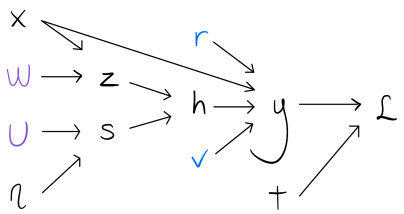
# CSC311 Summer 2024
# Homework 3

Maria Ma (1009054924)

1. (a) Below is the computation graph:




(b) $\frac{d\sigma(x)}{dx} = \frac{d}{dx}\left(\frac{1}{1+e^{-x}}\right) = \frac{e^{-x}}{(1+e^{-x})^2} = \frac{1}{1+e^{-x}}\left(1 - \frac{1}{1+e^{-x}}\right) = \sigma(x)(1-\sigma(x))$

(c) $\overline{\mathcal{L}} = 1$

$\overline{y} = \overline{\mathcal{L}}\frac{d\mathcal{L}}{dy} = \overline{\mathcal{L}}\frac{d}{dy}(t\log y + (1-t)\log(1-y)) = \overline{\mathcal{L}}(\frac{t}{y} - \frac{1-t}{1-y})$

$\overline{h_i} = \overline{y}\frac{dy}{dh_i} = \overline{y}\frac{dy}{d(v^Th+r^Tx)}\frac{d(v^Th+r^Tx)}{dh_i} = \overline{y}\sigma(v^Th+r^Tx)(1-\sigma(v^Th+r^Tx))v_i$

$\overline{v_i} = \overline{y}\frac{dy}{dv_i} = \overline{y}\frac{dy}{d(v^Th+r^Tx)}\frac{d(v^Th+r^Tx)}{dv_i} = \overline{y}\sigma(v^Th+r^Tx)(1-\sigma(v^Th+r^Tx))h_i$

$\overline{r_i} = \overline{y}\frac{dy}{dr_i} = \overline{y}\frac{dy}{d(v^Th+r^Tx)}\frac{d(v^Th+r^Tx)}{dr_i} = \overline{y}\sigma(v^Th+r^Tx)(1-\sigma(v^Th+r^Tx))x_i$

$\overline{s_i} = \overline{h_i}\frac{dh_i}{ds_i} = \overline{h_i}z_i$

$\overline{z_i} = \overline{h_i}\frac{dh_i}{dz_i} = \overline{h_i}s_i$

$\overline{U_{ij}} = \overline{s_i}\frac{ds_i}{dU_j} = \overline{s_i}\eta_j$

$\overline{w_{ij}} = \overline{z_i}\frac{dz_i}{dw_j} = \overline{z_i}x_j$

$\overline{\eta_j} = \sum_i \overline{s_i}\frac{ds_i}{d\eta_j} = \sum_i \overline{s_i}U_{ij}$

$\overline{x_j} = \sum_i \overline{z_i}\frac{dz_i}{dx_j} + \overline{y}\frac{dy}{dx_j} = \sum_i \overline{z_i}w_{ij} + \overline{y}\sigma(v^Th+r^Tx)(1-\sigma(v^Th+r^Tx))r_j$

2. (a) Decomposing the Log-Likelihood:

By the joint probability given in the question, $l(\theta,\pi) = \sum_{i=1}^N \log p(x^{(i)}, c^{(i)}|\theta,\pi) =$
$\sum_{i=1}^N \log p(c^{(i)}|\theta,\pi)p(x^{(i)}|c^{(i)},\theta,\pi) = \sum_{i=1}^N \log p(c^{(i)}|\pi)\prod_{j=1}^{784} p(x_j^{(i)}|c^{(i)},\theta_{jc}) =$
$\sum_{i=1}^N[\log p(c^{(i)}|\pi) + \sum_{j=1}^{784}\log p(x_j^{(i)}|c^{(i)},\theta_{jc})] = \sum_{i=1}^N[\log p(c^{(i)}|\pi) + \sum_{i=1}^N\sum_{j=1}^{784}\log p(x_j^{(i)}|c^{(i)},\theta_{jc})]$.

