

## 1 Probability Review [15pts]

A group of travellers find themselves lost in a cave. They come upon 3 tunnels  $A$ ,  $B$ ,  $C$ . Both tunnels  $A$  and  $B$  are closed loops that do not lead to an exit and in fact lead right back to the entrance of the 3 tunnels. Tunnel  $C$  is the tunnel which leads to the exit. If they go through tunnel  $A$ , then it takes 2 days to go through the tunnel. If they go through tunnel  $B$ , then it takes 1 day to go through the tunnel. If they go through tunnel  $C$ , then they immediately leave the cave. Suppose the travellers choose tunnels  $A$ ,  $B$  and  $C$  with constant probability 0.6, 0.3, 0.1 every time. (For the following questions please round your answer up to 4 digits.)

1. [6 pts] Suppose we record down the travellers choices into a sequence (e.g.,  $ABBA\dots C$ ). What is the probability that the pattern  $AAB$  appears in the sequence before any  $BAA$  appears?

Note: You should also count cases where  $AAB$  appears in the sequence and  $BAA$  does not.

Final Answer

0.27

Work for Final Answer

The pattern  $AAB$  appears before  $BAA = \text{no } B \text{ before } AAB \text{ appears}$

$\Rightarrow$  The overall pattern should be  $AAA\dots\dots AAB$

$$P(AAB) + P(AAAB) + P(AAAAB) + \dots$$

$$= P(AAB) + P(A)P(AAB) + P(A)P(A)P(AAB) + \dots$$

$$= P(AAB) (1 + P(A) + [P(A)]^2 + \dots + [P(A)]^\infty)$$

$$= P(A)P(A)P(B) [1 + P(A) + P(A)^2 + \dots + P(A)^\infty]$$

$$= 0.6 \times 0.6 \times 0.3 \times (1 + 0.6 + 0.6^2 + \dots + 0.6^\infty) = 0.108 \times \frac{1}{1-0.6} = \frac{0.108}{0.4} = 0.27$$

2. [3 pts] What is the expected number of days that the travellers will be lost in the cave?

Final Answer

15

Work for Final Answer

The expected days wasted for one wrong choice is

$$P(\text{Not } C) \times \left[ \left( \frac{P(A)}{1-P(C)} \times 2 \text{ days} \right) + \left( \frac{P(B)}{1-P(C)} \times 1 \text{ day} \right) \right] = 0.9 \times \left[ \frac{0.6}{0.9} \times 2 + \frac{0.3}{0.9} \times 1 \right] = 0.9 \times \frac{5}{3}$$

n Wrong choices and 1 right choice in the end

$$= \left( \sum_{n=1}^{\infty} n \times \frac{5}{3} \times (0.9)^n \right) \times 0.1 = 0.1 \times \frac{5}{3} \times \sum_{n=1}^{\infty} n (0.9)^n$$

$$= 0.1 \times \frac{5}{3} \times \frac{0.9}{(1-0.9)^2} = \frac{5}{3} \times 9 = \frac{45}{3} = 15 *$$

3. [6 pts] What is the variance of days that the travellers will be lost in the cave? (Hint: To compute  $\text{Var}(T)$  for a random variable  $T$ , you can either compute  $E[T^2]$  first and then  $\text{Var}(T)$  or directly compute the variance using the law of total variance.)

Final Answer

250

Work for Final Answer

$$\text{Var}(T) = E[T^2] - E[T]^2$$

$$E[T]^2 = 15^2 = 225$$

$$E[T^2] = \sum P(T)T^2$$

$$= 0.1 \times \left( \frac{5}{3} \right)^2 \times \sum_{n=1}^{\infty} n^2 (0.9)^n = 0.1 \times \frac{25}{9} \times 0.9 \times \frac{1+0.9}{(1-0.9)^2} = 475$$

$$\Rightarrow \text{Var}(T) = 475 - 225 = 250 *$$

## 2 MLE and MAP [25pts]

Please note that for this section (MLE and MAP) you do not need to show proof that the optimum point is a maximum.

