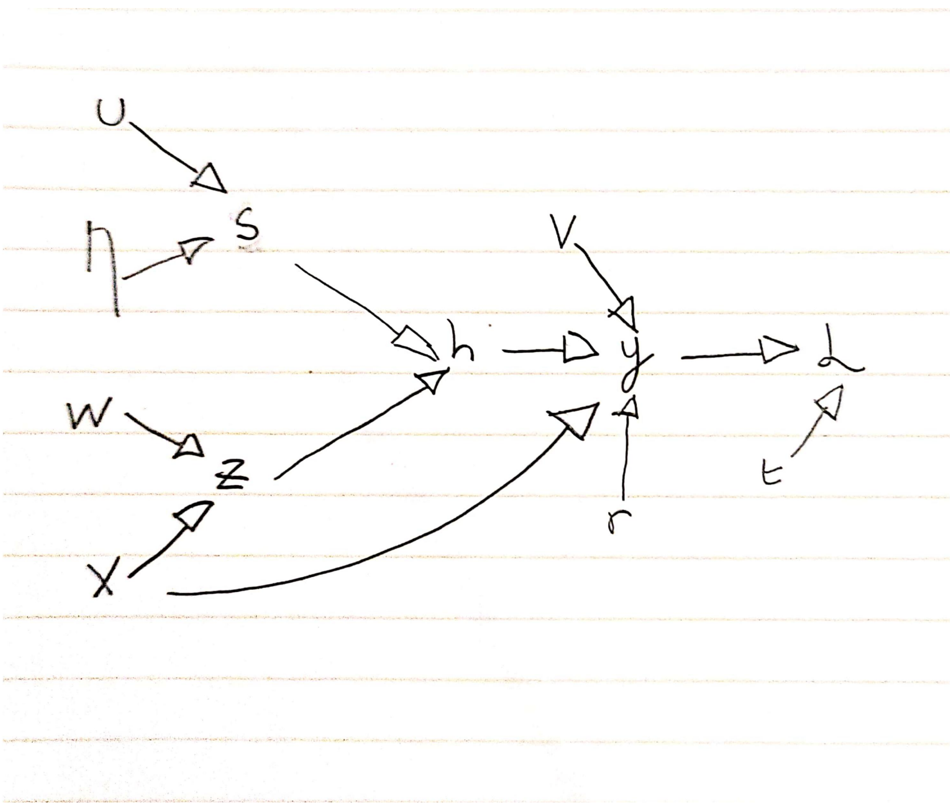
HW3 Writeup

Shiven Taneja 1005871013

November 2021

1 Backprop

1.a



1.b

Backprop Formulas:

$\overline{\mathcal{L}} = 1$

$\overline{y} = \overline{\mathcal{L}} \frac{\partial \mathcal{L}}{\partial y}$

$= \overline{\mathcal{L}}(y - t)$

$\overline{\mathbf{h}} = \overline{y} \frac{\partial y}{\partial \mathbf{h}}$

$= \overline{y} \mathbf{v}^\top$

$\overline{\mathbf{s}} = \overline{\mathbf{h}} \frac{\partial \mathbf{h}}{\partial \mathbf{s}}$

$= \overline{\mathbf{h}} \mathbf{z} \odot \sigma'(\mathbf{s})$

$\overline{\mathbf{U}} = \overline{\mathbf{s}} \frac{\partial \mathbf{s}}{\partial \mathbf{U}}$

$= \overline{\mathbf{s}} \eta^\top$

$\overline{\eta} = \overline{\mathbf{s}} \frac{\partial \mathbf{s}}{\partial \eta}$

$= \overline{\mathbf{s}} U^\top$

$\overline{\mathbf{z}} = \overline{\mathbf{h}} \frac{\partial \mathbf{h}}{\partial \mathbf{z}}$

$= \overline{\mathbf{h}} \sigma(\mathbf{s})$

$\overline{\mathbf{W}} = \overline{\mathbf{z}} \frac{\partial \mathbf{z}}{\partial \mathbf{W}}$

