Homework 5 - Mathetmatical Foundations of Machine Learning Engineers

kipngeno koech - bkoech

November 24, 2024

1 Probability (70 points)

1.1 MAP (30 Points)

In this question, we will use the Geometric distribution (the distribution of independent Bernoulli trials until the first success) to model an airport baggage claim. Let B be the number of your bag (meaning you watched $B-1$ bags pass before yours arrived). The probability distribution of B is given by

$$Pr(B = b) = (1-p)^{b-1}p \quad \text{for } b = 1, 2, \ldots$$

where $0 < p \leq 1$ is the parameter of the distribution.

We now will use a Beta prior to get a MAP estimate of the distribution parameter p. Specifically,

$$Pr(p|\alpha, \beta) = p^{\alpha-1}(1-p)^{\beta-1}$$

where $\alpha > 0$ and $\beta > 0$ are fixed, given constants.

1. Assume that over n trips, you noted your bag's number as $b_1, b_2, \ldots, b_n$ respectively, and the number of your bag on each of these trips is independent. Derive the log posterior probability of recording these numbers $b_1, b_2, \ldots, b_n$:

   $$\log Pr(p|b_1, \ldots, b_n)$$

   (10 Points)

   Hint: you can use Bayes' Rule.
   Baye's rule: Baye's rule in this case is the posterior probability of p given $b_1, \ldots, b_n$ is given by:
   geometric distribution:

   $$Pr(B = b) = (1-p)^{b-1}p$$

   beta prior:

   $$Pr(p|\alpha, \beta) = p^{\alpha-1}(1-p)^{\beta-1}$$

   Bayes theorem:

   $$Pr(A|B) = \frac{Pr(B|A)Pr(A)}{Pr(B)}$$

   where A and B in this case are the parameters p and $b_1, \ldots, b_n$ respectively. Therefore, the posterior probability of p given $b_1, \ldots, b_n$ is:

   $$Pr(p|b_1, \ldots, b_n) = \frac{Pr(b_1, \ldots, b_n|p)Pr(p)}{Pr(b_1, \ldots, b_n)}$$

   here $Pr(b_1, \ldots, b_n)$ is the marginal probability of $b_1, \ldots, b_n$ which can be calculated as:

   $$Pr(b_1, \ldots, b_n) = \int Pr(b_1, \ldots, b_n|p)Pr(p)dp$$

   where $Pr(b_1, \ldots, b_n|p)$ is the likelihood of $b_1, \ldots, b_n$ given p and $Pr(p)$ is the prior probability of p. Therefore, the log posterior probability of recording these numbers $b_1, b_2, \ldots, b_n$ is:

   $$\log Pr(p|b_1, \ldots, b_n) = \log\left(\frac{Pr(b_1, \ldots, b_n|p)Pr(p)}{Pr(b_1, \ldots, b_n)}\right)$$

   since the log of a fraction is the difference of the logs of the numerator and the denominator:

   $$= \log Pr(b_1, \ldots, b_n|p) + \log Pr(p) - \log Pr(b_1, \ldots, b_n)$$

   but the $Pr(b_1, \ldots, b_n)$ does not depend on p so we can ignore it:

   $$= \log Pr(b_1, \ldots, b_n|p) + \log Pr(p) + C$$

   where C is a constant (to represent what doesn't depend on p)
   we find the product of $Pr(b_1, \ldots, b_n)$ because it is a product of independent events:

   $$= \log\left(\prod_{i=1}^{n}(1-p)^{b_i-1}p\right) + \log\left(p^{\alpha-1}(1-p)^{\beta-1}\right) + C$$

   so our log posterior probability of recording these numbers $b_1, b_2, \ldots, b_n$ is:

   $$= \sum_{i=1}^{n}\log\left((1-p)^{b_i-1}p\right) + \log\left(p^{\alpha-1}(1-p)^{\beta-1}\right) + C$$

   if we expand this further:

   $$= \sum_{i=1}^{n}\left((b_i-1)\log(1-p) + \log p\right) + (\alpha-1)\log p + (\beta-1)\log(1-p) + C$$