### 2.1 MLE with Exponential Family

1. [7pts] The exponential family of distributions has the form  $P(x | \theta^*) = h(x) \exp[\theta^* \cdot \phi(x) - A(\theta^*)]$ . It may look unfamiliar, but in fact many well-known distributions including Gaussian, Bernoulli, Geometric and Laplace distributions belong to this family<sup>2</sup>. Suppose we are given  $n$  i.i.d samples  $X = \{x^1, x^2, \dots, x^n\}$  drawn from the distribution  $P(x | \theta^*)$ . Derive the Maximum Likelihood Estimator  $\hat{\theta}_{MLE}$  for this true parameter  $\theta^*$ .

Here you can assume that  $A$  is convex and differentiable, and that the derivative  $A'$  is invertible. Your answer should be in terms of  $A$ ,  $\phi(x)$  and  $h(x)$ .

Final Answer

$$\frac{1}{n} (A')^{-1} \sum_{x=1}^n \phi(x)$$

Work for Final Answer

$$P(x | \theta^*) = h(x) \exp[\theta^* \cdot \phi(x) - A(\theta^*)]$$

$$MLE : \hat{\theta}_{MLE} = \arg \max_{\theta^*} P(x | \theta^*)$$

$$= \arg \max_{\theta^*} \prod_{x=1}^n P(x | \theta^*)$$

$$\log L(\hat{\theta}) = \arg \max_{\theta^*} \log \prod_{x=1}^n P(x | \theta^*)$$

$$= \sum_{x=1}^n \log (h(x) \cdot e^{\theta^* \cdot \phi(x) - A(\theta^*)})$$

$$= \sum_{x=1}^n [\log h(x) + \log e^{\theta^* \cdot \phi(x) - A(\theta^*)}]$$

$$= \sum_{x=1}^n [\log h(x) + \theta^* \cdot \phi(x) - A(\theta^*)]$$

$$\Rightarrow \frac{dL(\theta)}{d\theta} = \sum_{x=1}^n (\phi(x) - A'(\theta^*)) = \sum_{x=1}^n \phi(x) - n A'(\theta^*) = 0$$

$$\Rightarrow \sum_{x=1}^n \phi(x) = n A'(\theta^*)$$

$$\Rightarrow \hat{\theta}_{MLE} = \frac{(A')^{-1} \sum_{x=1}^n \phi(x)}{n}$$

<sup>2</sup>To see the parameter setting for each of these distributions, which makes them become special cases of exponential distributions you can check [https://en.wikipedia.org/wiki/Exponential\\_family#Table\\_of\\_distributions](https://en.wikipedia.org/wiki/Exponential_family#Table_of_distributions).

## 2.2 MLE and MAP with Pareto Distribution

1. [5 pts] The Pareto distribution has the form

$$f(x) = \frac{ab^a}{x^{a+1}}, \quad x \geq b$$

with the parameters  $a, b > 0$ . For our purposes, assume that  $b$  is known. We obtain  $n$  i.i.d. data points  $x^1, x^2, \dots, x^n$  from the Pareto distribution. Find the MLE estimate  $\hat{a}$ .

Final Answer

$$\frac{n}{\sum_{i=1}^n \log x_i - n \log b}$$

Work for Final Answer

$$L(a) = \prod_{i=1}^n \frac{ab^a}{x_i^{a+1}} = (ab^a)^n \times \prod_{i=1}^n x_i^{-(a+1)}$$

$$\begin{aligned} \Rightarrow \log L(a) &= n \log ab^a + \sum_{i=1}^n \log x_i^{-(a+1)} \\ &= n \log a + na \log b - (a+1) \sum_{i=1}^n \log x_i \end{aligned}$$

$$\Rightarrow \frac{d \log L(a)}{da} = n \cdot \frac{1}{a} + n \log b - \sum_{i=1}^n \log x_i = 0$$

$$\frac{n}{a} = \sum_{i=1}^n \log x_i - n \log b$$

$$a = \frac{n}{\sum_{i=1}^n \log x_i - n \log b}$$

2. [10 pts] Now suppose  $a$  has a  $\text{Gamma}(\alpha, \beta)$  prior distribution with probability density function:

$$f(a) = \frac{\beta^\alpha}{\Gamma(\alpha)} a^{\alpha-1} e^{-\beta a}$$

The parameters  $\alpha > 0, \beta > 0$  are both known. Find the posterior distribution of  $a$  given  $x$ , and find the MAP estimate  $\tilde{a}$ .

