HOMEWORK 3:
WRITTEN EXERCISE PART

1 Multinomial Naïve Bayes [25/2 pts]

Consider the Multinomial Naïve Bayes model. For each point $(\mathbf{x}, y)$, $y \in \{0, 1\}$, $\mathbf{x} = (x_1, x_2, \dots, x_M)$ where each x_j is an integer from $\{1, 2, \dots, K\}$ for $1 \le j \le M$. Here K and M are two fixed integer.
Suppose we have N data points $\{(\mathbf{x}^{(i)}, y^{(i)}) : 1 \le i \le N\}$, generated as follows.

for $i \in \{1, \dots, N\}$:
  $y^{(i)} \sim$ Bernoulli(ϕ)
  for $j \in \{1, \dots, M\}$:
    $x_j^{(i)} \sim$ Multinomial$(\theta_{y^{(i)}}, 1)$

Here $\phi \in \mathbb{R}$ and $\theta_k \in \mathbb{R}^K (k \in \{0, 1\}$ are parameters. Note that $\sum_l \theta_{k,l} = 1$ since they are the parameters of a multinomial distribution.
Derive the formula for estimating the parameters ϕ and θ_k, as we have done in the lecture for the Bernoulli Naïve Bayes model. Show the steps.
since $y^{(i)}$ is calculated from the bernoulli distribution.

$$\phi = \frac{\sum_{i=1}^N I(y^{(i)} = 1)}{N}$$

it is given that x_j can take values from $[1, 2....., K]$ and $y^(i)$ can take values [0,1].θ is dependent on the values taken by x_j according to the bernoulli naive bayes model. Hence we can find θ_{k,x_j} when k=0 and k=1

representing θ_{k,x_j} for the following values of y

when y = 0

$$\theta_{0,x} = \frac{\sum_{i=1}^N I(y^{(i)} = 0) \bigwedge x_1^{(i)} \in [1, 2, ...K] \bigwedge x_2^{(i)} \in [1, 2, ...K] ... \bigwedge x_j^{(i)} \in [1, 2, ...K])}{\sum_{i=1}^N I(y^{(i)} = 0)} \text{ where } j \in [1, 2, ..., M]$$

using similar reasoning, we can derive formula for when y =1

$$\theta_{1,x} = \frac{\sum_{i=1}^N I(y^{(i)} = 1) \bigwedge x_1^{(i)} \in [1, 2, ...K] \bigwedge x_2^{(i)} \in [1, 2, ...K] ... \bigwedge x_j^{(i)} \in [1, 2, ...K])}{\sum_{i=1}^N I(y^{(i)} = 1)} \text{ where } j \in [1, 2, ..., M]$$

2 Logistic Regression [25/2 pts]

Suppose for each class $i \in \{1, \dots, K\}$, the class-conditional density $p(\mathbf{x}|y = i)$ is normal with mean $\mu_i \in \mathbb{R}^d$ and identity covariance:

$$p(\mathbf{x}|y = i) = N(\mathbf{x}|\mu_i, \mathbf{I}).$$

Prove that $p(y = i|\mathbf{x})$ is a softmax over a linear transformation of $\mathbf{x}$. Show the steps.
we are given $p(x|y = i)$, which can be represented as

$$p(y = i|x) = \frac{p(x|y = i)p(y = i)}{\sum_j \ p(x|y = i)p(y = i)} = \frac{exp(a_i)}{\sum_j \exp(a_i)}.$$

We can also represent $a_i = ln[p(x|y = i)p(y = i)]$

given that $p(x|y=i) = N(x|m_i, I)$, $a_i = ln[N(x|m_i, I)p(y=i)] and$

$N(x|m_i, I) = \frac{1}{2\pi^{(d/2)}} exp(\frac{-1}{2}||x - m_i||^2)$

Therefore $a_i = ln[\frac{1}{2\pi^{(d/2)}} exp(\frac{-1}{2}||x - m_i||^2).p(y=i)]$

$a_i = ln(\frac{1}{2\pi^{(d/2)}}) + ln(exp(\frac{-1}{2}||x - m_i||^2)) + ln(p(y=i))$

$ln(exp(\frac{-1}{2}||x - m_i||^2))$ can be expressed as $\frac{-1}{2}||x - m_i||^2$, and this expression can further be expressed as $\frac{-1}{2}||x - m_i||^2 = \frac{-1}{2}(x^T x + m_i^T m_i) + m_i^T x.$

Eliminating $\frac{-1}{2}x^T x$ from the main expression, a_i can be represented as follows

$a_i = ln(\frac{1}{2\pi^{(d/2)}}) - \frac{1}{2}m_i^T m_i + m_i^T x + ln(p(y=i))$

Furthermore, a_i can be represented in the form of $(w^i)^T x + b_i$, wherein $(w^i) = m_i$ and $b_i = ln(\frac{1}{2\pi^{(d/2)}}) - \frac{1}{2}m_i^T m_i + ln(p(y=i))$

Hence, $p(y=i|x) = \frac{exp(a_i)}{\sum_j \exp(a_i)} = \frac{exp((w^i)^T x + b_i)}{\sum_j \exp((w^i)^T x + b_i)}$, which proves that $p(y=i|x)$ is a softmax over linear transformation