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

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.

let, x^i = $\theta_{p,k}$
and, A_k^i is frequency of x^i be k

As y is a Bernoulli distribution, hence
$p(y|\phi) = \prod_{i=1}^{n} \phi^{y^(i)}(1-\phi)^{1-y^{(i)}}$

now, ϕ can be computed by minimizing the above expression.
as the expression is a product, hence we can take log on both sides for our convenience.

computing the derivative of log to be 0, we get

$\frac{\partial log(p(y|\phi)}{\partial \phi} = \frac{\partial}{\partial \phi}(-\sum_{i=1}^{n}(y^{(i)}\log(\phi) + (1-y^{(i)})\log(1-\phi))$
$\implies -\sum_{i=1}^{n} \frac{y^{(i)}}{\phi} - \frac{(1-y^{(i)})}{(1-\phi)} = 0$

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

now,
$p(y|x, \theta, \phi) \propto p(y|\phi)p(x|\theta, \phi)$

as per above derivation,
$p(y|\phi) = \prod_{i=1}^{n} p(y^i|\phi)$

also,
$p(x^i|\theta) = \prod_{k=1}^{K} \theta_{y^{(i)},k}^{A_k^i}$

so, posterior distribution of following has to be estimated
$p(y|x, \theta, \phi) = \prod_{i=1}^{n} p(y^i|\phi) \prod_{k=1}^{K} \theta_{y^{(i)},k}^{A_k^i}$

provided, $\sum_{k=1}^{K} \theta_{p,k} = 1$

$\implies -\log(posterior(L)) = [\sum_{i=1}^{n} y^{(i)} \log(\phi) + (1-y^i)\log(1-\phi)] \sum_{k=1}^{K} A_k^i \log(\theta_{y^{(i)},k})$

$\implies = [\sum_{i=1}^{n} y^{(i)} \log(\phi) + (1-y^i)\log(1-\phi) \sum_{k=1}^{K} A_k^i \log(\theta_{y^{(i)},k})] - \lambda_0(1-\sum_{k=1}^{K}\theta_{0,k}) - \lambda_1(1-\sum_{k=1}^{K}\theta_{1,k})$

summarizing above, we have taken added the Adding Lagrange multiplier
now, for best estimation, we need to minimize this expression

$\frac{\partial L}{\partial \theta}$ can be subdivided into

$\frac{\partial L}{\partial \theta_{0,k}}$; subject to $y^{(i)} = 0$, and $\frac{\partial L}{\partial \theta_{1,k}}$ subject to $y^{(i)} = 1$

now, $\frac{\partial L}{\partial \theta_{0,k}}$
$= \frac{\partial}{\partial \theta_{0,k}}[\sum_{i=1}^{n} y^{(i)} \log(\phi) + (1-y^i)\log(1-\phi) \sum_{k=1}^{K} A_k^i \log(\theta_{y^{(i)},k})] + \lambda_0(1-\sum_{k=1}^{K}\theta_{0,k}) + \lambda_1(1-\sum_{k=1}^{K}\theta_{1,k})$
$\implies = \frac{\partial}{\partial \theta_{0,k}} \sum_{i=1}^{n} 0 * \log(\phi) + (1-0)\log(1-\phi) \sum_{k=1}^{K} A_k^i \log(\theta_{y^{(i)},k}) + \lambda_0(1-\sum_{k=1}^{K}\theta_{0,k}) + \lambda_1(1-\sum_{k=1}^{K}\theta_{1,k})$
$\implies = \sum_{i=1,y^{(i)}=0}^{n} \phi * \frac{A_k^i}{\theta_{0,k}} - \lambda_0 = 0$

$$\implies \theta_{0,k} = \sum_{i=1,y^{(i)}=0}^{n} \phi * \frac{A_k^i}{\lambda_0}$$

using, $\sum_{k=1}^{K} \theta_{p,k} = 1$

$\implies \sum_{i=1,y^{(i)}=0}^{n} \sum_{k=1}^{K} \phi * \frac{A_k^i}{\lambda_0} = 1$

now, as A_k^i is frequency of x^i, therefore
$\sum_{k=1}^{K} A_j^i = M$

$\implies \frac{\lambda_0}{\phi} = M \sum_{i=1}^{n} I(y^i = 0)$

as derived before, $\theta_{0,k} = \sum_{i=1,y^{(i)}=0}^{n} \phi * \frac{A_k^i}{\lambda_0}$