2. Use your answer in part (1) to derive the maximum-a-posterior estimate (MAP) of the parameter:

$$\hat{p} = \arg\max_{0<p\leq 1} \log Pr(p|b_1, \ldots, b_n)$$

(10 Points)

$$\hat{p} = \arg\max_{0<p\leq 1} \log Pr(p|b_1, \ldots, b_n)$$

our log posterior probability of recording these numbers $b_1, b_2, \ldots, b_n$ is:

$$= \sum_{i=1}^{n} ((b_i - 1)\log(1-p) + \log p) + (\alpha - 1)\log p + (\beta - 1)\log(1-p) + C$$

to find the maximum of this function, we take the derivative with respect to p and set it to zero:

$$\frac{d}{dp}\log Pr(p|b_1, \ldots, b_n) = \frac{d}{dp}\left(\sum_{i=1}^{n} ((b_i - 1)\log(1-p) + \log p) + (\alpha - 1)\log p + (\beta - 1)\log(1-p) + C\right)$$

$$= \sum_{i=1}^{n}\left(\frac{d}{dp}(b_i - 1)\log(1-p) + \frac{d}{dp}\log p\right) + \frac{d}{dp}(\alpha - 1)\log p + \frac{d}{dp}(\beta - 1)\log(1-p) + \frac{d}{dp}C$$

$$= \sum_{i=1}^{n}\left(\frac{d}{dp}(b_i - 1)\log(1-p) + \frac{d}{dp}\log p\right) + \frac{d}{dp}(\alpha - 1)\log p + \frac{d}{dp}(\beta - 1)\log(1-p) + 0$$

$$= \sum_{i=1}^{n}\left(\frac{d}{dp}(b_i - 1)\log(1-p) + \frac{d}{dp}\log p\right) + \frac{\alpha - 1}{p} - \frac{\beta - 1}{1-p}$$

the derivatives inside the summation are:

$$\frac{d}{dp}(b_i - 1)\log(1-p) = (b_i - 1)\frac{d}{dp}\log(1-p) = (b_i - 1)\frac{-1}{1-p} = \frac{1 - b_i}{1-p}$$

$$\frac{d}{dp}\log p = \frac{1}{p}$$

so the derivative of the log posterior probability of recording these numbers $b_1, b_2, \ldots, b_n$ is:

$$= \sum_{i=1}^{n}\left(-\frac{b_i - 1}{1-p} + \frac{1}{p}\right) + \frac{\alpha - 1}{p} - \frac{\beta - 1}{1-p}$$

simplifying this:

$$= \sum_{i=1}^{n}\left(\frac{1}{p} - \frac{b_i - 1}{1-p}\right) + \frac{\alpha - 1}{p} - \frac{\beta - 1}{1-p}$$

setting this to zero to find the maximum:

$$\sum_{i=1}^{n}\left(\frac{1}{p} - \frac{b_i - 1}{1-p}\right) + \frac{\alpha - 1}{p} - \frac{\beta - 1}{1-p} = 0$$

simplifying this:

$$\frac{\alpha - 1 + n}{p} - \frac{\beta - 1 + \sum_{i=1}^{n}(b_i - 1)}{1-p} = 0$$

take the first term to the other side:

$$\frac{\alpha - 1 + n}{p} = \frac{\beta - 1 + \sum_{i=1}^{n}(b_i - 1)}{1-p}$$

cross multiply:

$$(\alpha - 1 + n)(1-p) = p(\beta - 1 + \sum_{i=1}^{n}(b_i - 1))$$

expand the left side:

$$\alpha - 1 + n - p(\alpha - 1 + n) = p(\beta - 1 + \sum_{i=1}^{n}(b_i - 1))$$

take the p terms to the righ side:

$$\alpha - 1 + n = p(\alpha - 1 + n + \beta - 1 + \sum_{i=1}^{n}(b_i - 1))$$

so:

$$\hat{p} = \frac{\alpha - 1 + n}{\alpha - 2 + n + \beta + \sum_{i=1}^{n}(b_i - 1)}$$

3. Suppose that $n = 5$ and the values of $b_1, b_2, \ldots, b_n$ are 10, 9, 5, 28, 7. Also, $\alpha = 14$ and $\beta = 590$. Using your answer in part (2), what is the MAP estimate of the p parameter for this data and prior? (10 Points)

$$\hat{p} = \frac{\alpha - 1 + n}{\alpha - 2 + n + \beta + \sum_{i=1}^{n}(b_i - 1)}$$

$$\hat{p} = \frac{14 - 1 + 5}{14 - 2 + 5 + 590 + 10 + 9 + 5 + 28 + 7 - 5}$$

$$\hat{p} = \frac{18}{14 + 5 + 590 + 10 + 9 + 5 + 28 + 7 - 5}$$

$$\hat{p} = \frac{18}{663} = \frac{6}{221} = \mathbf{0.027149}$$

1.2 Naive Bayes (20 Points)

You are asked to build a Naïve Bayes classifier using the training dataset in Table 1, where each instance is assigned to one out of 3 classes ("healthy" (H), "influenza" (I), or "salmonella poisoning" (S)).

