Hacettepe University

Computer Engineering Department

BBM 406 - Fundamentals of Machine Learning - 2022 Fall

Written Assignment - 1

November 21, 2022

Student names:	*Student Number:*
Numan Kafadar	b21946242
Umut Güngör	b21946198

1 Question 1

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 "∅" 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).

1.1 Provide a hand trace of the FIND-S algorithm learning from the above training examples and hypothesis language.

H0=⟨⟨∅ ∅ ∅ ∅⟩⟨∅ ∅ ∅ ∅⟩⟩
H1=⟨⟨male brown tall US⟩ ⟨female black short US⟩⟩
H2=⟨⟨male brown ? ?⟩ ⟨female black short US⟩⟩
H3=⟨⟨male brown ? ?⟩ ⟨female black short US⟩⟩
H4=⟨⟨male brown ? ?⟩ ⟨female ? short ?⟩⟩⟩

1.2 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.

S0=⟨⟨∅ ∅ ∅ ∅⟩⟨∅ ∅ ∅ ∅⟩⟩
G0=⟨⟨? ? ? ?⟩ ⟨? ? ? ?⟩⟩

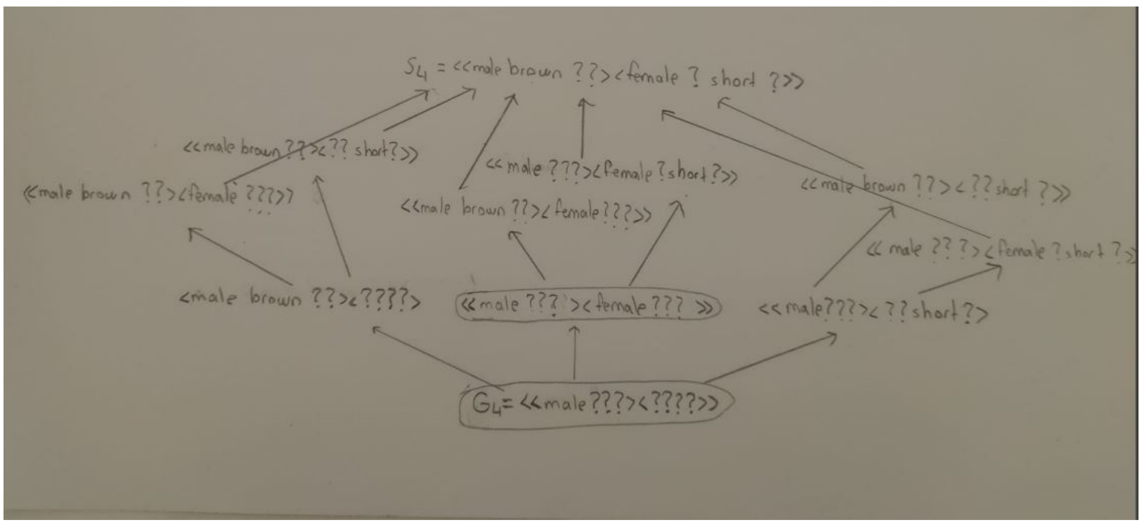
S1=⟨⟨male brown tall US⟩ ⟨female black short US⟩
G1=⟨⟨? ? ? ?⟩ ⟨? ? ? ?⟩⟩

S2=⟨⟨male brown ? ?⟩ ⟨female black short US⟩⟩
G2=⟨⟨? ? ? ?⟩ ⟨? ? ? ?⟩⟩

S3=⟨⟨male brown ? ?⟩ ⟨female black short US⟩⟩
G3=(⟨⟨male ? ? ?⟩ ⟨? ? ? ?⟩⟩,⟨⟨? ? ? ?⟩ ⟨? ? ? US⟩⟩)

S4=⟨⟨male brown ? ?⟩ ⟨female ? short ?⟩⟩
G4=(⟨⟨male ? ? ?⟩ ⟨? ? ? ?⟩⟩)

1.3 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⟩⟩



If we examine the figure above, the hypothesises that consistent with test example are :
⟨⟨male ? ? ?⟩ ⟨? ? ? ?⟩⟩
⟨⟨male ? ? ?⟩ ⟨female ? ? ?⟩⟩
So ,there are 2 distinct hypotheses that consistent with test example.

