

# VE 492 Homework9

Due: 23:59, July. 22<sup>th</sup>

## Question 1: Maximum Likelihood Estimation

We will begin with a short derivation. Consider a probability distribution with a domain that consists of  $|X|$  different values. We get to observe  $N$  total samples from this distribution. We use  $n_i$  to represent the number of the  $N$  samples for which outcome  $i$  occurs. Our goal is to estimate the probabilities  $\theta_i, i = 1, 2, \dots, |X| - 1$  of each of the events. The probability of the last outcome,  $|X|$ , equals  $1 - \sum_{i=1}^{|X|-1} \theta_i$ .

In *maximum likelihood estimation*, we choose the  $\theta_i$  that maximize the likelihood of the observed samples,

$$L(\text{samples}, \theta) \propto (1 - \theta_1 - \theta_2 - \dots - \theta_{|X|-1})^{n_{|X|}} \prod_{i=1}^{|X|-1} \theta_i^{n_i}$$

For this derivation, it is easiest to work with the log of the likelihood. Maximizing log-likelihood also maximizes likelihood, since the quantities are related by a monotonic transformation. Taking logs we obtain

$$\theta^{\text{ML}} = \underset{\theta}{\operatorname{argmax}} n_{|X|} \log(1 - \theta_1 - \theta_2 - \dots - \theta_{|X|-1}) + \sum_{i=1}^{|X|-1} n_i \log \theta_i$$

Setting derivatives with respect to  $\theta_i$  equal to zero, we obtain  $|X| - 1$  equations in the  $|X| - 1$  unknowns,  $\theta_1, \theta_2, \dots, \theta_{|X|-1}$ :

$$\frac{-n_{|X|}}{1 - \theta_1^{\text{ML}} - \theta_2^{\text{ML}} - \dots - \theta_{|X|-1}^{\text{ML}}} + \frac{n_i}{\theta_i^{\text{ML}}} = 0$$

Multiplying by  $\theta_i(1 - \theta_1 - \theta_2 - \dots - \theta_{|X|-1})$  makes the original  $|X| - 1$  nonlinear equations into  $|X| - 1$  linear equations:

$$-n_{|X|} \theta_i^{\text{ML}} + n_i (1 - \theta_1^{\text{ML}} - \theta_2^{\text{ML}} - \dots - \theta_{|X|-1}^{\text{ML}}) = 0$$

That is, the maximum likelihood estimation of  $\theta$  can be found by solving a linear system of  $|X| - 1$  equations in  $|X| - 1$  unknowns. Doing so shows that the maximum likelihood estimate corresponds to simply the count for each outcome divided by the total number of samples. I.e., we have that:

$$\theta_i^{\text{ML}} = \frac{n_i}{N}$$

**Notice:** Please write each sub-question in one row, that is, there will be 3 rows for this question. And please use irreducible fractions for your answer.

**Sample Answer:**

*1, 1/2, 1/3, 1/4*

*2/5 (instead of 4/10), 1/3, 4/7*

*3/8, 3/7, 3/5*

### Part 1.

Now, consider a sampling process with 3 possible outcomes: R, G, and B. We observe the following sample counts:

outcome	R	G	B
count	3	1	7

- 1) What is the total sample count  $N$ ?
- 2) What are the maximum likelihood estimates for the probabilities of each outcome?

$$\theta_R^{ML} =$$

$$\theta_G^{ML} =$$

$$\theta_B^{ML} =$$

### Part 2.

Now, use *Laplace smoothing* with strength  $k = 3$  to estimate the probabilities of each outcome.

$$\theta_R^{LAP,3} =$$

$$\theta_G^{LAP,3} =$$

$$\theta_B^{LAP,3} =$$

### Part 3.

Now, consider Laplace smoothing in the limit  $k \rightarrow \infty$ . Fill in the corresponding probability estimates.