$= \overline{\mathbf{z}} \mathbf{x}^\top$

$\overline{\mathbf{x}} = \overline{\mathbf{z}} \frac{\partial \mathbf{z}}{\partial \mathbf{x}} + \overline{y} \frac{\partial y}{\partial \mathbf{x}}$

$= \overline{\mathbf{z}} \mathbf{W}^\top + \overline{y} \mathbf{r}^\top$

2 Fitting a Naive Bayes Model

2.a

For $\hat{\theta_{jc}}$:
We start by finding the log likelihood:

$$\ell(\theta) = log(p(x^i, c^i|\theta, \pi)) = \sum_{i=1}^{N} log(p(c^i|\pi) * \Pi_{j=1}^{784} p(x_j^i|c^i, \theta_{jc}))$$

$$= \sum_{i=1}^{N}(log(\pi_{c^i}) + log(\Pi_{j=1}^{784}(\theta_{jc^i}^{x_j^i}(1-\theta_{jc})^{1-x_j^i}))$$

$$= \sum_{i=1}^{N}(log(\pi_{c^i}) + \sum_{j=1}^{784} log(\theta_{jc^i}^{x_j^i}(1-\theta_{jc})^{1-x_j^i}))$$

$$= \sum_{i=1}^{N}(log(\pi_{c^i}) + \sum_{j=1}^{784}(x_j^i log(\theta_{jc^i}) + (1-x_j)log(1-\theta_{jc^i}))$$

Taking the derivative of θ_{jc} and setting to 0:

$$\frac{\partial \ell}{\partial \theta_{jc}} = \sum_{i=1}^{N}(\frac{x_j^i}{\theta_{jc^i}} - \frac{(1-x_j^i)}{(1-\theta_{jc^i})}) = 0$$

$$= \sum_{i=1}^{N}(x_j^i - x_j^i\theta_{jc^i} - \theta_{jc^i} + x_j^i\theta_{jc^i}) = 0$$

$$= \sum_{i=1}^{N}(x_j^i - \theta_{jc^i}) = 0$$

$= \sum_{i=1}^{N}(x_j^i - \theta_{jc^i}) * 1\mathrm{I}[t_c^i = c] = 0$, where c is a number from 0-9

$$\theta_{jc^i} = \frac{\sum_{i=1}^{N} x_j^i 1\mathrm{I}[t_c^i=c]}{\sum_{i=1}^{N} 1\mathrm{I}[t_c^i=c]}$$

This is the number of data points that are the digit 'c' and have the j^{th} pixel over the number of data points that are the digit 'c'

For $\hat{\pi}$:
We start by finding the log likelihood:

$$\ell(\pi) = log(p(t^i|\pi_j)) = \sum_{i=1}^{N} log(\Pi_{j=0}^{9}\pi_j^{t_j^i})$$

$$= \sum_{i=1}^{N}\sum_{j=0}^{9} t_j^i log(\pi_j)$$

$$= \sum_{i=1}^{N}(\sum_{j=0}^{8} t_j^i log(\pi_j) + t_9^i log(1 - \sum_{j=0}^{8} t_j^i))$$

Taking the partial derivative of π and setting it to 0:

$$\frac{\partial \ell}{\partial \pi} = \sum_{i=1}^{N}(\frac{t_j^i}{\pi_j} - \frac{t_9^i}{1-\sum_{j=0}^{8} t_j^i}) = \sum_{i=1}^{N}(\frac{t_j^i}{\pi_j} - \frac{t_9^i}{\pi_9}) = 0$$

$$= \sum_{i=1}^{N}(\frac{t_j^i}{\pi_j}) = \sum_{i=1}^{N}(\frac{t_9^i}{\pi_9})$$

$$= \sum_{i=1}^{N}(\frac{t_j^i}{t_9^i}) = \frac{\pi_j}{\pi_9}$$

We know that $\hat{\pi_j}$ sums to 1 for $j = 1, ..., 9$ so for $j = 1, ..., 8$, $\hat{\pi_j} = 1 - \hat{\pi_9}$:

