CAP-5615 Intro to Neural Networks Summer 2018
Homework 2
Christian Garbin

Question 1 [1.5 ps: 0.25/each]

Please use your own language to briefly explain the following concepts (Must use your own language. No credit if descriptions are copied from external sources):

Product Rule in Probability Theory:

The Product Rule calculates the probability of two events occurring together (their joint probability). For dependent events it results in $P(AB) = P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$ (the conditional probability modified by the prior probability). For independent events it simplifies to $P(AB) = P(A)P(B)$.

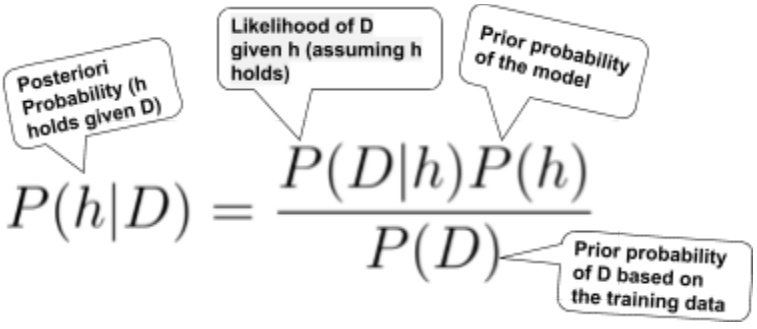
Sum Rule in Probability Theory:

The Sum Rule is used to calculate the probability of any one of several events occurring. For two events: $P(A + B) = P(A \cup B) = P(A) + P(B) - P(AB)$. The last term in the equation removes the instances where A and B overlap (their joint probability) to prevent double-counting.

Bayes Rule (in terms of priori probability, likelihood, posteriori probability):

The Bayes Rule allows us to calculate the *a posteriori* probability of a hypothesis, given an observed instance or set of instances, using the prior probability of the model, the prior probability of the observed instance and the likelihood of the observed instance, given the hypothesis.
In the picture below:

- P(h) is *a priori* probability of the hypothesis, based on what we know about the data or assumed to be uniformly distributed if we don't have any other piece of information
- P(D) is a priori probability of the instance, based on what we know about the training data (independent of a hypothesis)
- P(D|h) is the likelihood of the instance, given the hypothesis (i.e. assuming that the hypothesis holds)
- P(h|D) is the *a posteriori* probability of the hypothesis (given the instance, what is the probability that the hypothesis holds)



Maximum A Posteriori Estimation (MAP):

MAP (maximum *a posteriori*) is the most probable hypothesis for a given training data. It's found by calculating the *a posteriori* probability of each possible hypothesis in a set, and choosing the one with highest probability.

Conditional Independence:

Conditional Independence happens when a set of features are independent from each other, given the presence of a condition. In other words, if we cannot make any prediction about event B given event A when a condition is present, then A and B and conditionally independent *in the presence of that condition*.

Naive Bayes Classification:

Naive Bayes Classification is a simplification of the Bayes Optimal Classifier to make the calculation of hypothesis computationally feasible. The Bayes Optimal Classifier uses the joint probability of all attributes, while the Naive Bayes Classification assumes that the attributes are independent given a condition (conditionally independent) and uses the product of the probability of each attribute. This assumption makes it simpler to calculate.

Question 2 [2 pts]

A patient takes a lab test and the result comes back positive. Assume the test returns a correct positive result in only 95% of the cases in which the disease is actually present, and a correct negative result in only 95% of the cases in which the disease is not present. Assume further that 0.001 of the entire population have this cancer. Please use Bayes Rule to derive the probability of the patient having the disease given that his/her lab test is positive (please list the major steps).

$$P(disease) = 0.001, P(\neg disease) = 0.999$$
$$P(+|disease) = 0.95, P(-|disease) = 0.05$$
$$P(+|\neg disease) = 0.05, P(-|\neg disease) = 0.95$$

$$P(+) = P(+|disease)P(disease) + P(+|\neg disease)P(\neg disease)$$
$$P(+) = 0.95 \times 0.001 + 0.05 \times 0.999 = 0.0509$$

$$P(disease|+) = \frac{P(+|disease)P(disease)}{P(+)} = \frac{0.95 \times 0.001}{0.0509} = 0.019$$

Question 3 [2.5 pts]

Given three boxes: Black (B), Red (R), Green (G), where Black box has 3 apples and 4 oranges, Red box has 5 apples 1 oranges, and Green box has 2 apples 5 oranges. Assume a person picks Black box, Red Box, or Green box with 50%, 30%, and 20% probability, respectively, and then pick a fruit from the selected box.

1. What is the overall chance that an Apple will be selected? [1 pt]

$P(B) = 0.5, P(R) = 0.3, P(G) = 0.2$

$$P(Apple) = P(B)P(Apple|B) + P(R)P(Apple|R) + P(G)P(Apple|G)$$
$$P(Apple) = 0.5 \times \frac{3}{3+4} + 0.3 \times \frac{5}{5+1} + 0.2 \times \frac{2}{2+5} = 0.521$$

2. If the fruit selected is an Apple, what is the probability that the Apple was selected from the Green box? [1.5 pts]

$$P(G|Apple) = \frac{P(Apple|G)P(G)}{P(Apple)} = \frac{\frac{2}{2+5} \times 0.2}{0.521} = 0.11$$

Question 4 [3 pts]

In database showing in Table 1, please manually construct a Naïve Bayes Classifier.