Learning the Prior over Class (maximizing $\sum_{i=1}^N \log p(c^{(i)}|\pi)$):

By the categorical distribution given in the question, $\sum_{i=1}^N \log p(c^{(i)}|\pi) = \sum_{i=1}^N \log\prod_{j=0}^9 \pi_j^{t_j^{(i)}} =$
$\sum_{i=1}^N\sum_{j=0}^9 \log \pi_j^{t_j^{(i)}} = \sum_{i=1}^N\sum_{j=0}^9 t_j^{(i)}\log\pi_j = \sum_{i=1}^N[\sum_{j=0}^8 t_j^{(i)}\log\pi_j + t_9^{(i)}\log(1-\sum_{j=0}^8\pi_j)]$.

Setting the derivative of the log-likelihood with respect to $\pi_j$ to zero gives $\sum_{i=1}^N \frac{t_j^{(i)}}{\pi_j} - \frac{t_9^{(i)}}{1-\sum_{j=0}^8\pi_j} =$
$\sum_{i=1}^N \frac{t_j^{(i)}}{\pi_j} - \frac{t_9^{(i)}}{\pi_9} = 0$, which yields $\sum_{i=1}^N \frac{t_j^{(i)}}{\pi_j} = \sum_{i=1}^N \frac{t_9^{(i)}}{\pi_9} \Rightarrow \frac{\sum_{i=1}^N t_j^{(i)}}{\sum_{i=1}^N t_9^{(i)}} = \frac{\hat{\pi}_j}{\hat{\pi}_9} \Rightarrow \hat{\pi}_j = \hat{\pi}_9\frac{\sum_{i=1}^N t_j^{(i)}}{\sum_{i=1}^N t_9^{(i)}}$.

Since $\sum_{j=0}^{9}\hat{\pi}_j = 1$, we know $1 = \sum_{j=0}^{9}\hat{\pi}_j = \sum_{j=0}^{8}\hat{\pi}_j + \hat{\pi}_9 = \sum_{j=0}^{8}(\hat{\pi}_9\frac{\sum_{i=1}^{N}t_j^{(i)}}{\sum_{i=1}^{N}t_9^{(i)}}) + \hat{\pi}_9\frac{\sum_{i=1}^{N}t_9^{(i)}}{\sum_{i=1}^{N}t_9^{(i)}} = \frac{\hat{\pi}_9}{\sum_{i=1}^{N}t_9^{(i)}}(\sum_{j=0}^{8}\sum_{i=1}^{N}t_j^{(i)} + \sum_{i=1}^{N}t_9^{(i)}) = \frac{\hat{\pi}_9}{\sum_{i=1}^{N}t_9^{(i)}}\sum_{j=0}^{9}\sum_{i=1}^{N}t_j^{(i)} = \frac{\hat{\pi}_9 \cdot N}{\sum_{i=1}^{N}t_9^{(i)}} \Rightarrow \hat{\pi}_9 = \frac{\sum_{i=1}^{N}t_9^{(i)}}{N}$ and $\hat{\pi}_j = \hat{\pi}_9\frac{\sum_{i=1}^{N}t_j^{(i)}}{\sum_{i=1}^{N}t_9^{(i)}} = \frac{\sum_{i=1}^{N}t_j^{(i)}}{N}$ for $j = 0, 1, ..., 8$.

Therefore we have $\hat{\pi}_j = \frac{\sum_{i=1}^{N}t_j^{(i)}}{N}$ for $j = 0, 1, ..., 9$, where the numerator represents the number of samples in class $j$ and the denominator represents the total number of samples.

