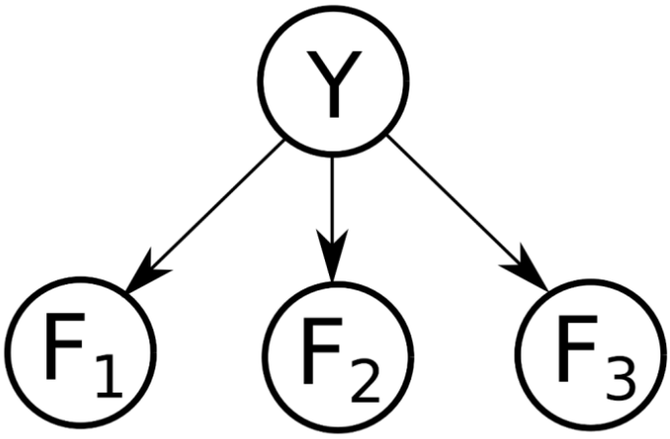
In this question, we will train a Naive Bayes classifier to predict class labels Y as a function of input features F_i . (**Keep 3 decimal places**)



We are given the following 15 training points:

F_1	0	1	1	1	0	1	1	1	1	1	0	1	1	1
F_2	0	0	1	0	1	0	1	1	1	1	1	1	0	1
F_3	1	0	1	1	1	1	1	1	1	1	1	0	1	0
Y	A	A	B	B	B	B	B	B	B	B	B	C	C	C	C

What is the maximum likelihood estimate of the prior P(Y)?

Y	P(Y)
A	
B	
C	

$\frac{2}{15} = 0.133$

$\frac{9}{15} = 0.600$

$\frac{4}{15} = 0.267$

What are the maximum likelihood estimates of the conditional probability distributions? Fill in the tables below (the second and third are done for you).

F_1	Y	$P(F_1 \mid Y)$
0	A	
1	A	
0	B	
1	B	
0	C	
1	C	

$\frac{1}{9} = 0.111$

F_2	Y	$P(F_2 \mid Y)$
0	A	1.000
1	A	0.000
0	B	0.222
1	B	0.778
0	C	0.250
1	C	0.750

$\frac{2}{9}$

F_3	Y	$P(F_3 \mid Y)$
0	A	0.500
1	A	0.500
0	B	0.000
1	B	1.000
0	C	0.500
1	C	0.500

Bayes:

$P(Y, F_1 \cdots F_n) = P(Y) \prod P(F_i \mid Y)$

Following question 1, Now consider a new data point ($F_1 = 0, F_2 = 0, F_3 = 1$). Use your classifier to determine the joint probability of causes Y and this new data point, along with the posterior probability of Y given the new data: (**Keep 3 decimal places**)

Y	$P(Y, F_1 = 0, F_2 = 0, F_3 = 1)$
A	
B	
C	

$P(A) \cdot P(F_1|A) \cdot P(F_2|A) \cdot P(F_3|A)$

normalize

Y	$P(Y \mid F_1 = 0, F_2 = 0, F_3 = 1)$
A	0.587
B	0.267
C	0.143

What label does your classifier give to the new data point? (Break ties alphabetically). Enter capital letters only

The training data is repeated here for your convenience:

F_1	0	1	1	1	0	1	1	1	1	1	1	0	1	1	1
F_2	0	0	1	0	1	0	1	1	1	1	1	1	1	0	1
F_3	1	0	1	1	1	1	1	1	1	1	1	1	0	1	0
Y	A	A	B	B	B	B	B	B	B	B	B	C	C	C	C

Following the previous questions, now use Laplace Smoothing with strength k = 3 to estimate the prior P(Y) for the same data. (**Keep 3 decimal places**)

Y	P(Y)
A	
B	0.500
C	

$\frac{9+3}{15+3\times3}$

$\frac{4+3}{15+3\times3}$

$\frac{2}{15} \rightsquigarrow \frac{2+3}{15+3\times3}$

Use Laplace Smoothing with strength k = 3 to estimate the conditional probability distributions below (again, the second two are done for you).

F_1	Y	$P(F_1 \mid Y)$
0	A	
1	A	
0	B	
1	B	
0	C	
1	C	

F_2	Y	$P(F_2 \mid Y)$
0	A	0.625
1	A	0.375
0	B	0.333
1	B	0.667
0	C	0.400
1	C	0.600

F_3	Y	$P(F_3 \mid Y)$
0	A	0.500
1	A	0.500
0	B	0.200
1	B	0.800
0	C	0.500
1	C	0.500

Now consider again the new data point $F_1 = 0, F_2 = 0, F_3 = 1$. Use the Laplace-Smoothed version of your classifier to determine the joint probability of causes Y and this new data point, along with the posterior probability of Y given the new data: (**Keep 3 decimal places**)

Y	$P(Y, F_1 = 0, F_2 = 0, F_3 = 1)$
A	
B	
C	

Y	$P(Y \mid F_1 = 0, F_2 = 0, F_3 = 1)$
A	
B	
C	

What label does your (Laplace-Smoothed) classifier give to the new data point? (Break ties alphabetically). Enter a single capital letter.

When training a classifier, it is common to split the available data into a training set, a hold-out set, and a test set, each of which has a different role.

Which data set is used to learn the conditional probabilities?

- (●) Training Data
- ( ) Hold-Out Data
- ( ) Test Data

问题 6

Which data set is used to tune the Laplace Smoothing hyperparameters?

- ( ) Training Data
- (●) Hold-Out Data
- ( ) Test Data

问题 7

Which data set is used to apply early stopping when training a neural net?

- ( ) Training data
- (●) Hold-Out data
- ( ) Test data

Which data set is used to apply early stopping when training a neural net?

