# Homework 2

Your Name: NARENDRA RAGAV VENKATESAN

Student ID: A20480789

**1. (30 points) Classification: We are going to make a decision about whether an animal is useful (P) or useless (N) in our experiments. We measure their age in days, whether fat or not, and the size of their soles of the feet.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Left Back** | **Right Back** | **Left Front** | **Right Front** | **Fat** | **Age** | **Label** |
| 4.1 | 4.8 | 1.6 | 3.2 | Yes | 100 | P |
| 4.6 | 4.2 | 1.4 | 0.2 | Yes | 40 | N |
| 4.3 | 5.0 | 1.5 | 4.2 | No | 160 | N |
| 5 | 1.3 | 1.4 | 2.2 | Yes | 90 | P |
| 5 | 1.2 | 4.7 | 1.4 | No | 40 | N |
| 4.4 | 3.2 | 4.5 | 1.5 | No | 80 | P |
| 4.9 | 3.1 | 4.9 | 1.5 | No | 100 | N |
| 2.5 | 1.3 | 4 | 1.3 | Yes | 110 | P |
| 4.5 | 2.8 | 4.6 | 1.5 | Yes | 120 | P |
| 4.3 | 3.3 | 4.9 | 2.5 | Yes | 30 | N |
| 1.8 | 2.7 | 5.0 | 1.9 | Yes | 20 | P |
| 2.1 | 3 | 5.0 | 2.1 | No | 40 | N |
| 4.3 | 2.9 | 5.0 | 1.8 | No | 30 | P |
| 4.5 | 3 | 4.9 | 2.2 | No | 50 | N |
| **4.3** | **3.6** | **1.5** | **1.8** | **Yes** | **70** | **?** |

a). [5 points] Do we need normalization and discretization (data type transformation) to use KNN classifier? Why (use your own text/description)?

Ans: It require normalization and discretization to use KNN classifier as the model requires to calculate distance which can be done only through numeric data. The major requirement to normalization is derived from the fact of computed distance is majorly focused on large features. So, the ages in the table above should be normalized.

b). [5 points] If your answer is Yes in part 1), please apply normalization (to new scale [1,5]) and discretization. Give the process of preprocessing and the table of final data

Ans In the above table I have changed the data type of “Fat Column” from yes/no to numerical data wherein “Yes” is given a value of 1 and “No” is given a value of 0.

I have normalized the age column on a scale of [1,5].

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Left Back** | **Right Back** | **Left Front** | **Right Front** | **Fat** | **Age** | **Label** |
| 4.1 | 4.8 | 1.6 | 3.2 | 1 | 3.3 | P |
| 4.6 | 4.2 | 1.4 | 0.2 | 1 | 1.6 | N |
| 4.3 | 5.0 | 1.5 | 4.2 | 0 | 5 | N |
| 5 | 1.3 | 1.4 | 2.2 | 1 | 3 | P |
| 5 | 1.2 | 4.7 | 1.4 | 0 | 1.6 | N |
| 4.4 | 3.2 | 4.5 | 1.5 | 0 | 2.7 | P |
| 4.9 | 3.1 | 4.9 | 1.5 | 0 | 3.3 | N |
| 2.5 | 1.3 | 4 | 1.3 | 1 | 3.6 | P |
| 4.5 | 2.8 | 4.6 | 1.5 | 1 | 3.8 | P |
| 4.3 | 3.3 | 4.9 | 2.5 | 1 | 1.3 | N |
| 1.8 | 2.7 | 5.0 | 1.9 | 1 | 1 | P |
| 2.1 | 3 | 5.0 | 2.1 | 0 | 1.6 | N |
| 4.3 | 2.9 | 5.0 | 1.8 | 0 | 1.3 | P |
| 4.5 | 3 | 4.9 | 2.2 | 0 | 1.8 | N |
| **4.3** | **3.6** | **1.5** | **1.8** | **1** | **2.4** | **?** |

c). [20 points] Use first 10 rows as training, the next 4 rows as testing set. Apply KNN Classifier to the new data table in part b). In other words, build your KNN classifier by the following requirements based on the knowledge in the table, and then predict which class/label the object (in red) belongs to:

* Distance measures: Manhattan distance
* K = 1, 3, 5

Find the best K value by examining F1 score on the test set. Finally apply the best setting and predict the label for the unseen data.