$$\theta_R^{LAP,\infty} =$$

$$\theta_G^{LAP,\infty} =$$

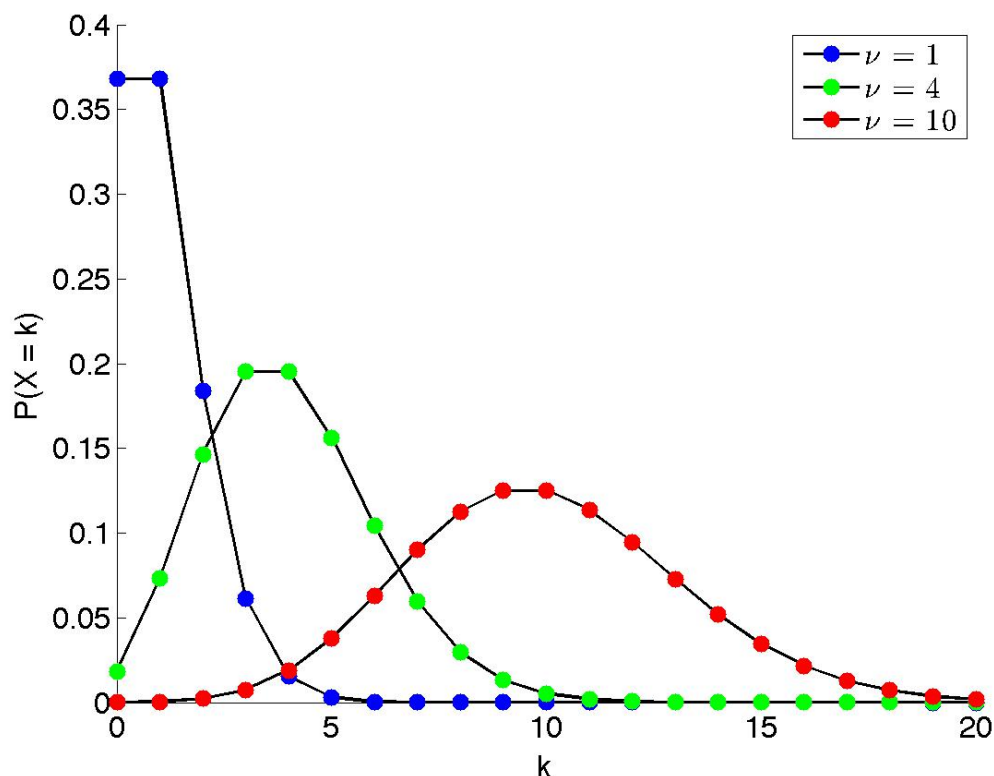
$$\theta_B^{LAP,\infty} =$$

## Question 2: Poisson Parameter Evaluation

We will now consider maximum likelihood estimation in the context of a different probability distribution. Under the Poisson distribution, the probability of an event occurring  $X = k$  times is:

$$P(X = k) = \frac{v^k e^{-v}}{k!}$$

Here  $v$  is the parameter we wish to estimate. The distribution is plotted for several values of  $v$  below.



On a sheet of scratch paper, work out the maximum likelihood estimate for  $v$ , given observations of several  $k_i$ .

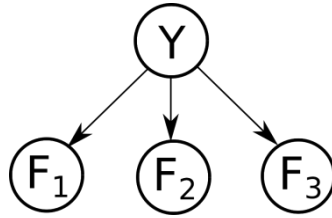
**Hints:** start by taking the product of the equation above over all the  $k_i$ , and then taking the log. Then, differentiate with respect to  $v$ , set the result equal to 0, and solve for  $v$  in terms of the  $k_i$ .

You observe the samples  $k_1 = 6, k_2 = 3, k_3 = 8, k_4 = 4, k_5 = 2$ . What is your maximum likelihood estimate of  $v$ ?

**Sample Answer (rounded to 3 decimal places):**  
**0.160**

## Question 3: Naive Bayes

In this question, we will train a Naive Bayes classifier to predict class labels  $Y$  as a function of input features  $F_i$ .



We are given the following 15 training points:

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

**Note:** Please write your answer for each table in one row, that is, there will be 10 rows for this question. Besides, please use values rounded to 3 decimal places.

**Sample Answer:**

**0.160,0.170,0.160**

...

...

...

**A**

...

**0.200,0.211,...**

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

$Y$	$P(Y)$
A	
B	
C	

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

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

$F_2$	$Y$	$P(F_2 Y)$
0	A	0.800
1	A	0.200
0	B	1.000
1	B	0.000
0	C	0.000
1	C	1.000

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

- 3) 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:

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

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

- 4) What label does your classifier give to the new data point? (Break ties alphabetically). Write capital letters only.
- 5) Now use Laplace Smoothing with strength  $k = 3$  to estimate the prior  $P(Y)$  for the same data.