Learning Pr. Feature Given Class (maximizing $\sum_{i=1}^{N}\sum_{j=1}^{784}\log p(x_j^{(i)}|c^{(i)}, \theta_{jc})$):

By the Bernoulli distribution given in the question, $\sum_{i=1}^{N}\sum_{j=1}^{784}\log p(x_j^{(i)}|c^{(i)}, \theta_{jc}) = \sum_{i=1}^{N}\sum_{j=1}^{784}\mathbb{I}(c^{(i)} = c)\{x_j^{(i)}\log\theta_{jc} + (1 - x_j^{(i)})\log(1 - \theta_{jc})\}$.

Setting the derivative of the log-likelihood with respect to $\theta_{jc}$ to zero gives $\sum_{i=1}^{N}\mathbb{I}(c^{(i)} = c)\{\frac{x_j^{(i)}}{\theta_{jc}} - \frac{1-x_j^{(i)}}{1-\theta_{jc}}\} = 0$, which yields $\sum_{i=1}^{N}\mathbb{I}(c^{(i)} = c)\{\frac{x_j^{(i)}}{\theta_{jc}}\} = \sum_{i=1}^{N}\mathbb{I}(c^{(i)} = c)\{\frac{1-x_j^{(i)}}{1-\theta_{jc}}\} \Rightarrow \frac{\sum_{i=1}^{N}\mathbb{I}(c^{(i)}=c)x_j^{(i)}}{\theta_{jc}} = \frac{\sum_{i=1}^{N}\mathbb{I}(c^{(i)}=c)(1-x_j^{(i)})}{1-\theta_{jc}} \Rightarrow \hat{\theta}_{jc} = \frac{\sum_{i=1}^{N}\mathbb{I}(c^{(i)}=c)x_j^{(i)}}{\sum_{i=1}^{N}\mathbb{I}(c^{(i)}=c)} \Rightarrow \hat{\theta}_{jc} = \frac{\sum_{i=1}^{N}t_c^{(i)}x_j^{(i)}}{\sum_{i=1}^{N}t_c^{(i)}}$.

Therefore we have $\hat{\theta}_{jc} = \frac{\sum_{i=1}^{N}t_c^{(i)}x_j^{(i)}}{\sum_{i=1}^{N}t_c^{(i)}}$, where the numerator represents the number of pixels $x_j$ in class $c$ and the denominator represents the total number of samples in class $c$.

(b) We know $p(t|x, \theta, \pi) = \frac{p(t,x|\theta,\pi)}{p(x|\theta,\pi)} = \frac{p(t,x|\theta,\pi)}{\sum_{j=0}^{9}p(t_j=1,x|\theta,\pi)} = \frac{p(t,x|\theta,\pi)}{\sum_{i=0}^{9}p(t_i=1|\pi)\prod_{j=1}^{784}p(x_j|t_i=1,\theta_{jc})} = \frac{p(t,x|\theta,\pi)}{\sum_{i=0}^{9}(\pi_i\prod_{j=1}^{784}\theta_{ij}^{x_j}(1-\theta_{ij})^{1-x_j})}$.

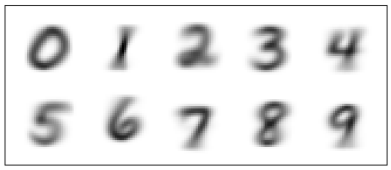
Then taking the log would yield $\log p(t|x, \theta, \pi) = \log p(t, x|\theta, \pi) - \log\sum_{i=0}^{9}(\pi_i\prod_{j=1}^{784}\theta_{ij}^{x_j}(1 - \theta_{ij})^{1-x_j}) = \sum_{i=0}^{9}t_i\log\pi_i + \sum_{j=1}^{784}(x_j\log\theta_{jc} + (1 - x_j)\log(1 - \theta_{jc})) - \log\sum_{i=0}^{9}(\pi_i\prod_{j=1}^{784}\theta_{ij}^{x_j}(1 - \theta_{ij})^{1-x_j})$.

(c)

```
Average log-likelihood for MLE is  nan
```

The average log-likelihood is underfined since some classes have a likelihood ($\hat{\theta}_{jc}$) of 0 but taking the log of 0 yields NAN.

