Homework 2

Gareth (Zhangchi) Sun
NetID: zs389

Course Name: CS5785 Fall 2023
Professor's Name: Prof. Kyra Gan

---

1 Naive Bayes with Binary Features

Consider a group of 50 Cornell Students. 20 of them are Master's students, while the rest 30 of them are PhD students. There are 5 Master's students who bike, and there are 5 Master's students who ski. On the other hand, 20 PhD students bike, and 15 PhD students ski.

We can formulate this as a machine learning problem by modeling the students with features $x = (x_1, x_2) \in \{0, 1\}^2$, where x_1 is a binary indicator of whether the students bike and x_2 is a binary indicator of whether they ski, and the target y equals 1 if they are PhD students and 0 if they are Master's students.

1. Please elaborate in this context what is the Naive Bayes assumption.

   Answer:

   The features of our samples are independent from each other. It means that whether you are a Master student or you are a PhD student, it will not influence your preference for bike or ski.On the other hand, whether you choose ski or bike is also independent. So, whether choose to ski won't influence one's probability of choosing to bike.

2. With the Naive Bayes assumption, find the probability of a student in this group who neither bikes or skis being a Master's student.

   Answer:

   The probability of a student in this group who neither bikes or skis being a Master's student is $\frac{9}{13}$ (0.6923).

$$
\begin{aligned}
P(x_1 = 0|y = 0) = P(x_2 = 0|y = 0) &= \frac{15}{20} = \frac{3}{4} \\
P(x_1 = 0|y = 1) &= \frac{1}{3} \\
P(x_2 = 0|y = 1) &= \frac{1}{2} \\
P(y = 0) &= \frac{20}{50} = \frac{2}{5} \\
P(y = 1) &= \frac{30}{50} = \frac{3}{5} \\
P(x = (0, 0)|y = 0) = P(x_1 = 0|y = 0) \times P(x_2 = 0|y = 0) \times P(y = 0) &= \frac{9}{16} \times \frac{2}{5} = \frac{9}{40} \\
P(x = (0, 0)|y = 1) = P(x_1 = 0|y = 1) \times P(x_2 = 0|y = 1) \times P(y = 1) &= \frac{1}{6} \times \frac{3}{5} = \frac{1}{10} \\
P(x = (0, 0)) = P(x = (0, 0)|y = 1) + P(x = (0, 0)|y = 1) &= \frac{13}{40} \\
P(y = 0|x = (0, 0)) = \frac{P(x = (0, 0)|y = 0)}{P(x = (0, 0))} &= \frac{9}{13} \approx 0.6923
\end{aligned}
$$

3. Suppose we know that every PhD who skis also bikes. Does it make sense to still assume that probability of biking and skiing are conditionally independent for a PhD student? If not, how would your answer to part (b) change with this knowledge (you can still assume probability of biking and skiing are conditionally independent for a Master's student)?

   Answer: This will influence the independence of the assumption of the Naive Bayes model. Now, if we know a PhD skis, then he will definitely bike. These two options are not independent anymore. It doesn't make sense to still assume that probability of biking and skiing are conditionally independent. Also, since we will use the probability of $P(x = (0, 0)|y = 1)$ in the part (b) to calculate the $P(x = (0, 0))$. So, the answer will also change.

2 Categorical Naive Bayes

In the Categorical Naive Bayes algorithm, we model this data via a probabilistic model $P_\theta(x, y)$.

- The distribution $P_\theta(y)$ is Categorical with parameters $\phi = (\phi_1, \ldots, \phi_K)$ and

$$P_\theta(y = k) = \phi_k$$

- The distribution of each feature x_j conditioned on $y = k$ is a Categorical distribution with parameters $\psi_{jk} = \left(\psi_{jk1}, \ldots, \psi_{jkL}\right)$, where

$$P_\theta\left(x_j = \ell \mid y = k\right) = \psi_{jk\ell}.$$

- The distribution over a vector of features x is given by

$$P_\theta(x \mid y = k) = \prod_{j=1}^{d} P_\theta\left(x_j \mid y = k\right),$$

  which is just the Naive Bayes factorization of $P_\theta(x \mid y = k)$.

In other words, the prior distribution $P_\theta(y)$ in this model is the same as in Bernoulli Naive Bayes. The distribution $P_\theta(x \mid y = k)$ is a product of Categorical distributions, whereas in Bernoulli Naive Bayes it was the product of Bernoulli distributions.