$\sum_{i=1}^{N}(\frac{t_j^i}{t_9^i}) = \frac{1-\pi_9}{\pi_9} = \frac{1}{\pi_9} - 1$

$\sum_{i=1}^{N}(\frac{t_j^i}{t_9^i}) + 1 = \frac{1}{\pi_9}$

$\sum_{i=1}^{N}(\frac{t_j^i + t_9^i}{t_9^i}) = \frac{1}{\pi_9}$

$\sum_{i=1}^{N}(\frac{t_9^i}{t_j^i}) = \pi_9$

We know that $\hat{\pi_j} = 1 - \hat{\pi_9}$, so:

$\hat{\pi_j} = 1 - \sum_{i=1}^{N}(\frac{t_9^i}{t_j^i})$

$\hat{\pi_j} = \sum_{i=1}^{N}(\frac{t_j^i - t_9^i}{t_j^i})$

$\hat{\pi_j} = \sum_{i=1}^{N}(\frac{\sum_{j=0}^{8} t_j^i}{\sum_{j=0}^{9} t_j^i})$

Since we chose $j = 8$ randomly, we could get the same result for any value of j and thus we get:
$\hat{\pi_j} = \frac{1}{N} * \sum_{i=1}^{N} \sum_{j=0}^{9} t_j^i$
This is each class in the dataset divided by N

2.b

$$\log(P(t \mid x, \theta, \pi)) = \log\left(\frac{P(c \mid x, \theta)\, P(x \mid c, \theta, \pi)}{\sum_{i=1}^{9} P(c_i \mid \pi)\, P(x \mid c_i, \theta, \pi)}\right)$$

[Where $t \in \{0,1\}^{10}$ is the one-hot encoding for $c \in \{0, \dots, 9\}$]

$$= \log\left(\frac{P(c \mid \pi) \prod_{j=1}^{784} p(x_j \mid c, \theta, \pi)}{\sum_{i=1}^{9} P(c_i \mid \pi) \prod_{j=1}^{784} P(x_j \mid c_i, \theta, \pi)}\right) = \log\left(\frac{\pi_c \prod_{j=1}^{784} \theta_{jc}^{x_j} (1-\theta_{jc})^{(1-x_j)}}{\sum_{i=1}^{9} \pi_{c_i} \prod_{j=1}^{784} \theta_{jc_i}^{x_j} (1-\theta_{jc_i})^{(1-x_j)}}\right)$$

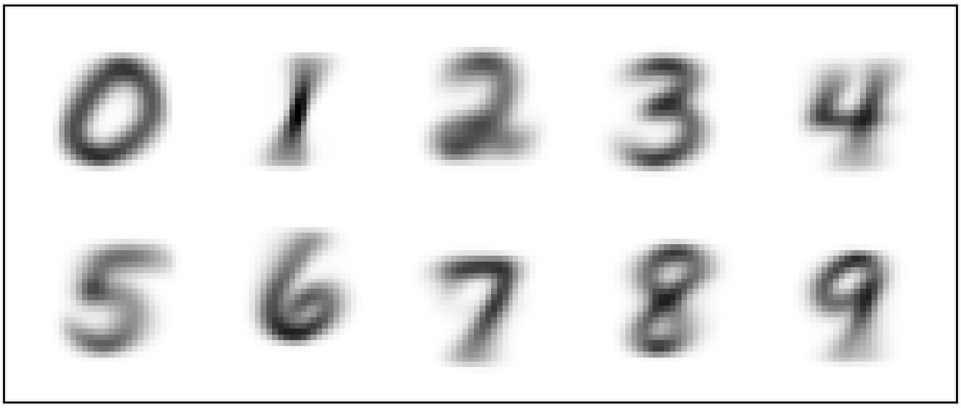
$$= \log(\pi_c) \cdot \sum_{j=1}^{784}\left(x_j \log(\theta_{jc}) + (1-x_j) \log(1-\theta_{jc})\right) - \sum_{i=1}^{9}\left(\log(\pi_{c_i}) + \sum_{j=1}^{784}\left(x_j \log(\theta_{jc_i}) + (1-x_j)\log(1-\theta_{jc_i})\right)\right)$$