- ( ) Training data
- (●) Hold-Out data
- ( ) Test data

问题 8

The K-means algorithm:

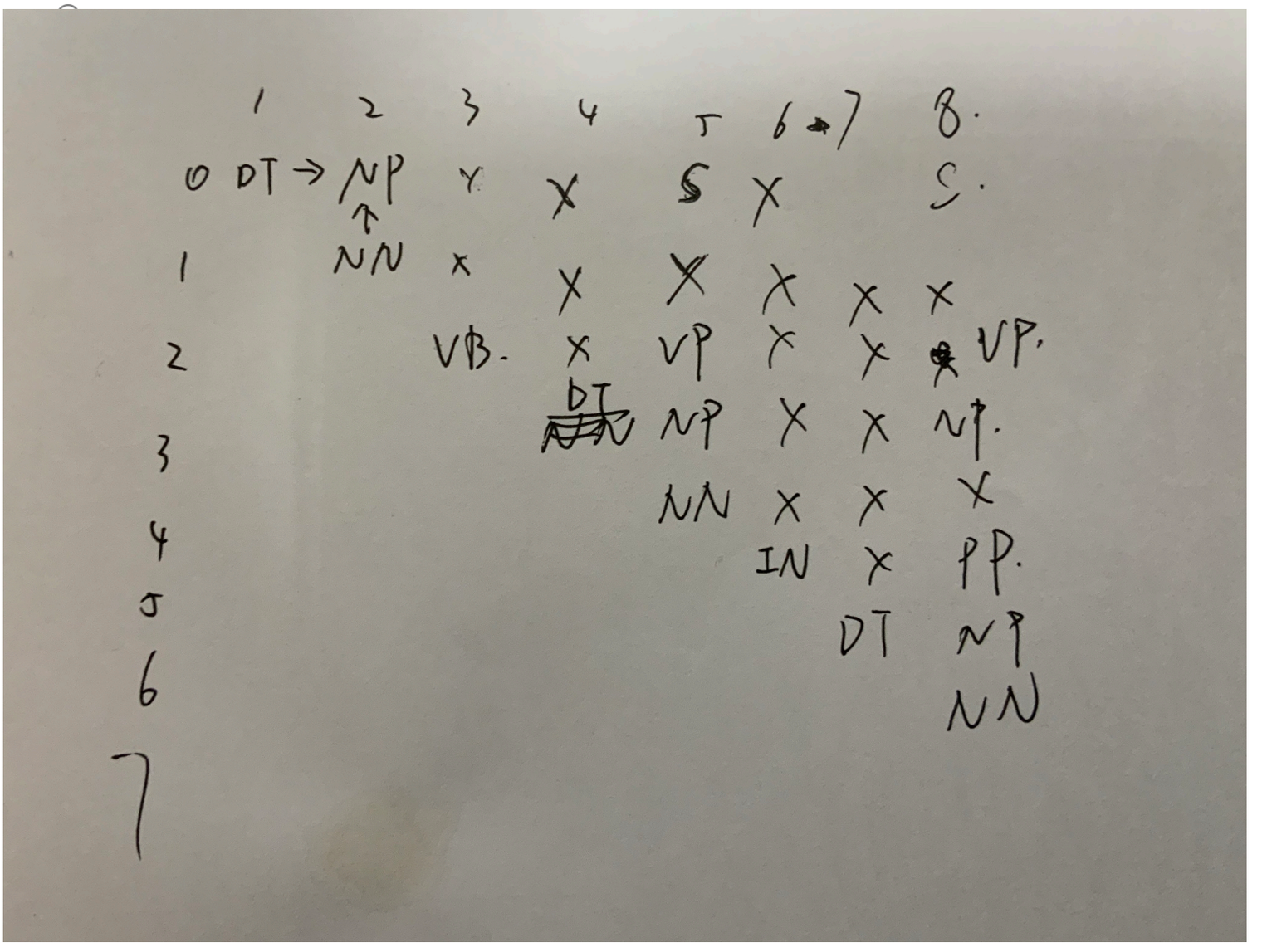
- [ ] Requires the dimension of the feature space to be no bigger than the number of samples
- [ ] Has the smallest value of the objective function when K = 1
- [ ] Converges to the global optimum if and only if the initial means are chosen as some of the samples themselves
- [x] None of the above

Consider a context-free grammar with the following rules (assume that S is the start symbol):

S → NP VP
NP → DT NN
NP → NP PP
PP → IN NP
VP → VB NP
DT → the
NN → man
NN → dog
NN → cat
NN → park
VB → saw
IN → in
IN → with
IN → under

How many parse trees are there under this grammar for the sentence: **the man saw the dog in the park?**

- ○ 4
- ○ 3
- ◉ 2



"Less important parameters go close to zero when we increase the value of tuning parameters" in which of the following regressions?

- ○ LASSO
- ◉ Ridge
- ○ Both of above
- ○ None of above

问题 11

What is overfitting?

- ○ When a predictive model is accurate but takes too long to run
- ◉ When the model learns specifics of the training data that can't be generalized to a larger data set
- ○ When you apply a powerful deep learning algorithm to a simple machine learning problem
- ○ When you perform hyperparameter tuning and performance degrades

问题 12

You are presented with a dataset that has hidden/missing variables that influences your data. You are asked to use Expectation Maximizatio algorithm to best capture the data. How would you define the E and M in Expectation Maximization?

- ◉ Estimate the Missing/Latent Variables in the Dataset, Maximize the likelihood over the parameters in the model
- ○ Estimate the number of Missing/Latent Variables in the Dataset, Maximize the likelihood over the parameters in the model
- ○ Estimate likelihood over the parameters in the model, Maximize the number of Missing/Latent Variables in the Dataset
- ○ Estimate the likelihood over the parameters in the model, Maximize the number of parameters in the model