Considering the Training Data set Manhattan distance formula is

dist(X, Y) = |X1-Y1|+|X2-Y2|…..|Xn-Yn|

Since K =1,

P1’KNN = {P2}

= |4.1-4.6| + |4.8-4.2|+|1.6-1.4|+|3.2-0.2|+|1-1|+|3.3-1.6|

= 0.5+0.6+0.2+3+0+1.7

=6

K=1 is 6.

The label for P1’KNN = {P2} is “N”

When K =3,

P2’KNN= {P3, P5, P1}

{P3}

= |4.6-4.3|+|4.2-5.0|+|1.4-1.5|+|0.2-4.2|+|1-0|+|1.6-5|

= 6.3

{P5}

=|4.6-5|+|4.2-1.2|+|1.4-4.7|+|0.2-1.4|+|1-0|+|1.6-1.6|

=8.9

{P1}

=|4.6-4.1|+|4.2-4.8|+|1.4-1.6|+|0.2-3.2|+|1-1|+|1.6-3.3|

=6.6

Therefore, K =3 is {6.3,8.9,6.6}

Ranking is {P3, P1, P5} i,e {N,P,N} frequently most repeated is “N”

Prediction for P2 is “N”.

K=5,

P5’KNN= {P1, P2, P4, P7, P9}

{P1}

=|5.0-4.1|+|1.2-4.8|+|4.7-1.6|+|1.4-3.2|+|0-1|+|1.6-3.3|

=12.5

{P2}

=|5.0-4.6|+|1.2-4.2|+|4.7-1.4|+|1.4-0.2|+|0-1|+|1.6-1.6|

=8.9

{P4}

=|5.0-5.0|+|1.2-1.3|+|4.7-1.4|+|1.4-2.2|+|0-1|+|1.6-3.0|

=6.6

{P7}

=|5.0-4.9|+|1.2-3.1|+|4.7-4.9|+|1.4-1.5|+|0-0|+|1.6-3.3|

=4

{P9}

=|5.0-4.5|+|1.2-2.8|+|4.7-4.6|+|1.4-1.5|+|0-1|+|1.6-3.8|

=5.5

Therefore, value for K5 is {12.5,8.9,6.6,4,5.5}

Ranking is {P7, P9, P4, P2, P1} which is {N, P, P, N, P} and frequent label is P. Therefore, prediction for P5 is “P”

Testing data set:

K=1

P12-{P13}

=|2.1-4.3|+|3.0-2.9|+|5.0-5.0|+|2.1-1.8|+|0-0|+|1.6-1.3|

=2.2+0.1+0+0.3+0+0.3

=2.9

K=3

P11- {P12, P13, P15}

P12

=|1.8-2.1|+|2.7-3.0|+|5.0-5.0|+|1.9-2.1|+|1-0|+|1.0-1.6|

=0.3+0.3+0+0.2+1+0.6

=2.4

P13

=|1.8-4.3|+|2.7-2.9|+|5.0-5.0|+|1.9-1.8|+|1-0|+|1.0-1.3|

=2.5+0.2+0+0.1+1+0.3

=4.1

P15

=|1.8-4.3|+|2.7-3.6|+|5.0-1.5|+|1.9-1.8|+|1-1|+|1.0-2.4|

=2.5+0.9+3.5+0.1+0+1.4

=8.4

Hence, K=3 is {2.4,4.1,8.4}

Ranking for the same is {P12, P13, P15} Which is {N, P, N} and most frequent label is ‘N’

Therefore, Prediction for P15 is “N”

F1 score is given by F=2\* where,

Precision = TP/(TP+FP)

Recall = TP/(TP+FN)

**Make sure that you clearly show each step. You cannot simply give me the final answer without showing how did you do that.**

2. (30 points) Use Naïve Bayes Classifier to classify the objects

We conducted a survey to collect people’s daily diets and try to build a model to predict whether their diets result in healthy conditions or not. The final results could be Yes, No, Unsure

|  |  |  |  |
| --- | --- | --- | --- |
| **Breakfast** | **Lunch** | **Dinner** | **Healthy?** |
| Ham | Carnivorous | Beef | Y |
| Milk | Carnivorous | Beef | N |
| Bread | Veggie | Pork | U |
| Bread | Veggie | Veggie | Y |
| Ham | Veggie | Veggie | Y |
| Bread | Carnivorous | Beef | N |
| Ham | Veggie | Pork | N |
| Milk | Veggie | Pork | U |
| Milk | Carnivorous | Veggie | U |
| Noddle | Carnivorous | Pork | ? |

1). [10 points] What is laplace smoothing? And why we need it in the Naïve Bayesian classifier?