$$= \log(\pi_c) \cdot \sum_{j=1}^{784}\left(x_j \log(\theta_{jc}) + (1-x_j) \log(1-\theta_{jc})\right) - \log \sum_{i=1}^{9} \exp\left(\log(\pi_{c_i}) + \sum_{j=1}^{784}\left(x_j \log(\theta_{jc_i}) + (1-x_j)\log(1-\theta_{jc_i})\right)\right)$$

2.c

The average log-likelihood for MLE is nan. The reason for this may be due to the zero values in θ making the average close to 0.

2.d



2.e

We Know $\Theta \sim Beta(3,3)$

$l(\Theta) =$

$$\log(p(\Theta|x,c,\pi)) = \log(p(\Theta)) + \log(p(x,c|\Theta,\pi))$$

$$= \log(\Theta_{jc}^{3-1}\cdot(1-\Theta_{jc})^{3-1}) + \sum_{i=0}^{N}\log\left(p(c|\pi)\prod_{j=1}^{784} p(x_j^i|c,\Theta_{jc})\right)$$

$$= \log(\Theta_{jc}^{2}) + 2\cdot\log(1-\Theta_{jc}) + \sum_{i=0}^{N}\left(\log(\pi_c) + \sum_{j=1}^{784}\log\left(\Theta_{jc}^{x_j^i}(1-\Theta_{jc})^{(1-x_j^i)}\right)\right)$$

$$= 2\log(\Theta_{jc}) + 2\log(1-\Theta_{jc}) + \sum_{i=0}^{N}\left(\log(\pi_c) + \sum_{j=1}^{784}\left(x_j^i\log(\Theta_{jc}) + (1-x_j^i)\log(1-\Theta_{jc})\right)\right)$$

We can now take the derivative and set it to 0:

$$\frac{\partial l}{\partial \Theta_{jc}} = \frac{2}{\Theta_{jc}} - \frac{2}{1-\Theta_{jc}} + \sum_{i=1}^{N}\left(\frac{x_j^i}{\Theta_{jc}} - \frac{(1-x_j^i)}{(1-\Theta_{jc})}\right)$$

$$0 = 2-2\Theta_{jc}-2\Theta_{jc}+\sum_{i=1}^{N}(x_j^i-\Theta_{jc})$$

$$0 = 2-2\Theta_{jc}-2\Theta_{jc}+\sum_{i=1}^{N}\mathbb{1}[t_c^{(i)}=c]\cdot(x_j^i-\Theta_{jc})$$

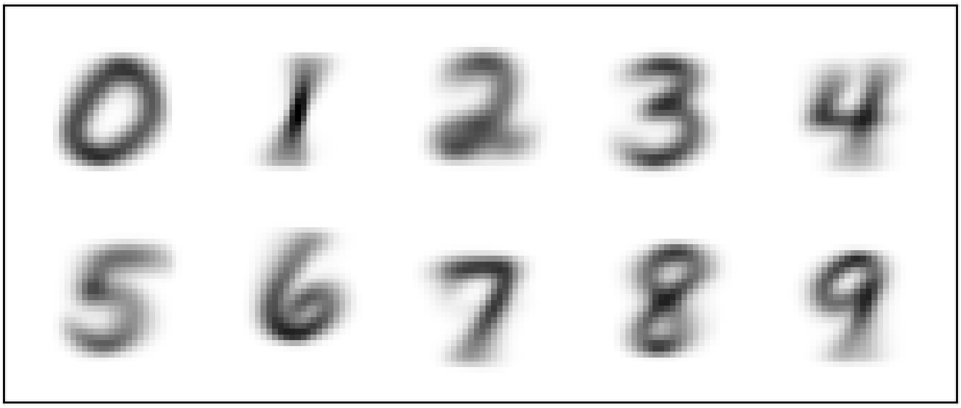
$$4\Theta_{jc}+\sum_{i=1}^{N}\mathbb{1}[t_c^{(i)}=c]\,\Theta_{jc} = 2+\sum_{i=1}^{N}\mathbb{1}[t_c^{(i)}=c]\,x_j^i$$

$$\hat{\Theta}_{jc} = \frac{2+\sum_{i=1}^{N}\mathbb{1}[t_c^{i}=c]\,x_j^i}{4+\sum_{i=1}^{N}\mathbb{1}[t_c^{i}=c]}$$