The total set of parameters of this model is $\theta = (\phi_1, \ldots \phi_K, \psi_{111}, \ldots \psi_{dKL})$. We learn the parameters via maximum likelihood:

$$\max_\theta \frac{1}{n} \sum_{i=1}^{n} \log P_\theta\left(x^{(i)}, y^{(i)}\right)$$

1. Show that the maximum likelihood estimate for the parameters ϕ_k is

$$\phi_k^* = \frac{n_k}{n}$$

   where n_k is the number of data points with class k.

$$L(\theta) = \max \frac{1}{n} \sum_{i=1}^{n} \log P_\theta(x^{(i)}, y^{(i)})$$

$$L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log P_\theta(x^{(i)} | y^{(i)}) + \frac{1}{n} \sum_{i=1}^{n} \log P_\theta(y^{(i)})$$

$$L(\phi_k) = \arg\max_{\phi_k} \left(\frac{1}{n} \sum_{i=1}^{n} \log P_\theta(x^{(i)} | y^{(i)}) + \frac{1}{n} \sum_{i=1}^{n} \log P_\theta(y^{(i)}) \right)$$

Because the parameter ϕ_k only appears in the $\left(\frac{1}{n}\sum_{i=1}^{n}\log P_\theta(y^{(i)})\right)$, so we can simplify the $L(\phi_k)$ to the following form:

$$L(\phi_k) = \frac{1}{n}\sum_{i=1}^{n}\log P_\theta(y^{(i)})$$
$$L(\phi_k) = \frac{1}{n}\sum_{i:y^{(i)}=k}^{n_k}\log P_\theta(y^{(i)}) + \frac{1}{n}\sum_{i:y^{(i)}\neq k}^{n-n_k}\log P_\theta(y^{(i)})$$
$$L(\phi_k) = \frac{1}{n}\sum_{i:y^{(i)}=k}^{n_k}\log \phi_k + \frac{1}{n}\sum_{i:y^{(i)}\neq k}^{n-n_k}\log(1-\phi_k)$$
$$L(\phi_k) = \frac{n_k}{n}\log \phi_k + \frac{n-n_k}{n}\log(1-\phi_k)$$

So, In order to achieve the maximum value of the $L(\phi_k)$, we need to differentiate it and set its derivative to zero in order to find its maximum value.

$$L'(\phi_k) = \frac{n_k}{n\times\phi_k} - \frac{n-n_k}{n(1-\phi_k)} = 0$$

$$\phi_k^* = \frac{n_k}{n}$$

2. Show that the maximum likelihood estimate for the parameters $\psi_{jk\ell}$ is

$$\psi_{jk\ell}^* = \frac{n_{jk\ell}}{n_k}$$

where $n_{jk\ell}$ is the number of data points with class k for which the j-th feature equals ℓ.

By the same logic of last problem, we can know that as for the parameter ψ_{jkl}, if we want to find its maximum likelihood for it, we should focus max $\left(\frac{1}{n}\sum_{i=1}^{n}\log P_\theta(xi, yi)\right)$, so we can get:

$$L(\theta) = max\frac{1}{n}\sum_{i=1}^{n}\log P_\theta(x^{(i)}|y^{(i)}) + \frac{1}{n}\sum_{i=1}^{n}\log P_\theta(y^{(i)})$$
$$L(\psi_{jkl}) = max\frac{1}{n}\sum_{i=1}^{n}\log P_\theta(x^{(i)}|y^{(i)})$$
$$L(\psi_{jkl}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_j}\log P_\theta(x_j^{(i)}|y^{(i)}=k)$$
$$L(\psi_{jkl}) = \frac{1}{n}\sum_{i:y^{(i)}=k}^{n_k}\sum_{j=1}^{n_j}\log P_\theta(x_j^{(i)}|y^{(i)}=k)$$
$$L(\psi_{jkl}) = \frac{1}{n}\sum_{i:y^{(i)}=k}^{n_k}\sum_{j=1}^{n_j}\sum_{m:x_{jm}=l}^{n_l}\log P_\theta(x_{jm}^{(i)}|y^{(i)}=k) + \frac{1}{n}\sum_{i:y^{(i)}=k}^{n_k}\sum_{j=1}^{n_j}\sum_{m:x_{jm}\neq l}^{n-n_l}\log P_\theta(x_j^{(i)}|y^{(i)}=k)$$

Since we have $P_\theta\left(x_j = \ell \mid y = k\right) = \psi_{jk\ell}$, so we can get:

$$L(\psi_{jkl}) = \frac{1}{n} \sum_{i:y^{(i)}=k}^{n_k} \sum_{j=1}^{n_j} \sum_{m:x_{jm}=l}^{n_l} \log \psi_{jkl} + \frac{1}{n} \sum_{i:y^{(i)}=k}^{n_k} \sum_{j=1}^{n_j} \sum_{m:x_{jm}\neq l}^{n-n_l} \log(1-\psi_{jkl})$$

And since $n_{jk\ell}$ is the number of data points with class k for which the j-th feature equals ℓ, so the number of data points with class k for which the j-th feature not equals ℓ is $n - n_{jk\ell}$:

$$L(\psi_{jkl}) = \frac{n_j kl}{n} \log \psi_{jkl} + \frac{n - n_j kl}{n} \log(1-\psi_{jkl})$$

So, In order to achieve the maximum value of the $L(\psi_{jkl})$, we need to differentiate it and set its derivative to zero in order to find its maximum value. Then we will get:

$$L'(\psi_{jkl}) = \frac{n_{jkl}}{n} \times \frac{1}{\psi_{jkl}} - \frac{n - n_{jkl}}{n} \times \frac{1}{1-\psi_{jkl}} = 0$$
$$\psi^*_{jkl} = \frac{n_{jkl}}{n}$$