$Y$	$P(Y)$
A	
B	
C	

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

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

$F_2$	$Y$	$P(F_2 Y)$
0	A	0.688
1	A	0.312
0	B	0.700
1	B	0.300
0	C	0.429
1	C	0.571

$F_3$	$Y$	$P(F_3 Y)$
0	A	0.812
1	A	0.188
0	B	0.500
1	B	0.500
0	C	0.571
1	C	0.429

- 7) 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:

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

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

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

## Question 4: Datasets

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.

*Sample Answer:*

A

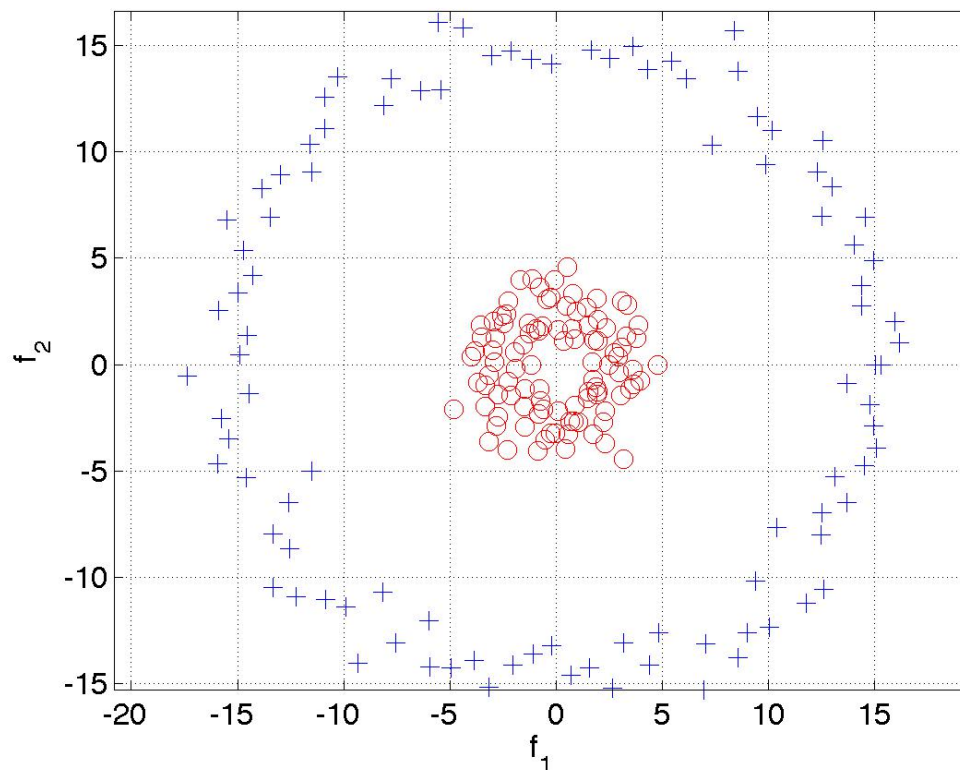
A

A

- 1) Which data set is used to learn the conditional probabilities?
  - A. Training Data
  - B. Hold-Out Data
  - C. Test Data
- 2) Which data set is used to tune the Laplace Smoothing hyperparameters?
  - A. Training Data
  - B. Hold-Out Data
  - C. Test Data
- 3) Which data set is used for quantifying performance results?
  - A. Training Data
  - B. Hold-Out Data
  - C. Test Data

## Question 5: Linear Separability

Consider the data in the figure below.



The data is plotted as a function of two features,  $f_1$  and  $f_2$ . As plotted, the data is not linearly separable. Which of the following candidate features  $f_3$ , when added, would cause the data to be linearly separable? Choose all possible answer(s).

- ☒ A.  $f_3 = |f_1| + |f_2|$
- ☐ B.  $f_3 = \sin(f_1)$
- ☒ C.  $f_3 = f_1^2 + f_2^2$
- ☐ D.  $f_3 = f_1^2$
- ☐ E.  $f_3 = f_1$
- ☐ F.  $f_3 = 1$
- ☐ G.  $f_3 = f_1 f_2$
- ☒ H.  $f_3 = 1$  if  $f_1 \in [-7, 7]$  and  $f_2 \in [-7, 7]$ , 0 otherwise