739-Assignment 2

Xu Sun | xs6925

Question 1 - PAC learning analysis

Consider the Boolean data set posted in MyCourses named q1.csv. We may choose to model this as an arbitrary Boolean function f(x1..xn)=y. If we do so, we can use PAC learning to predict the expected accuracy of our model. Answer the following questions in your write up. Show your work.

a) How many different hypotheses are there (give an exact number or numeric expression)?
Answer: There are 2 classes and 6 variables, so totally there are 2^{2^6} different hypotheses.

b) With what probability can we guarantee a classifier that is 90% accurate (i.e with no more than 10% classification error)? How about one that is 80% accurate?
Answer:
90% accurate:
H = 2^{2^6} = 18446744073709551616 , $\epsilon = 1 - 0.9 = 0.1$, N = 200
$\sigma = \mathrm{H} * (1 - \epsilon)^N \approx 13014323873$
$1 - \sigma \leq 0$
So there is no possible to guarantee the classifier has a 90% accurate.

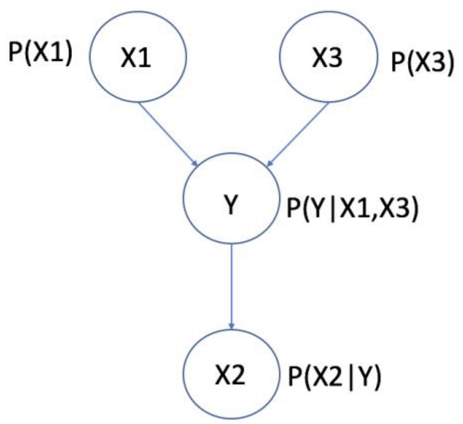
80% accurate:
H = 2^{2^6} = 18446744073709551616 , $\epsilon = 1 - 0.8 = 0.2$, N = 200
$\sigma = \mathrm{H} * (1 - \epsilon)^N \approx 0.765$
$1 - \sigma = 0.235$
So there is 23.5% possible to guarantee the classifier has a 80% accurate.

c) How many additional samples would we want to get a classifier with 90% accuracy, 80% of the time?
Answer:
$1 - \sigma = 0.8$, $\sigma = 0.2$, $\epsilon = 1 - 0.9 = 0.1$
$\mathrm{M} \geq \frac{1}{\epsilon}(\ln|H| + ln\frac{1}{\sigma}) \approx 456$
456 – 200 = 256
So we should have additional 256 samples.

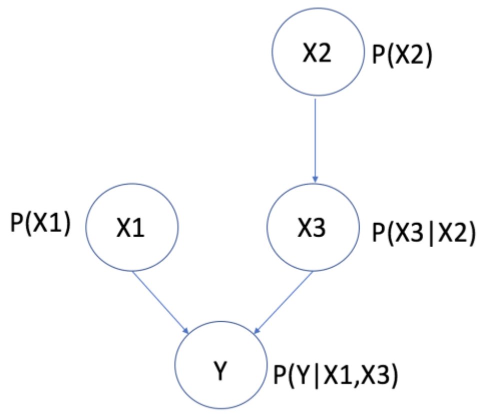
d)
Answer:
H = 3^6 = 729
$1 - \sigma = 0.8$, $\sigma = 0.2$, $\epsilon = 1 - 0.9 = 0.1$
$\mathrm{M} \geq \frac{1}{\epsilon}(\ln|H| + ln\frac{1}{\sigma}) \approx 81.9$
So we should have at least 82 samples.

Question 2 - Bayes nets
Consider a problem with four Boolean variables (one of which is the class we are trying to predict). Here are two different sets of assumptions about how the variables are related:

For each set of assumptions:
a) Draw the corresponding Bayes network, including the symbolic parameters that are required for a full representation of the problem.
Answer:



model 1



model 2

b) Write an exact symbolic expression for P(Y=True | x1=True ^ x2=False ^ x3 = True), using only the parameters of the model.
Answer:
Model 1:
P(Y=T | X1=T, X2=F, X3=T)

$$= \frac{P(Y=T,\ X1=T,\ X2=F,\ X3=T)}{P(X1=T,\ X2=F,\ X3=T)}$$

$$= \frac{P(X1=T)P(X3=T)P(Y=T|X1=T,\ X3=T)P(X2=F|Y=T)}{\sum_{Y=(T,F)} P(X1=T)P(X3=T)P(Y|X1=T,\ X3=T)P(X2=F|Y)}$$

Model 2:
P(Y=T | X1=T, X2=F, X3=T)

$$= \frac{P(Y=T,\ X1=T,\ X2=F,\ X3=T)}{P(X1=T,\ X2=F,\ X3=T)}$$

$$= \frac{P(X1=T)P(X2=F)P(X3=T|X2=F)P(Y=T|X1=T,\ X3=T)}{\sum_{Y=(T,F)} P(X1=T)P(X2=F)P(X3=T|X2=F)P(Y=T|X1=T,\ X3=T)}$$

c) Based on these data points, compute the maximum-likelihood values for the parameters needed by each network.
Answer:

Model 1:
P(X1=T) = 0.315 P(X1=F) = 0.685
P(X3=T) = 0.595 P(X3=F) = 0.405
P(Y=T| X1=T,X3=T) = 0.079 P(Y=F| X1=T,X3=T) = 0.921
P(Y=T| X1=T,X3=F) = 0.440 P(Y=F| X1=T,X3=F) = 0.560
P(Y=T| X1=F,X3=T) = 0.889 P(Y=F| X1=F,X3=T) = 0.111
P(Y=T| X1=F,X3=F) = 0.142 P(Y=F| X1=F,X3=F) = 0.857

