing only accuracy metric is not enough for measuring classification performance of your model.
b) Give an example showing that using only recall or precision metric is not enough and we should use both of them.

Q5) Imagine that we develop an algorithm to predict spam e-mails. Based on the previous experience, we know that 97% of the mails are legitimate and 3% are spam. If an e-mail is spam, there is a 95% chance that the algorithm predict it as spam. If an e-mail is legit, the algorithm classifies it as spam with 50% chance. What is the probability that an e-mail is actually spam if the algorithm predict it as spam?

Q6)

Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	Hiigh	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	High	Drug B
p8	Young	F	Normal	Normal	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	Normal	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A
p15	Middle-age	F	Low	Normal	?

With respect to the dataset table above;

a) Construct a decision tree model and classify the p15 sample with the decision tree model you constructed
b) Construct a naive bayes model and classify the p15 sample with the naive bayes model you constructed