Ans Laplace Smoothing is the process that helps solve the problem of zero probability. The higher value of alpha will drive the likelihood towards 0.5, i,e the probability of a term which is equal to 0.5 for positive and negative tests.

The formula for Laplace smoothing is given by P(w’|positive) as P(w’|positive) =

Where,

Alpha is smoothing parameter,

K is the number of dimensions in the data

N represents the number of reviews with y = positive

2). [20 points] Using the Naive Bayesian Classification Hint: you may need to use laplace smoothing (use the formula in our slide) if you do have zero-conditional probabilities. Use the setting in the slide to solve the problems in this case. Note, only apply laplace smoothing to the ones you have zero-conditional probabilities.

Steps for Prior Probability:

Probability (Healthy = Y) = 3/9 = 0.33

Probability (Healthy = N) = 3/9 = 0.33

Probability (Healthy = N) = 3/9 = 0.33

Conditional Probability:

|  |
| --- |
| **Breakfast:** **Y N U** |
| Ham 2/3 1/3 0 |
| Milk 0 1/3 2/3 |
| Bread 1/3 1/3 1/3 |
| **Lunch**   **Y N U** |
| Carnivorous 1/3 2/3 1/3 |
| Veggie 2/3 1/3 2/3 |
| **Dinner**  **Y N U** |
| Beef 1/3 2/3 0 |
| Pork 0 1/3 2/3 |
| Veggie 2/3 0 1/3 |

Let’s Consider e1 = Breakfast

e2 = Lunch

e3 = Dinner

c1 = Y

c2 = N

c3 = U

To find a fresh Noodle for Breakfast, Carnivorous for Lunch and Pork for Dinner instance.

Therefore, Prop of new instance to be yes p(e1|c1)

P(e1|c1) = 0/3 = 0

As Zero probability error occurs, we now use Laplace smoothing for the below calculations.

= (0+9(1/3))/3+9 = 0.25

P(e2|c1) =1/3 = 0.33

P(e3|c1) = 0

Since a zero-probability error has occurred Laplace smoothing needs to be applied

P(e1|c1) = 0.25\*0.33\*0.25+0.020

N-P(e2|c2) is the probability of new instance

P(e1|c2) = 0

Now Laplace smoothing we get 0.25

P(e2|c2) = 0.25

P(e2|c2) = 2/3 = 0.66

P(e3|c2) = 1/3 = 0.33

P(e2|c2) = 0.25\*0.66\*0.33 = 0.054

U-P(e3|c3) is the probability of new instance

P(e1|c3) = 0 on applying laplace smoothing and got 0.25

P(e2|c3) = 1/3 = 0.33

P(e3|c3) = 2/3 = 0.66

P(E) = P(c1) \*P(e1|c1) +P(e2|c2) +P(e3|c3)

= 1/3\*0.020+1/3\*0.054+1/3\*0.054

= 0.04

P(c1|E) = (1/3\*P(e1|c1))/P(E)

= 1/3\*0.020/0.04

=0.17

P(c2|E) = 1/3\*0.054/0.04

= 0.45

P(c3|E) = 1/3\*0.054/0.04

= 0.45

With respect to both N and U for the new instance of Breakfast as Noodle, Carnivorous for Lunch and Pork for Dinner as P(c2|E) and P(c3|E) have values.