ID	Outlook	Temperature	Humidity	Wind	Class
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Mild	Normal	Weak	No
14	Rain	Hot	High	Strong	Yes
15	Rain	Mild	High	Strong	No

Table 1

Please report priori probabilities and the conditional probabilities [2 pts].

P(Yes) = 9/15
P(No) = 6/15

Outlook	
P(Overcast|Yes) = 3/9	P(Overcast|No) = 1/6
P(Rain|Yes) = 4/9	P(Rain|No) = 2/6
P(Sunny|Yes) = 2/9	P(Sunny|No) = 3/6
Temperature	
P(Cool|Yes) = 3/9	P(Cool|No) = 1/6
P(Hot|Yes) = 2/9	P(Hot|No) = 2/6
P(Mild|Yes) = 4/9	P(Mild|No) = 3/6
Humidity	
P(High|Yes) = 4/9	P(High|No) = 4/6
P(Normal|Yes) = 5/9	P(Normal|No) = 2/6
Wind	
P(Strong|Yes) = 4/9	P(Strong|No) = 3/6
P(Weak|Yes) = 5/9	P(Weak|No) = 3/6

Please use your Naïve Bayes classifier to determine whether a person should play tennis or not, under conditions that “Outlook=Overcast & Temperature=Hot & Humidity =Normal& Wind=Weak”. Why or why not? [1 pt]

$$P(Yes)P(Overcast|Yes)P(Hot|Yes)P(Normal|Yes)P(Weak|Yes) = \frac{9}{15} \times \frac{3}{9} \times \frac{2}{9} \times \frac{5}{9} \times \frac{5}{9} = 0.0137$$

$$P(No)P(Overcast|No)P(Hot|No)P(Normal|No)P(Weak|No) = \frac{6}{15} \times \frac{1}{6} \times \frac{2}{6} \times \frac{2}{6} \times \frac{3}{6} = 0.0037$$

The person should play tennis, given that the conditional probability for Class = Yes > conditional probability for Class = No.

Question 5 [2 pts]

In database showing in Table 1, please manually construct a Naïve Bayes Classifier, by using m-estimate to calculate the conditional probabilities (m=1, and p equals to 1 divided by the number of attribute values for each attribute).

Please report priori probabilities and the conditional probabilities [1 pt].

P(Yes) = 9/15
P(No) = 6/15

Outlook p = 1/3	
P(Overcast|Yes) = (3 + 1/3)/(9 + 1) = 10/30	P(Overcast|No) = (1 + 1/3)/(6 + 1) = 4/21
P(Rain|Yes) = (4 + 1/3)/(9 + 1) = 13/30	P(Rain|No) = (2 + 1/3)/(6 + 1) = 7/21
P(Sunny|Yes) = (2 + 1/3)/(9 + 1) = 7/30	P(Sunny|No) = (3 + 1/3)/(6 + 1) = 10/21
Temperature p = 1/3	
P(Cool|Yes) = (3 + 1/3)/(9 + 1) = 10/30	P(Cool|No) = (1 + 1/3)/(6 + 1) = 4/21
P(Hot|Yes) = (2 + 1/3)/(9 + 1) = 7/30	P(Hot|No) = (2 + 1/3)/(6 + 1) = 7/21
P(Mild|Yes) = (4 + 1/3)/(9 + 1) = 13/30	P(Mild|No) = (3 + 1/3)/(6 + 1) = 10/21
Humidity p = 1/2	
P(High|Yes) = (4 + 1/2)/(9 + 1) = 9/20	P(High|No) = (4 + 1/2)/(6 + 1) = 9/14
P(Normal|Yes) = (5 + 1/2)/(9 + 1) = 11/20	P(Normal|No) = (2 + 1/2)/(6 + 1) = 5/14
Wind p = 1/2	
P(Strong|Yes) = (4 + 1/2)/(9 + 1) = 9/20	P(Strong|No) = (3 + 1/2)/(6 + 1) = 7/14
P(Weak|Yes) = (5 + 1/2)/(9 + 1) = 11/20	P(Weak|No) = (3 + 1/2)/(6 + 1) = 7/14

Please use your Naïve Bayes classifier to determine whether a person should play tennis or not, under conditions that “Outlook=Overcast & Temperature=Hot & Humidity =Normal& Wind=Weak”. Why or why not? [1 pt]

$$P(Yes)P(Overcast|Yes)P(Hot|Yes)P(Normal|Yes)P(Weak|Yes) = \frac{9}{5} \times \frac{10}{30} \times \frac{7}{30} \times \frac{11}{20} \times \frac{11}{20} = 0.014$$

$$P(No)P(Overcast|No)P(Hot|No)P(Normal|No)P(Weak|No) = \frac{6}{15} \times \frac{4}{21} \times \frac{7}{21} \times \frac{5}{14} \times \frac{7}{14} = 0.005$$

The person should play tennis, given that the conditional probability for Class = Yes > conditional probability for Class = No.