(d)



(e) We know $p(\theta_{jc}) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta_{jc}^{a-1}(1 - \theta_{jc})^{b-1} \propto \theta_{jc}^{a-1}(1 - \theta_{jc})^{b-1}$.

Then taking the log would yield $\log p(\theta_{jc}) = \log\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} + (a - 1)\log\theta_{jc} + (b - 1)\log(1 - \theta_{jc})$.

Setting the derivative of the log-prior with respect to $\theta_{jc}$ to zero gives $\frac{a-1}{\theta_{jc}} - \frac{b-1}{1-\theta_{jc}} + \sum_{i=1}^{N}t_c^{(i)}\frac{x_j^{(i)}}{\theta_{jc}} - \sum_{i=1}^{N}t_c^{(i)}\frac{1-x_j^{(i)}}{1-\theta_{jc}} = 0$, which yields $\frac{a-1}{\theta_{jc}} + \sum_{i=1}^{N}t_c^{(i)}\frac{x_j^{(i)}}{\theta_{jc}} = \frac{b-1}{1-\theta_{jc}} + \sum_{i=1}^{N}t_c^{(i)}\frac{1-x_j^{(i)}}{1-\theta_{jc}} \Rightarrow$

$\frac{a-1+\sum_{i=1}^{N} t_c^{(i)} x_j^{(i)}}{\theta_{jc}} = \frac{b-1+\sum_{i=1}^{N} t_c^{(i)}(1-x_j^{(i)})}{1-\theta_{jc}} \Rightarrow \hat{\theta}_{jc_{MAP}} = \frac{a-1+\sum_{i=1}^{N} t_c^{(i)} x_j^{(i)}}{a+b-2+\sum_{i=1}^{N} t_c^{(i)}}$.

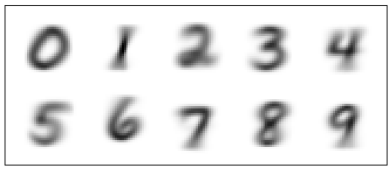
Taking $\alpha = 3, \beta = 3$, $\hat{\theta}_{jc_{MAP}} = \frac{3-1+\sum_{i=1}^{N} t_c^{(i)} x_j^{(i)}}{3+3-2+\sum_{i=1}^{N} t_c^{(i)}} = \frac{2+\sum_{i=1}^{N} t_c^{(i)} x_j^{(i)}}{4+\sum_{i=1}^{N} t_c^{(i)}}$.

Compared to the MLE estimator, the MAP estimator has constants added to the numerator (2) and the denominator (4), which depend on the $\alpha$ and $\beta$ values. The remaining part is identical.

(f)

```
Average log-likelihood for MAP is  -3.357063137860284
Training accuracy for MAP is  0.8352166666666667
Test accuracy for MAP is  0.816
```

(g)



(h) One advantage of Naive Bayes algorithm:

It is simple and easy to implement, and it is computationally efficient for both training and testing.

One disadvantage of Naive Bayes algorithm:

It assumes that the features are conditionally independent given the class, which may not be true in practice.

3. (a) Since $p(y=1|x,\theta) = \sigma(x^T\theta)$ as given in the question, the likelihood function is $L(\theta) = \prod_{i=1}^{N} p(y^{(i)}|x^{(i)},\theta) = \prod_{i=1}^{N} \sigma(x^{(i)T}\theta)^{y^{(i)}}(1-\sigma(x^{(i)T}\theta))^{1-y^{(i)}}$.

The log-likelihood is $\mathcal{L}(\theta) = \log L(\theta) = \sum_{i=1}^{N}[y^{(i)} \log \sigma(x^{(i)T}\theta) + (1-y^{(i)}) \log(1-\sigma(x^{(i)T}\theta))]$.