2 Question 2

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

2.1 Show that if in both 1-NN[(S1) and 1-NN(S2) the label of point x is positive, then in 1-NN(S1 U S2) the label of x is positive.

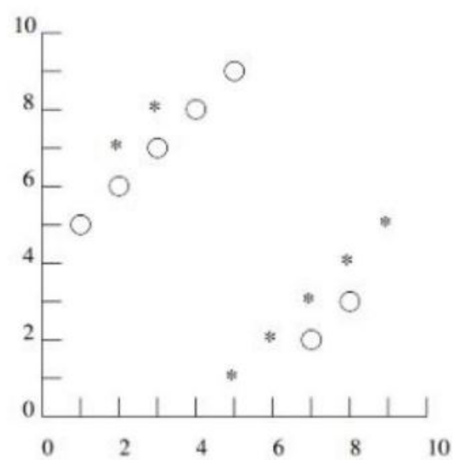
We assume that, the nearest sample of x in S1 is z1 and in S2 is z2. Labels of both is z1, z2 are positive because, for the label of x to be positive in S1, z1 must have positive level. Likewise for z2. So, in sample set S1 U S2 label of the x is either label of z1 or label of z2. Since labels of both z1 and z2 are positive, label of x in S1 U S2 must be be positive.

2.2 Show an example such that in both 3-NN[(S1) and 3-NN(S2) the label of x is positive, and in 3-NN(S1 U S2) the label of x is negative.

We assume that, the 3-nearest sample of x in S1 is y1, z1 and t1. And the 3-nearest sample of x in S2 is y2, z2 and t2. Distance between x and y1 is 4, x and z1 is 3, x and t1 is 2. Labels of y1 and z1 are positive but label of t1 is negative. The label of x is positive in S1 because the majority of votes are positive. Distance between x and y2 is 1, x and z2 is 5, x and t2 is 6. Labels of z2 and t2 are positive but label of y2 is negative. The label of x is positive in S2 because majority of votes are positive. However , in sample set S1 U S2, the 3-nearest sample of x are t1, z1, and y2. Labels of t1 and y2 are negative but label of z1 is positive. Majority of votes are negative. Therefore label of x in S1 U S2 must be negative.

3 Question 3

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).



3.1 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?

k=1 minimizes the training set error.

Resulting training set error when we choose k=1 is 0.Training set error is not a reasonable estimate of test set error because test set instances are new data and they are not in training set. If we use 1 as a k-value, in the test set there will be many misclassified instances.Because, since we use 1 as a k-value, the model predicts by looking at the closest sample, the model is going to be error-prone. The model is very complex and this causes over fitting.

3.2 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?

K=5 minimizes the LOOCV error.

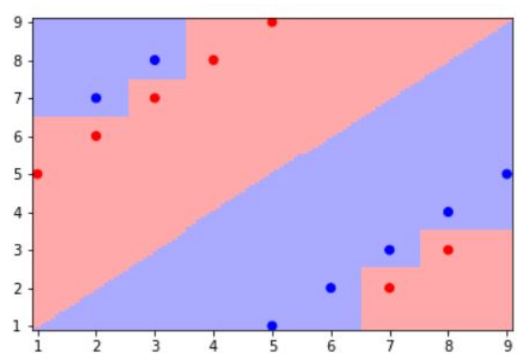
Resulting error when we choose k=5 is 0.29. LOOCV is a better measure of test set performance because in LOOCV validation data is not in the training set.So, the instance that is going to be predict is an unknown instance for the model. It is similar to the test set evaluation.

3.3 Why might using too large values k be bad in this dataset? Why might too small values of k also be bad?

Using too large k-values causes under fitting. The model should be more complex.Under fitting decreases the accuracy of the model. Under fitting simply means that the model does not fit the data well enough.

Using too small k-values causes over fitting. The model should be less complex Over fitting decreases the accuracy of the model too. The model does not label data correctly, because of too many details and noise.

3.4 Sketch the 1-nearest neighbor decision boundary for this dataset.



4 Question 4

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.

4.1 Give an example showing that observing only accuracy metric is not enough for measuring classification performance of your model.

	Diabetes	Healthy
Diabetes	TP=0	FP=0
Healthy	FN=50	TN=950

According to accuracy matrix,the model is good.However,it is quite bad because it classifies every patient as healthy.Recall value of this model is zero.So ,model is really bad since it cannot find any patient with diabetes. Recall value should be high.