Question 6 [2 pts]

Please download mtcars.header.binary.categorical.txt dataset from Canvas (the “cyl” and “gear” are changed as categorical attributes, instead of numerical numbers). Please use R to implement tasks below:

Please use all instances from mtcars.header.binary.categorical.txt to train a Naïve Bayes classifier, and use the classifier to predict all instances in mtcars.header.binary.categorical.txt, and report the classification accuracy [1 pt]

Please refer to file cap5615-homework2-question6.R for the source code

The accuracy is

$$\frac{16+13}{16+1+2+13} = 0.906$$

The confusion matrix for the predictions:

```
p    0  1
  0 16  1
  1  2 13
```

Please report the conditional probabilities for attributes “cyl” and “gear”, and explain the meaning of the conditional probability values [1 pt]

```
> cars_nb$tables$cyl
   cyl
Y       eight      four       six
  0 0.6111111 0.1666667 0.2222222
  1 0.0000000 0.7857143 0.2142857

> cars_nb$tables$gear
   gear
Y        five      four     three
  0 0.1666667 0.1111111 0.7222222
  1 0.1428571 0.7142857 0.1428571
```

The rows are the possible values for the class we want to predict. The columns are all the values for the attribute under consideration. Each cell is the prior probability of the attribute value, given the class. For example, P(cyl=four|1) = 0.7857143, i.e. ~78% of the four-cylinder cars are high-mileage cars (in the training data).

Question 7 [3 pts]

Please download housing.header.binary.txt dataset from Canvas, and use R to implement tasks below (a brief description of this dataset is available from the following URL) https://archive.ics.uci.edu/ml/datasets/housing [The Medv attribute in housing.header.binary.txt is binarized with Medv value of the house greater than 200k being 1, or 0 otherwise.]

Please use 80% of instances in the “housing.header.binary.txt” dataset to build a Naïve Bayes Classifier. Report the performance of the NB classifier on the remaining 20% of instances in the “housing.header.binary.txt”

Report source code to build the NB classifier and answer the following tasks [0.5 pt]

Please see attached file cap5615-homework2-question7.R

Report confusion table, TPR, FPR, and the Accuracy [0.5 pt]

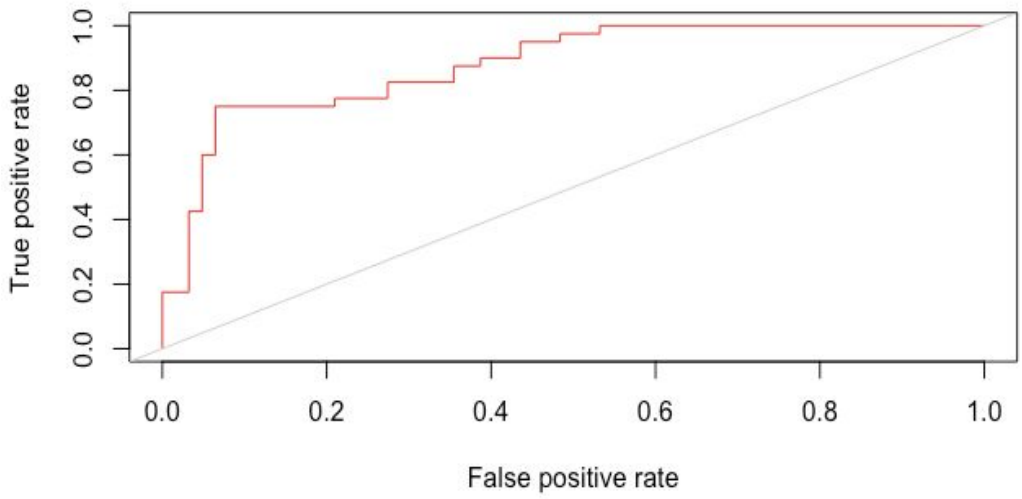
```
> table(p, housing_test$Medv)

p    0  1
  0 40  7
  1 22 33
```

$$Accuracy = \frac{40+33}{40+7+22+33} = 0.716$$

$$TPR = \frac{33}{33+22} = 0.6 \quad FPR = \frac{7}{7+40} = 0.149$$

Report the ROC curve [0.5 pt]



Report the AUC value [0.5 pt]

```
> auc@y.values[[1]]
[1] 0.8802419
```

Create a new instance with “Crim=0.03, Zn=13, Indus=3.5, Chas=0.3, Nox=0.58, Rm=4.1, Age=68, Dis=4.98, Rad =3, Tax=225, Ptratio=17, B=396, Lstat=7.56”, and predict the Medv value of the instance. Report the posterior probability, and the classification result [1.0 pt]

```
> predict(housing_nb, newdata = new_inst, type = "raw") # probability
                0         1
[1,] 0.001128189 0.9988718

> predict(housing_nb, newdata = new_inst, type = "class") # classification
[1] 1
```