2.f

Average log-likelihood for MAP is -5.509798896626949
Training accuracy for MAP is 0.9383166666666667
Test accuracy for MAP is 0.9368

2.g



3 Categorial Distribution

3.a

Using Bayes Rule we Know:

$$P(\theta|D) = \frac{P(\theta)\,P(D|\theta)}{\int P(\theta')\,P(D|\theta)\,d\theta'}$$

Thus:

$$P(\theta|D) \propto P(\theta)P(D|\theta)$$

Where:

$$P(\theta) \propto \theta_1^{a_1-1} \cdots \theta_K^{a_K-1}$$

And since the data is independent and individually distributed:

$$P(X|\theta) = P(D|\theta) = \prod_{K=1}^{K} \theta_K^{x_K}$$

Therefore

$$P(\theta|D) \propto \prod_{K=1}^{K} \theta_K^{x_K^i}$$

The Dirichlet distribution is not a conjugate pair for the categorical distribution as when we combine the categorical and the Dirichlet prior, it does not have the same structure as a Dirichlet distribution

3.b

MAP Estimation

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(\theta|D)$$
$$= \arg\max_{\theta} p(\theta)p(D|\theta)$$
$$= \arg\max_{\theta} \log(p(\theta)) + \log(p(D|\theta))$$

$$\log(P(\theta|D)) = \log(p(\theta)) + \log(p(D|\theta))$$
$$= (a_k - 1)\log(\theta_k) + \sum_{k=1}^{K}(x_k^i \log(\theta_k))$$
$$= (a_k - 1)\log(\theta_k) + \sum_{k=1}^{K-1}(x_k^i \log(\theta_k) + x_K^i \log(1 - \sum_{k=1}^{K-1}\theta_k)$$

Taking the partial derivative and setting to 0:

$$\frac{\partial l}{\partial \theta} = \frac{a_k - 1}{\theta_k} + \frac{x_k^i}{\theta_k} - \frac{x_K^i}{\theta_K} = 0$$
$$= \frac{a_k - 1 + x_k^i}{\theta_k} - \frac{x_K^i}{\theta_K} = 0$$
$$= \theta_K(a_k - 1) + \theta_K x_k^i - \theta_k x_K^i = 0$$
$$= \frac{\theta_k}{\theta_K} = \frac{a_k - 1 + x_k^i}{x_K^i}$$
$$= \frac{1 - \theta_K}{\theta_K} = \frac{a_K - 1 + x_K^i}{x_K^i}$$

$$\frac{1}{\theta_K} - 1 = \frac{a_K - 1 + x_K^i}{x_K^i}$$

$$\frac{1}{\theta_K} = \frac{a_K - 1 + x_K^i}{x_K^i} + 1 = \frac{a_{K-1} + x_K^i + x_K^i}{x_K^i} = \frac{a_{K-1} + \sum_{k=1}^{K} x_K^i}{x_K^i}$$

$$\hat{\theta}_K = \frac{x_K^i}{a_{K-1} + x_K^i}$$

Since $k = K$ was chosen arbitrarily, we could do this for any valu of k and thus $\hat{\theta}_k$ would then be:

$$\hat{\theta}_k = \frac{a_k - 1 + x_k^i}{N}$$

3.c

4 Gaussian Discriminant Analysis

4.a

Average conditional log likelihood of training set: -0.12462443666863293
Average conditional log likelihood of test set: -0.19667320325525828

4.b

Test accuracy for the training set: 0.9814285714285714
Test accuracy for the test set: 0.97275

4.c