1. Build a decision Tree [40]

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Color** | **Size** | **Act** | **Age** | **Inflated** |
| YELLOW | SMALL | STRETCH | Young | T |
| YELLOW | SMALL | STRETCH | Old | T |
| YELLOW | SMALL | STRETCH | Old | T |
| YELLOW | SMALL | DIP | Kid | F |
| YELLOW | SMALL | DIP | Kid | T |
| YELLOW | LARGE | STRETCH | Old | T |
| YELLOW | LARGE | STRETCH | Old | T |
| YELLOW | LARGE | DIP | Young | T |
| YELLOW | LARGE | DIP | Young | T |
| YELLOW | LARGE | DIP | Young | F |
| PURPLE | SMALL | STRETCH | Young | F |
| PURPLE | SMALL | STRETCH | Old | T |
| PURPLE | SMALL | STRETCH | Old | F |
| PURPLE | SMALL | DIP | Kid | T |
| PURPLE | SMALL | DIP | Kid | F |
| Blue | LARGE | STRETCH | Kid | T |
| Blue | LARGE | DIP | Young | F |
| Blue | LARGE | DIP | Young | F |
| Blue | LARGE | DIP | Old | T |
| Blue | LARGE | DIP | Young | T |

1). The data above is used to classify whether a balloon is inflated or not. [30]

In this question, you need to build a decision tree, but only build the top-1 and top-2 levels – in other words, build the root note and build the next level. On the second level, you need to figure out the features to be filled too, but you do not need to find the features to be filled in the 3rd level

You should show the process and calculations about how to build the tree.

Note: if two variables have the same largest information gain, you should choose the one with less number of the values in the variable (i.e., fewer branches). For example, if you find that act and age have the same information gain value, and the value is the largest one, then you choose act instead of age, since there are only two values in the variable act

Consider Entire data set,

S = [13+, 7-]

Entropy (S) = -

= 0.93

Values(color) = Yellow, Purple, Blue

Sy = [8+,2-] = -

= 0.72

Sp = [2+,3-] = -

= 0.9709

Sb = [3+,2-] = -

= 0.9709

Gain (S, Color) = 0.93-

= 0.085

Values (Size) = small, Large

S = [6+,7-]

Ss = [6+4-] =-

= 0.9709

S1 = [7+3-] = -

= 0.8812

Gain (S, Size) = 0.93-

= 0.005

Values (Act) = Stretch, Dip

Sst = [7+,2-] = -

= 0.764

Sd = [6+,5-] = -

= 0.994

Gain (Act) = 0.93-

= 0.0395

Values (Age) = Young, Old, Kid

Sy [4+,4-] = 1

So [6+, 1-] = -

=0.5916

Sk[3+,2-] = -

= 0.9709

Gain (S, Age) = 0.93-

= 0.082

2). On the 3rd level of the tree, use the label on this level – you can simply use the majority of the label in each group as the predicted label. Given a new data “YELLOW, SMALL, DIP, YOUNG”, make your prediction [5]

Sc [13+,7-]

Therefore,

Color

Yellow. Purple. Blue

Yellow

E (S, Sy) = 0.72

E (Y, Ss) = [4+,1-] = -

= 0.7219

Size = 0.72-

= 0

E (Sy, stretch) = [5+,0-] = 0

E [Sy, Old) [3+, 2-] = -

= 0.97

Act = 0.72-7

= 0.235

E (Sy, Young) = [3+,1-] = -

= 0.8112

E (Sy, Old) = [4 +, 0-]

= 0

E (Sy, Kid) = [1+, 1]

= 1

Age = 0.72-

= 0.19

We can infer that from above computation Act has highest information Gain.

Purple-

E (S, Sp) = 0.9709

Size = E(Small) = [2+,3-] = -

= 0.9709

E(large) = 0

Size = 0.97-0.9709

= 0

Act = Stretch = [1+,2-] = -

= 0.918

Dip = [1+,1-] = 1

0.98-\*0.918-

= 0.0192

Age = Young [0+, 1-] = 0

Old [1+,1-] = 1

Young [1+, 1-] = 1

0.97-\*0 –

= 0.17

From the above calculation we can infer that Age has the Highest Information Gain

Blue = E (S, Sb) = 0.9709

Size = Large [3+,2-] = 0.9709

Small = 0

E (Blue, Ss) = 0.9709-\*0.9709

= 0

DIP [2+,2-] = 1

Act = 0.9709 -\*0 -\*1

= 0.1709

Age

Sk [1+,0-]

= 0

Sy [1+,2-] = 0.918

So [1+, 0-] = 1

0.9709 - \*0.918

= 0.4192

We can infer that Age has the Highest Information Gain

3). List at least three solutions to alleviate overfitting in decision tree [5]

The three solutions to alleviate overfitting in decision tree are:

1. Pre-pruning
2. Post-Pruning
3. Estimation of errors /Testing for significance- Chi-square test/tree trimming.