Training	Fever (F)	Vomiting (V)	Diarrhea (D)	Classification
D1	no	no	no	Healthy (H)
D2	average	no	no	Influenza (I)
D3	high	no	no	Influenza (I)
D4	high	yes	yes	Salmonella poisoning (S)
D5	average	no	yes	Salmonella poisoning (S)

Table 1: Health Classification based on Symptoms

1. Using the Naïve Bayes model, find the prior probabilities $P(H)$, $P(I)$, and $P(S)$ given the training data above. (4 points)

$$P(H) = \frac{\mathbf{1}}{\mathbf{5}}, \quad P(I) = \frac{2}{5}, \quad P(S) = \frac{2}{5}$$

2. Fill in Table 2 below to finish building your Naïve Bayes classifier. Use Laplace smoothing with parameter $\alpha = 2$ for your conditional probability estimates. (8 points)

$$P(X|Y) = \frac{N_{X,Y} + \alpha}{N_Y + \alpha \cdot |X|}$$

where $N_{X,Y}$ is the number of occurrences of X in class Y, N_Y is the number of occurrences of class Y, and $|X|$ is the number of possible values of X.

$$P(X|Y) = \frac{N_{X,Y} + 2}{N_Y + 2 \cdot |X|}$$

for X and Y in the table below.

x = high fever:

$$P(x|H) = \frac{0+2}{1+2\cdot 3} = \frac{\mathbf{2}}{\mathbf{7}}$$
$$P(x|I) = \frac{1+2}{2+2\cdot 3} = \frac{\mathbf{3}}{\mathbf{8}}$$
$$P(x|S) = \frac{1+2}{2+2\cdot 3} = \frac{\mathbf{3}}{\mathbf{8}}$$

x = average fever:

$$P(x|H) = \frac{0+2}{1+2\cdot 3} = \frac{\mathbf{2}}{\mathbf{7}}$$
$$P(x|I) = \frac{1+2}{2+2\cdot 3} = \frac{\mathbf{3}}{\mathbf{8}}$$
$$P(x|S) = \frac{1+2}{2+2\cdot 3} = \frac{\mathbf{3}}{\mathbf{8}}$$

x = fever no:

$$P(x|H) = \frac{1+2}{1+2\cdot 3} = \frac{\mathbf{3}}{\mathbf{7}}$$
$$P(x|I) = \frac{0+2}{2+2\cdot 3} = \frac{\mathbf{1}}{\mathbf{4}}$$
$$P(x|S) = \frac{0+2}{2+2\cdot 3} = \frac{\mathbf{1}}{\mathbf{4}}$$

x = vomiting yes:

$$P(x|H) = \frac{0+2}{1+2\cdot 2} = \frac{\mathbf{2}}{\mathbf{5}}$$
$$P(x|I) = \frac{0+2}{2+2\cdot 2} = \frac{\mathbf{2}}{\mathbf{6}}$$
$$P(x|S) = \frac{1+2}{2+2\cdot 2} = \frac{\mathbf{3}}{\mathbf{6}}$$

x = vomiting no:

$$P(x|H) = \frac{1+2}{1+2\cdot 2} = \frac{\mathbf{3}}{\mathbf{5}}$$
$$P(x|I) = \frac{2+2}{2+2\cdot 2} = \frac{\mathbf{4}}{\mathbf{6}}$$
$$P(x|S) = \frac{1+2}{2+2\cdot 2} = \frac{\mathbf{3}}{\mathbf{6}}$$

x = diarrhea = yes:

$$P(x|H) = \frac{0+2}{1+2\cdot 2} = \frac{\mathbf{2}}{\mathbf{5}}$$
$$P(x|I) = \frac{0+2}{2+2\cdot 2} = \frac{\mathbf{2}}{\mathbf{6}}$$
$$P(x|S) = \frac{2+2}{2+2\cdot 2} = \frac{\mathbf{4}}{\mathbf{6}}$$

x = diarrhea = no:

$$P(x|H) = \frac{1+2}{1+2\cdot 2} = \mathbf{\frac{3}{5}}$$
$$P(x|I) = \frac{2+2}{2+2\cdot 2} = \mathbf{\frac{4}{6}}$$
$$P(x|S) = \frac{0+2}{2+2\cdot 2} = \mathbf{\frac{2}{6}}$$