Posterior Distribution	MAP estimate $\tilde{a}$
$\prod_{i=1}^n \frac{ab^a}{x_i^{\alpha+1}} \times \frac{\beta^\alpha}{\Gamma(\alpha)} a^{\alpha-1} e^{-\beta a}$	$\frac{\alpha-1+n}{\beta - n \log b + \sum_{i=1}^n \log x_i}$

Work for Final Answer

$$\begin{aligned}
 P(a|x) &= P(x|a)P(a) \\
 &= \prod_{i=1}^n \frac{ab^a}{x_i^{\alpha+1}} \times \frac{\beta^\alpha}{\Gamma(\alpha)} a^{\alpha-1} e^{-\beta a} \\
 &= \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot e^{-\beta a} \cdot a^{\alpha-1} \cdot \prod_{i=1}^n \frac{ab^a}{x_i^{\alpha+1}} \\
 &= \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot e^{-\beta a} \cdot a^{\alpha-1} \cdot a^n \cdot \prod_{i=1}^n \frac{b^a}{x_i^{\alpha+1}} \\
 &\stackrel{\log}{\rightarrow} \alpha \log \beta - \log \Gamma(\alpha) + \log b^a + \log a^{\alpha-1+n} + \sum a \log b - (\alpha+1) \sum \log x_i \\
 &= \alpha \log \beta - \log \Gamma(\alpha) - \beta a + (\alpha-1+n) \log a + a \sum \log b - (\alpha+1) \sum \log x_i \\
 \frac{d \log P(a|x)}{da} &= -\beta + \frac{\alpha-1+n}{a} + \sum (\log b - \log x_i) = 0 \\
 \Rightarrow \frac{\alpha-1+n}{a} &= \beta - \sum (\log b - \log x_i) \\
 a &= \frac{\alpha-1+n}{\beta - \sum (\log b - \log x_i)} = \frac{\alpha-1+n}{\beta - n \log b + \sum \log x_i}
 \end{aligned}$$

3. [3 pts] Assume  $\sum_{i=1}^n \log \frac{x_i}{b} \rightarrow \infty$  as  $n \rightarrow \infty$  for the Pareto distribution. Compare the MLE ( $\hat{\alpha}$ ) and the MAP ( $\tilde{\alpha}$ ) as  $n \rightarrow \infty$  and describe your findings.

$$\sum_{i=1}^n \log \frac{x_i}{b} = \sum_{i=1}^n (\log X_i - \log b) = \sum_{i=1}^n \log X_i - n \log b$$

$$\text{MLE } (\hat{\alpha}) = \frac{n}{\sum_{i=1}^n \log X_i - n \log b} \xrightarrow{\frac{d}{d\alpha}} \frac{1}{-\log b} \quad \leftarrow \text{Same}$$

$$\text{MAP } (\tilde{\alpha}) = \frac{\alpha + n - 1}{\beta - n \log b + \sum_{i=1}^n \log X_i} \xrightarrow{\frac{d}{d\alpha}} \frac{1}{-\log b} \quad \leftarrow \text{Same}$$

The more data points we have for MAP, the less the effect of prior to the posterior.  
 $\Rightarrow$  "washed out"

### 3 K-Nearest Neighbors: Black Box [10 Points]

- [6 pts] In a KNN classification problem, assume that the distance measure is not explicitly specified to you. Instead, you are given a "black box" where you input a set of instances  $P_1, P_2, \dots, P_n$  and a new example  $Q$ , and the black box outputs the nearest neighbor of  $Q$ , say  $P_i$  and its corresponding class label  $C_i$ . Is it possible to construct a KNN classification algorithm (w.r.t the unknown distance metrics) based on this black box alone? If so, how and if not, why not?

Yes, we can input the same  $Q$  for  $k$  times, and for each time we remove the output  $P_i$  from the instances and record the label  $C_i$ . Then we will have  $C_1, C_2, \dots, C_k$  in the order of nearest to the  $k^{\text{st}}$  nearest. Finally, we can classify  $Q$  as the class that has the highest amount in  $C_1, C_2, \dots, C_k$ .

