# HOMEWORK 3:
# WRITTEN EXERCISE PART

## 1 Multinomial Naïve Bayes [25/2 pts]

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

**for** $i \in \{1, \ldots, N\}$:
  $y^{(i)} \sim$ Bernoulli$(\phi)$
  **for** $j \in \{1, \ldots, M\}$:
    $x_j^{(i)} \sim$ 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 $\phi$ and $\theta_k$, as we have done in the lecture for the Bernoulli Naïve Bayes model. Show the steps.

Since the $y^{(i)}$ is estimated from a bernoulli distribution $\phi$, we can simply write this as

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

We know that $x_j$ can take any set of values from the set of integers $\{\ 1, 2, ...K\ \}$. This means that $\theta$ is dependent on the values that are taken by $x_j$ for some data. Since $y$ can be 0 or 1, $\theta_{k,x_j}$ can be estimated for both cases where $k = 1$ and $k = 0$. The following formula indicates the $\theta_{k,x_j}$ for when $y = 0$

$$\theta_{0,x} = \frac{\sum_{i=1}^{N} \mathbb{I}(y^{(i)} = 0 \wedge x_1^{(i)} \in \{1, 2, ...K\} \wedge x_2^{(i)} \in \{1, 2, ...K\} \ ... \ \wedge x_j^{(i)} \in \{1, 2, ...K\})}{\sum_{i=1}^{N} I(y^{(i)} = 0)}$$

Where $j \in \{1, 2...M\}$
Similarly, we for $y = 1$, we have

$$\theta_{1,x} = \frac{\sum_{i=1}^{N} \mathbb{I}(y^{(i)} = 1 \wedge x_1^{(i)} \in \{1, 2, ...K\} \wedge x_2^{(i)} \in \{1, 2, ...K\} \ ... \ \wedge x_j^{(i)} \in \{1, 2, ...K\})}{\sum_{i=1}^{N} I(y^{(i)} = 1)}$$

## 2 Logistic Regression [25/2 pts]

Suppose for each class $i \in \{1, \ldots, K\}$, 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 the class conditional probabilities for multiclass classification as $p(x|y = i)$, hence the individual class probabilities are $p(y = i)$. Using bayes rule, we get the following

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

Hence, from this we can note that

$$a_i = ln[p(x|y=i)p(y=i)]$$

We are given that $p(x|y=i) = N(x|\mu_i, \mathbf{I})$, hence

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

The term $-\frac{1}{2}||x-u_i||^2$ can be written as $-\frac{1}{2}x^Tx - \frac{1}{2}u_i^Tu_i + u_i^Tx$ Also, We know that the term $a_i$ can be written as $ln[p(x|y=i)p(y=i)]$, this can be further simplified as follows

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

This can be further simplified as

$$ln[\frac{1}{(2\pi)^{d/2}}] - \frac{1}{2}u_i^Tu_i - \frac{1}{2}x^Tx + u_i^Tx + ln[p(y=i)]$$

Cancelling the 3rd term from the above, we get the following:

$$ln[\frac{1}{(2\pi)^{d/2}}] - \frac{1}{2}u_i^Tu_i + u_i^Tx + ln[p(y=i)]$$

The above equation can take the form $a_i = (w^i)^Tx + b_i$ where $w^i = u_i$ and $b_i = \frac{1}{2}u_i^Tu_i + ln[p(y=i)] + ln[\frac{1}{(2\pi)^{d/2}}]$
Thus, $\frac{exp(a_i)}{\sum_j exp(a_j)}$ can be written as

$$p(y=i|x) = \frac{exp((w^i)^Tx + b_i)}{\sum_j exp(w^j)^Tx + b_j)}$$

Hence, the above equation is a softmax over a linear transformation of x