$\implies = \frac{\sum_{i=1,y^{(i)}=0}^{n} A_k^i}{M \sum_{i=1}^{n} I(y^i=0)}$

$$\implies \theta_{0,k} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{M} I(y^{(i)} = 0, x_j^i = k)}{M \sum_{i=1}^{n} I(y^{(i)} = 0)}$$

Now, for computation of $\theta_{1,k}$ we need following replacements
$0 \to 1$
$\phi \to (1-\phi)$

so, $\frac{\partial L}{\partial \theta_{1,k}}$
$= \frac{\partial}{\partial \theta_{1,k}}[\sum_{i=1}^{n} y^{(i)} \log(\phi) + (1-y^i)\log(1-\phi) \sum_{k=1}^{K} A_k^i \log(\theta_{y^{(i)},k})] + \lambda_0(1-\sum_{k=1}^{K}\theta_{0,k}) + \lambda_1(1-\sum_{k=1}^{K}\theta_{1,k})$

$\implies = \frac{\partial}{\partial \theta_{1,k}} \sum_{i=1}^{n} 1 * \log(\phi) + (1-1)\log(1-\phi) \sum_{k=1}^{K} A_k^i \log(\theta_{y^{(i)},k}) + \lambda_0(1-\sum_{k=1}^{K}\theta_{0,k}) + \lambda_1(1-\sum_{k=1}^{K}\theta_{1,k})$

$\implies = \sum_{i=1,y^{(i)}=1}^{n} (1-\phi) * \frac{A_k^i}{\theta_{1,k}} - \lambda_1 = 0$

$$\implies \theta_{1,k} = \sum_{i=1,y^{(i)}=0}^{n} (1-\phi) * \frac{A_k^i}{\lambda_1}$$

following similar steps, we can compute,

$$\implies \theta_{1,k} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{M} I(y^{(i)} = 1, x_j^i = k)}{M\sum_{i=1}^{n} I(y^{(i)} = 1)}$$

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 that class-conditional density, is normal.
$\implies$ class-conditional density = $p(\mathbf{x}|y = i)$
$\implies = N(x|\mu_i, I)$
$\implies = \frac{1}{2\pi^{d/2}} \exp\left(\frac{-1}{2}||x - \mu_i||^2\right)$

assume that, $m_i = \log(p(\mathbf{x}|y = i)p(y = i)$

$$\implies = \frac{-1}{2}||x - \mu_i||^2 + \log\frac{1}{2\pi^{d/2}} + \log(p(y = i))$$

and, we know,

$$\frac{-1}{2}||x - \mu_i||^2 = \frac{-1}{2}x^T x + \mu_i^T x + \frac{-1}{2}\mu_i^T \mu_i$$

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

assume,
$n_i = \frac{-1}{2}\mu_i^T \mu_i + \log\frac{1}{2\pi^{d/2}} + \log(p(y = i))$

$$\implies m_i = \frac{-1}{2}x^T x + (w^i)^T x + n_i$$

now, using the bayesian rule, $p(y = i|\mathbf{x})$ is :

$$p(y = i|\mathbf{x}) = \frac{p(\mathbf{x}|y = i)p(y = i)}{\sum_j p(\mathbf{x}|y = j)p(y = j)}$$

hence, $p(y = i|\mathbf{x}) = \frac{e^{m_i}}{\sum_j e^{m_j}}$
$\implies = \frac{e^{\left(\frac{-1}{2}x^T x + (w^i)^T x + n_i\right)}}{\sum_j e^{\left(\frac{-1}{2}x^T x + (w^j)^T x + n^j\right)}}$

$$\implies = \frac{e^{((w^i)^T x + n_i)}}{\sum_j e^{((w^j)^T x + n^j)}}$$

As this is a softmax over linear transformation of x, Hence Proved.