- [4 pts] If the black box returns the  $j$  nearest neighbors (and their corresponding class labels) instead of the single most nearest neighbor (assume  $j \neq k$ ), is it possible to construct a KNN classification algorithm based on the black box? If so how, and if not why not?

It is possible if  $j \leq k$ . We only remove the remainder of  $k \div j$  points for the first step, and remove  $j$  points for the rest of the steps until we have removed  $k$  points. Then we can classify  $Q$  the same way as above.

However, it is impossible if  $j > k$  because for each step, we can only know the  $j$  nearest points but don't know about the ranking, so we are unable to know which  $C_s$  are from the  $k$  nearest points.

## 4 Naive Bayes [20 Points]

Suppose we let  $X = (x_1, x_2, \dots, x_n)$  denote the features, and  $y \in \{0, 1\}$  denote the label. Note that in any generative model approach, we model the conditional label distribution  $P(y | X)$  via the conditional distribution of features given the label  $P(X | y)$ :

$$P(y | X) \propto P(X | y)P(y) \quad (1)$$

1. [2 pts] Rewrite the conditional distribution in (1) under the Naïve Bayes assumption that the features are conditionally independent given the label.

$$P(y | X) \propto \prod_{i=1}^n P(x_i | y)P(y)$$

2. Suppose that each feature  $x_i$  takes values in the set  $\{1, 2, \dots, K\}$ . Further, suppose that the label distribution is Bernoulli, and the feature distribution conditioned on the label is multinomial. Please give detailed step by step derivations for the following questions.

- (a) [2 pts] What is the total number of parameters of the model under the Naïve Bayes assumption?

$n$  features

$K$  values  $\Rightarrow P(X_i) \Rightarrow K-1$  for values

2 classes

$$\Rightarrow n(K-1) 2 + (2-1) = 2nk - 2n + 1$$

- (b) [2 pts] What is the total number of parameters of the model without the Naïve Bayes assumption?

$$(K^n - 1) \times 2 + (2-1) = 2K^n - 1$$

- (c) [2 pts] Suppose we change the set of values that  $y$  takes, so that  $y \in \{0, 1, \dots, M-1\}$ : How would your answers change in both cases (with/out Naïve Bayes assumption)?

w/ NB:

$n$  features,  $K$  values for each feature,  $M$  classes

$$\Rightarrow n(K-1) \cdot M + (M-1) = nKM - nM + M - 1$$

w/o NB:

$$(K^n - 1) \cdot M + (M-1) = MK^n - M + M - 1 = MK^n - 1$$

3. Suppose each feature is real-valued, with  $x_i \in \mathbb{R}$ , and  $P(x_i | y = c) \sim \mathcal{N}(\mu_{i,c}, 1)$  for  $i = 1, 2, \dots, n$  and  $c = 0, 1$ . Also suppose that the label distribution is Bernoulli with  $P(y = 1) = p$ . Solve the following problems under the Naïve Bayes assumption.

- (a) [6 pts] Given  $N$  observations  $\{(X^\ell, y^\ell)\}_{\ell=1}^N$ , derive the MLE estimators of  $p$  and  $\mu_{i,c}$ .

MLE estimator of $p$	MLE estimator of $\mu_{i,c}$
$\frac{\sum_{i=1}^N k_i}{N}$	$X_i$

Work for Final Answer

$y$  is Bernoulli, with  $N$  observation  $\Rightarrow$  use binomial

$$L(p) = \prod_{i=1}^N C_{k_i}^{n_i} p^{k_i} (1-p)^{n_i - k_i} \quad \begin{matrix} n: \text{observations} \\ k: \text{get } y=1 \end{matrix}$$

$$\log \rightarrow \sum_{i=1}^N \log C_{k_i}^{n_i} + \sum_{i=1}^N k_i \log p + \sum_{i=1}^N (n_i - k_i) \log (1-p)$$

$$\frac{d}{dp} \sum_{i=1}^N k_i \frac{1}{p} + \sum_{i=1}^N (n_i - k_i) \left(-\frac{1}{1-p}\right) = 0$$