P(X2=T| Y=T) = 0.755 P(X2=F| Y=T) = 0.245
P(X2=T| Y=F) = 0.207 P(X2=T| Y=F) = 0.792

Model 2:
P(X1=T) = 0.315 P(X1=F) = 0.685
P(X2=T) = 0.465 P(X2=F) = 0.535

P(X3=T| X2=T) = 0.677 P(X3=F| X2=T) = 0.323
P(X3=T| X2=F) = 0.523 P(X3=T| X2=F) = 0.477

P(Y=T| X1=T,X3=T) = 0.079 P(Y=F| X1=T,X3=T) = 0.921
P(Y=T| X1=T,X3=F) = 0.440 P(Y=F| X1=T,X3=F) = 0.560
P(Y=T| X1=F,X3=T) = 0.889 P(Y=F| X1=F,X3=T) = 0.111
P(Y=T| X1=F,X3=F) = 0.142 P(Y=F| X1=F,X3=F) = 0.857

d) For each network, two variables are specified as independent (i.e. with no incoming edges). Is this assumption supported by the data in each case?
Answer:
Model 1:
P(X1=T) = 0.315 P(X3=T) = 0.595
P(X1=T, X3=T) = 38/200 = 0.19 = P(X1=T)* P(X3=T)
So in model 1, X1 and X3 are independent.

Model 2:
P(X1=T) = 0.315 P(X2=T) = 0.465
P(X1=T, X2=T) = 46/200 = 0.23 ≠ P(X1=T)*P(X2=T)
So in model 1, X1 and X3 are not independent.

Question 3 - Learning and inference in Naive Bayes.
The data set given in q3.csv is synthetically generated, but labeled in an intentional way. Note that it includes both discrete and real-valued parameters. Use the Naive Bayes model for the data, with Spam being the parent class, and also assume that each real-valued parameter has a Gaussian distribution.
a) Using the data in q3.csv, compute the maximum-likelihood parameters for this network.

```
P(in html=False| is spam=False) = 0.41304347826086957
P(in html=True| is spam=False) = 0.5869565217391304
P( has emoji=False| is spam=False) = 0.8526570048309179
P( has emoji=True| is spam=False) = 0.1473429951690821
P( sent to list=False| is spam=False) = 0.6884057971014492
P( sent to list=True| is spam=False) = 0.3115942028985508
P( from .com=False| is spam=False) = 0.7246376811594203
P( from .com=True| is spam=False) = 0.2753623188405797
P( has my name=False| is spam=False) = 0.39855072463768115
P( has my name=True| is spam=False) = 0.6014492753623188
P( has sig=False| is spam=False) = 0.6763285024154589
P( has sig=True| is spam=False) = 0.32367149758454106
P( # sentences) = mean:6.190821256038648, var:6.400785315876683
P( # words) = mean:70.77053140096618, var:912.7661847417676
```

```
P(in html=False| is spam=True) = 0.2441860465116279
P(in html=True| is spam=True) = 0.7558139534883721
P( has emoji=False| is spam=True) = 0.8023255813953488
P( has emoji=True| is spam=True) = 0.19767441860465118
P( sent to list=False| is spam=True) = 0.9302325581395349
P( sent to list=True| is spam=True) = 0.06976744186046513
P( from .com=False| is spam=True) = 0.2558139534883721
P( from .com=True| is spam=True) = 0.7441860465116279
P( has my name=False| is spam=True) = 0.6511627906976745
P( has my name=True| is spam=True) = 0.34883720930232553
P( has sig=False| is spam=True) = 0.3372093023255814
P( has sig=True| is spam=True) = 0.6627906976744187
P( # sentences) = mean:3.9767441860465116, var:3.7203893996754998
P( # words) = mean:68.83720930232558, var:79.34559221200648
```

b) For each row of the file q3b.csv, use the values of the features other than Spam to compute P(Y | feature values) for each row. Compute an overall classification error rate based on a threshold P(Y) of 0.5.

```
classification error rate:15.5%
```

c) Playtime! Repeat the preceding analysis, but ignoring some or all of the features (columns), and compare the classification accuracy. Choose a subset of the features that is small but seems to give good results.

```
1 parameters ['in html']
classification error rate with 1 parameters: 20.0%

2 parameters ['in html', ' has emoji']
classification error rate with 2 parameters: 56.00000000000001%

3 parameters ['in html', ' has emoji', ' sent to list']
classification error rate with 3 parameters: 56.00000000000001%

4 parameters ['in html', ' has emoji', ' sent to list', ' from .com']
classification error rate with 4 parameters: 46.5%

5 parameters ['in html', ' has emoji', ' sent to list', ' from .com', ' has my name']
classification error rate with 5 parameters: 27.500000000000004%

6 parameters ['in html', ' has emoji', ' sent to list', ' from .com', ' has my name', ' has sig']
classification error rate with 6 parameters: 32.0%

7 parameters ['in html', ' has emoji', ' sent to list', ' from .com', ' has my name', ' has sig', ' # sentences']
classification error rate with 7 parameters: 27.500000000000004%
```

I just try each possible from 1 to 7 parameter from front to end and calculate each error rate. Then when I have 1 parameter, I get the error rate is 20%, or when 5 parameters , the error rate is 27.5%.