$P(X\|Y)$	$Y = H$	$Y = I$	$Y = S$
$X =$ (high fever)	$\frac{2}{7}$	$\frac{3}{8}$	$\frac{3}{8}$
$X =$ (average fever)	$\frac{2}{7}$	$\frac{3}{8}$	$\frac{3}{8}$
$X =$ (no fever)	$\frac{3}{7}$	$\frac{1}{4}$	$\frac{1}{4}$
$X = V = Yes$	$\frac{2}{5}$	$\frac{2}{6}$	$\frac{3}{6}$
$X = V = No$	$\frac{3}{5}$	$\frac{4}{6}$	$\frac{3}{6}$
$X = D = yes$	$\frac{2}{5}$	$\frac{2}{6}$	$\frac{4}{6}$
$X = D = no$	$\frac{3}{5}$	$\frac{4}{6}$	$\frac{2}{6}$

Table 2: Conditional Probability Table for $P(X|Y)$

3. Apply your Naïve Bayes Classifier to a person who is vomiting but has no fever or diarrhea. Determine the probabilities of this person being healthy, suffering from influenza, and salmonella poisoning. (8 points)
Let V be vomiting, F be fever, and D be diarrhea. The probabilities of this person being healthy, suffering from influenza, and salmonella poisoning are:
Naive Bayes classifier:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

where X is the evidence and Y is the class. We can ignore the denominator since it is constant for all classes:

$$P(Y|X) \propto P(X|Y)P(Y)$$

probability that the person is healthy:

$$P(H|V = \text{yes}, F = \text{no}, D = \text{no}) \propto P(V = \text{yes}|H)P(F = \text{no}|H)P(D = \text{no}|H)P(H)$$
$$= \frac{2}{5}\cdot\frac{3}{7}\cdot\frac{3}{5}\cdot\frac{1}{5} = \frac{18}{875}$$

probability that the person is suffering from influenza:

$$P(I|V = \text{yes}, F = \text{no}, D = \text{no}) \propto P(V = \text{yes}|I)P(F = \text{no}|I)P(D = \text{no}|I)P(I)$$
$$= \frac{2}{6}\cdot\frac{1}{4}\cdot\frac{4}{6}\cdot\frac{2}{5} = \frac{1}{45}$$

probability that the person is suffering from salmonella poisoning:

$$P(S|V = \text{yes}, F = \text{no}, D = \text{no}) \propto P(V = \text{yes}|S)P(F = \text{no}|S)P(D = \text{no}|S)P(S)$$
$$= \frac{3}{6}\cdot\frac{1}{4}\cdot\frac{2}{6}\cdot\frac{2}{5} = \frac{1}{60}$$

let us sum the probabilities:

$$P(H|V = \text{yes}, F = \text{no}, D = \text{no}) + P(I|V = \text{yes}, F = \text{no}, D = \text{no}) + P(S|V = \text{yes}, F = \text{no}, D = \text{no}) = \frac{18}{875} + \frac{1}{45} + \frac{1}{60} = \frac{1873}{31500}$$

Normalizing the probabilities:

$$P(H|V = \text{yes}, F = \text{no}, D = \text{no}) = \frac{18}{875}\cdot\frac{31500}{1873} = \frac{36}{1873} = \mathbf{0.345969}$$
$$P(I|V = \text{yes}, F = \text{no}, D = \text{no}) = \frac{1}{45}\cdot\frac{31500}{1873} = \frac{700}{1873} = \mathbf{0.373978}$$
$$P(S|V = \text{yes}, F = \text{no}, D = \text{no}) = \frac{1}{60}\cdot\frac{31500}{1873} = \frac{525}{1873} = \mathbf{0.280053}$$

The probabilities of this person being healthy, suffering from influenza, and salmonella poisoning are 0.345969, 0.373978, and 0.280053 respectively. This shows that this person is more likely to be suffering from influenza.

1.3 MLE (20 points)

- A variable y is called a count if it only takes non-negative integer values, i.e., $y \in \{0, 1, 2, \ldots\}$. A common distribution for handling such variables is the Poisson distribution, which has the following form:

$$\text{Poisson}(y; \lambda) = \frac{e^{-\lambda}\lambda^y}{y!}, \quad y \in \{0, 1, 2, \ldots\}$$

where $\lambda > 0$ is some known constant that characterizes the Poisson distribution.