Since $\sigma(x) = \frac{1}{1+e^{-x}}$, the derivative of the log-likelihood with respect to $\theta$ is $\frac{d\mathcal{L}(\theta)}{d\theta} = \sum_{i=1}^{N}[y^{(i)} \frac{1}{\sigma(x^{(i)T}\theta)}\sigma(x^{(i)T}\theta)(1-\sigma(x^{(i)T}\theta))x^{(i)} - (1-y^{(i)}) \frac{1}{1-\sigma(x^{(i)T}\theta)}\sigma(x^{(i)T}\theta)(1-\sigma(x^{(i)T}\theta))x^{(i)}]$

$= \sum_{i=1}^{N}[y^{(i)}(1-\sigma(x^{(i)T}\theta))x^{(i)} - (1-y^{(i)})\sigma(x^{(i)T}\theta)x^{(i)}] = \sum_{i=1}^{N}[y^{(i)}x^{(i)} - \sigma(x^{(i)T}\theta)x^{(i)}]$.

The resulting log-likelihood would be optimized using $\theta \leftarrow \theta + \alpha \frac{d\mathcal{L}(\theta)}{d\theta} = \theta + \alpha \sum_{i=1}^{N}[y^{(i)}x^{(i)} - \sigma(x^{(i)T}\theta)x^{(i)}]$.

(b) Since $p(\theta) \sim \mathcal{N}(0, \sigma^2 I)$ as given in the question, the prior is $p(\theta) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp(-\frac{1}{2\sigma^2}\theta^T\theta)$.

The posterior is $p(\theta|D) \propto p(D|\theta)p(\theta)$, therefore to maximize the posterior, we need to maximize the log-posterior $\log p(\theta|D) = \log p(D|\theta) + \log p(\theta)$.

Therefore $\log p(\theta|D) = \log p(D|\theta) + \log p(\theta) = \log \prod_{i=1}^{N} p(y^{(i)}|x^{(i)},\theta) + \log p(\theta) =$

$\sum_{i=1}^{N}[y^{(i)} \log \sigma(x^{(i)T}\theta) + (1-y^{(i)}) \log(1-\sigma(x^{(i)T}\theta))] + \log(\frac{1}{(2\pi\sigma^2)^{d/2}} \exp(-\frac{1}{2\sigma^2}\theta^T\theta)) =$

$\sum_{i=1}^{N}[y^{(i)} \log \sigma(x^{(i)T}\theta) + (1-y^{(i)}) \log(1-\sigma(x^{(i)T}\theta))] - \frac{d}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\theta^T\theta$.

And the derivative of the log-posterior with respect to $\theta$ is $\frac{d \log p(\theta|D)}{d\theta} = \sum_{i=1}^{N}[y^{(i)}x^{(i)} - \sigma(x^{(i)T}\theta)x^{(i)}] - \frac{1}{\sigma^2}\theta$.

(The resulting log-posterior would be optimized using $\theta \leftarrow \theta + \alpha \frac{d \log p(\theta|D)}{d\theta} = \theta + \alpha[\sum_{i=1}^{N}[y^{(i)}x^{(i)} - \sigma(x^{(i)T}\theta)x^{(i)}] - \frac{1}{\sigma^2}\theta]$.

4. (a)

```
The average conditional log-likelihood for the training dataset: -0.125.
The average conditional log-likelihood for the test dataset: -0.197.
```

(b)

```
The accuracy for the training dataset: 0.981.
The accuracy for the test dataset: 0.973.
```

(c)

```
(Diagonal Covariance Matrix) The average conditional log-likelihood for the training dataset: -1.23.
(Diagonal Covariance Matrix) The average conditional log-likelihood for the test dataset: -1.29.
(Diagonal Covariance Matrix) The accuracy for the training dataset: 0.850.
(Diagonal Covariance Matrix) The accuracy for the test dataset: 0.840.
```

The average conditional log-likelihoods are lower if it is a diagonal covariance matrix since the model is less flexible and cannot capture the data as well as a full covariance matrix (diagonal covariance matrix cannot model dependence between pixels). Consequently, the average accuracies are also lower.