4.2 Give an example showing that using only recall or precision metric is not enough and we should use both of them.

Figure-1	Diabetes	Healthy	Figure-2	Diabetes	Healthy
Diabetes	TP=50	FP=950	Diabetes	TP=1	FP=0
Healthy	FN=0	TN=0	Healthy	FN=49	TN=950

Figure-1:Precision=50/100=0.05 Recall=50/50=1
Figure-2:Precision=1/1=1 Recall=1/50=0.02

In Figure-1,the model classifies every patient as diabetes. According to the recall metric,the performance of the model should be good.However,since precision value is so low,model is bad.Precision value should be high.

In Figure-2, precision value is 1 but recall value is 0.02. The model is bad since it has low recall

5 Question 5

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?

Bayes' Theorem

$$P(A|B) = \frac{P(A) * P(B|A)}{P(B)}$$

A=an email is spam
B—A=the algorithm predicts as spam if an email is spam

B=the algorithm predicts as spam

$$P(A|B) = \frac{0.03 * 0.95}{0.03 * 0.95 + 0.97 * 0.50} = 0.06$$

6 Question 6

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	?

Figure 1: Dataset

With respect to the dataset table above;

6.1 Construct a decision tree model and classify the p15 sample with the decision tree model you constructed

First , we need to calculate each attributes information gain and then we can decide our root and childs nodes.

A formula of Entropy and Information Gain is:

$$Entropy(S) = \sum_{i=1}^{n} -p_i log_2 p_i$$

$$Gain(S,A) = Entropy(S) - \sum_{v:Values(A)} (|S_v|/|S|)Entropy(S_v)$$

So ,the information gain for "Age" Attribute is

$$Gain(Dataset, Age) = Entropy(Dataset) - (5/14)Entropy(Dataset_{Young})$$
$$-(4/14)Entropy(Dataset_{Middle-age}) - (5/14)Entropy(Dataset_{Senior})$$

Gain(Age) : 0.24674981977443933
Similarly;
Gain(Sex) : 0.15183550136234159
Gain(BP) : 0.02922256565895487
Gain(Cholesterol): 0.04812703040826949

Since "Age" has maximum information gain value,root is "Age". We do the same process for remained attributes and tree is as follows:

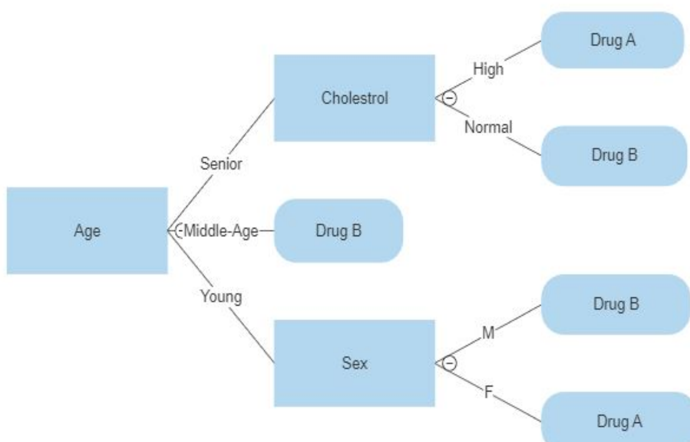


Figure 2: Decision Tree

If we classify p15, since its "Age" attribute value is "Middle-age",it classified as "Drug B"

6.2 Construct a naive bayes model and classify the p15 sample with the naive bayes model you constructed

Naive Bayes Classifier Formula is:

$$v_{NB} = argmax_{v_j:V} P(v_j) \prod_i P(a_i|v_j)$$

Since we are trying to predict p15, values will be as follows:

$$P(DrugA)P(Middle-Age|DrugA)P(F|DrugA)P(Low|DrugA)P(Normal|DrugA)$$
$$(5/14)*(0/5)*(4/5)*(1/5)*(2/5) = 0$$
$$P(DrugB)P(Middle-Age|DrugB)P(F|DrugB)P(Low|DrugB)P(Normal|DrugB)$$
$$(9/14)*(4/9)*(3/9)*(3/9)*(6/9) = 0.02116$$

Thus ,the naive Bayes classifier assigns the target value "Drug =Drug B" to this new instance, based on the probability estimates learned from the training data.