Given independent identically distributed (iid) observations $\{y_i\}_{i=1}^N$, $y_i \in \{0, 1, 2, \ldots\}$, calculate the MLE estimate of λ. (8 Points)

$$L_Y(\lambda) = \prod_{i=1}^{N} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} = \frac{e^{-N\lambda}\lambda^{\sum_{i=1}^N y_i}}{\prod_{i=1}^N y_i!}$$

$$\log L_Y(\lambda) = -N\lambda + \sum_{i=1}^{N} y_i \log \lambda - \sum_{i=1}^{N} \log y_i!$$

$$\frac{d}{d\lambda} \log L_Y(\lambda) = -N + \frac{1}{\lambda}\sum_{i=1}^{N} y_i$$

$$\frac{d}{d\lambda} \log L_Y(\lambda) = 0 \implies \lambda = \frac{1}{N}\sum_{i=1}^{N} y_i$$

$$\hat{\lambda}_{\text{MLE}} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

- In this problem, you will examine the task of estimating the probability density of the maximum height obtained by waves in the ocean. Scientists have recorded the maximum wave height on n days, obtaining samples $x_1, x_2, \ldots, x_n \in \mathbb{R}$. It is known that these are i.i.d. random variables following the Rayleigh distribution with parameter θ. Consider the following probability density function for the Rayleigh distribution:

$$f_X(x; \theta) = \frac{x}{\theta^2} \exp\left(-\frac{x^2}{2\theta^2}\right)$$

The likelihood function for your estimate is then $L_X(\theta) = f_X(x_1, \ldots, x_n|\theta)$. Your task is to estimate $\hat{\theta}_{\text{MLE}}$, the maximum likelihood estimate of θ. (12 Points)

The likelihood function is:

$$L_X(\theta) = \prod_{i=1}^{n} \frac{x_i}{\theta^2} \exp\left(-\frac{x_i^2}{2\theta^2}\right) = \frac{1}{\theta^{2n}} \exp\left(-\frac{1}{2\theta^2}\sum_{i=1}^{n} x_i^2\right) \prod_{i=1}^{n} x_i$$

$$\log L_X(\theta) = -2n \log \theta - \frac{1}{2\theta^2}\sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} \log x_i$$

dropping the values that do not depend on θ since their derivative will be zero:

$$\log L_X(\theta) = -2n \log \theta - \frac{1}{2\theta^2}\sum_{i=1}^{n} x_i^2$$

$$\frac{d}{d\theta} \log L_X(\theta) = -\frac{2n}{\theta} + \frac{1}{\theta^3}\sum_{i=1}^{n} x_i^2$$

we then set the derivative to zero to find the maximum likelihood estimate of θ:

$$-\frac{2n}{\theta} + \frac{1}{\theta^3}\sum_{i=1}^{n} x_i^2 = 0$$

$$\frac{2n}{\theta} = \frac{1}{\theta^3}\sum_{i=1}^{n} x_i^2$$

multiply both sides by θ^3:

$$2n\theta^2 = \sum_{i=1}^{n} x_i^2$$

divide both sides by $2n$:

$$\hat{\theta}_{\text{MLE}} = \sqrt{\frac{1}{2n}\sum_{i=1}^{n} x_i^2}$$

2 Text classification (30 Points)

Consider a text classification problem. In this case, you will try to classify text as either spam or ham. To do this, you will apply concepts of Likelihood, prior, and posterior given a dataset comprising pairs of text and labels. There are two types of labels: 1 (spam) and 0 (ham). Your goal is to create a simple classifier that, when given, determines if the text is spam or ham. You have been provided with the starter code and the data

1. Find the priors. What are the priors in this distribution? i.e find $P(ham)$ and $P(spam)$
2. Find the likelihoods for each word. For each word in the dataset, find the likelihood that the word is in spam and ham. This will represent the conditional probability $P(w|\text{spam})$ and $P(w|\text{ham})$ for w where $w \in V$. V is the vocabulary of the dataset.
3. Define a function that, when given a text sequence, returns the probability of the text being in spam. I.e., it returns $P(\text{spam}|\text{text})$. Note that this function calculates the likelihood using the Bayes rule. Do the same for ham.
4. Perform inference, i.e., given a string of text, determine if it is ham or spam based on the posterior probabilities calculated from the previous steps. Your function will determine the posterior probability of your text being in ham and spam and classify it as being the larger of the two.
5. Evaluate the data based on your test set and report the accuracy of your classifier. Your accuracy must be greater than 85%.