$$\Rightarrow p = \frac{\sum k_i}{\sum n_i} = \frac{\sum k_i}{N}$$

$$L(\mu_{i,c}) = \prod_{i=1}^N P(X_i | y=c) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i - \mu_{i,c})^2}{2}}$$

$$\log \rightarrow \sum_{i=1}^N \left[ \log \frac{1}{\sqrt{2\pi}} + \left( -\frac{(x_i - \mu_{i,c})^2}{2} \right) \right]$$

$$\frac{d}{d\mu_{i,c}} \sum_{i=1}^N \frac{-x_i(x_i - \mu_{i,c})}{2} = 0 \Rightarrow \mu_{i,c} = \bar{x}_i$$

- (b) [6 pts] Show that the decision boundary  $\{(x_1, x_2, \dots, x_n) : P(y=0 | x_1, x_2, \dots, x_n) = P(y=1 | x_1, x_2, \dots, x_n)\}$  is linear in  $x_1, x_2, \dots, x_n$ .

$$\begin{aligned}
 P(y|X) &\propto P(X|y)P(y) \Rightarrow P(X|y=0)P(y=0) = P(X|y=1)P(y=1) \\
 \left[ \prod_{i=1}^N \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i - \mu_{i,0})^2}{2}} \right] * P &= \left[ \prod_{i=1}^N \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i - \mu_{i,1})^2}{2}} \right] * (1-P) \\
 \log \sum_{i=1}^n \left( \log \frac{1}{\sqrt{2\pi}} - \frac{(x_i - \mu_{i,0})^2}{2} \right) + \log P &= \sum_{i=1}^n \left( \log \frac{1}{\sqrt{2\pi}} - \frac{(x_i - \mu_{i,1})^2}{2} \right) + \log(1-P) \\
 \Rightarrow \sum_{i=1}^n \left( -\frac{(x_i - \mu_{i,0})^2}{2} + \frac{(x_i - \mu_{i,1})^2}{2} \right) &= \log(1-P) - \log P = \log \frac{1-P}{P} \\
 \Rightarrow \sum_{i=1}^n \frac{2x_i\mu_{i,0} - \mu_{i,0}^2 - 2x_i\mu_{i,1} + \mu_{i,1}^2}{2} &= \log \left( \frac{1-P}{P} \right) \\
 \downarrow \\
 \text{is linear in } x_i \quad (i=1, 2, \dots, n)
 \end{aligned}$$

You will use the 2 given files:

- `census.csv`: Each line is a training data sample, with attributes listed in the same order as on the website and delimited by commas. The last is called income ( $>50K$ ,  $\leq 50K$ ). There should be 32,561 training data samples.
- `adult.test.csv`: Same format as `census.csv`, but only used in evaluation of the model (i.e. testing), so you shouldn't use the label for training your NB classifier. There should be 16,281 testing data samples.

Important: Because  $P(y) \prod_i P(x_i|y)$  can get extremely small, you should use log-posterior for your computations:

$$\log \left[ P(y) \prod_i P(x_i|y) \right] = \log P(y) + \sum_i \log P(x_i|y)$$

## 5.1 Report Parameters

For questions below, report only up to 4 significant digits after the decimal points. In addition, for the questions below in this section use the data set with imputation of missing values.

1. [2 pts] Report the prior probability of each class.

$\leq 50K$	$>50K$
0.7592	0.2408

2. [8 pts] For each class  $c$  and for each attribute  $i$  in [education-num, marital-status, race, capital-gain] print & report the following:

- If the attribute is discrete, report the value of  $\alpha_{i,c,j}$  for every possible value  $j$  in the boxes provided below!
- If the attribute is continuous, report the value of  $\mu_{i,c}$  and  $\sigma_{i,c}$  in their corresponding boxes.

(The values given below for age and workclass are what is expected. You should use these values to check correctness of your programming):

Class " $> 50K$ :

- age: mean=44.2498, var=110.6358
- workclass: Private=0.6569, Self-emp-not-inc=0.0926, Self-emp-inc=0.0795, Federal-gov=0.0473, Local-gov=0.0787, State-gov=0.0450, Without-pay=0.0, Never-worked=0.0,

Class " $\leq 50K$ :

- age: mean=36.7837, var=196.5549
- workclass: Private=0.7837, Self-emp-not-inc=0.0736, Self-emp-inc=0.0200, Federal-gov=0.0238, Local-gov=0.0598, State-gov=0.0382, Without-pay=0.0006, Never-worked=0.0003,

(a) Class “> 50K”:

- education-num:

Mean	Variance
11.6117	5.6881

- marital-status:

Married-civ-spouse	Divorced
0.8535	0.0590
Never-married	Separated
0.0626	0.0084
Widowed	Married-spouse-absent
0.0108	0.0043
Married-AF-spouse	
0.0013	

- race:

White	Asian-Pac-Islander
0.9077	0.0352
Amer-Indian-Eskimo	Other
0.0046	0.0032
Black	
0.0494	

- capital-gain:

Mean	Variance
4006.1425	212268867.6732

(b) Class “<= 50K”:

- education-num:

Mean	Variance
9.5951	5.9346

- marital-status:

Married-civ-spouse	Divorced
0.3351	0.1610
Never-married	Separated
0.4123	0.0388
Widowed	Married-spouse-absent
0.0367	0.0155
Married-AF-spouse	
0.0005	

- race:

White	Asian-Pac-Islander
0.8373	0.0309
Amer-Indian-Eskimo	Other
0.0111	0.0100

Black

0.1107

- capital-gain:

Mean

148.7525

Variance

927599.7996

3. [4 pts] Report the log-posterior values (i.e.  $\log[P(X|y)P(y)]$ ) for the first 10 test data (in the same order as the data), each rounding to 4 decimal places (have 4 numbers after decimal points, for example, 12.3456). Make sure for each of the 10 test data to report the log-posterior values for both  $\leq 50K$  and  $> 50K$ .

	$\leq 50K$	$> 50K$
1	-48.2512	-63.3939
2	-44.6070	-48.3967
3	-51.0270	-52.8532
4	-76.2501	-49.8035
5	-45.0767	-58.6450
6	-45.8555	-58.9968
7	-45.2053	-55.6472
8	-58.4537	-51.2869
9	-45.7080	-58.5134
10	-51.6230	-59.0251

## 5.2 Evaluation

1. [2 pts] Evaluate the trained model on the training data **without imputation**. What is the training accuracy of your NB model? Round your answer to 4 decimal places.

Final Answer

0.8329

2. [2 pts] Evaluate the trained model on the training data **with imputation**. What is the training accuracy of your NB model? Round your answer to 4 decimal places.

Final Answer

0.8328

3. [2 pts] Evaluate the trained model on the testing data **without imputation**. What is the testing accuracy of your NB model? Round your answer to 4 decimal places.

Final Answer

0.8298

4. [2 pts] Evaluate the trained model on the testing data with imputation. What is the testing accuracy of your NB model? Round your answer to 4 decimal places.

Final Answer

0.8297

5. [8 pts] Instead of training the NB using all training data, train only with the first  $m$  data by following these steps:

- Select the first  $m$  data points including lines with missing values and call this your training data.
- Remove lines with missing values from your training data (so you have  $m - m'$  rows where  $m'$  rows contain missing values).
- Train on the  $m - m'$  data and test on the entire testing data.
- Repeat step (a) - (c) for  $m = \{2^i \text{ for } i = 5, 6, 7, \dots, 13\}$  (i.e.  $m = 32, \dots, 8192$ )
- Report training accuracy over the  $m$  samples and testing accuracy over all of the test data.
- Plot training and testing accuracies calculated in (e) vs. # of training data.

(Important: Use " $\leq 50K$ " as a label if  $P_{l_{eq}} > P_{gr}$  else " $> 50K$ " to break ties.)

Do the steps above for both the data with imputation and the data without imputation, be sure to label which graph is which. Compare the results between using the data set with imputation and without, explain briefly what you observe. In addition at what values of  $m$  do testing accuracy and training accuracy attain their maximums, respectively for the datasets with and without imputation? In general, what would you expect to happen if we use only a few (say  $m < 3$ ) training data for Naive Bayes? Explain briefly (hint: we did not use smoothing). Please put your solutions the box on the next page.