CSC311 HW3 Write-up

Jaffa Romain (student #1005243407)

04/11/2020

Question 1

QUESTION 1

a) $t^{(i)} \in \{1,-1\}$

$\bullet\ h_i \leftarrow \arg\min \sum_{i=1}^{n} w_i \mathbb{I}\{h(x^{(i)}) \neq t^{(i)}\}$ $\quad\bullet\ \alpha_t = \frac{1}{2}\log \frac{1-\text{error}_t}{\text{err}_t}$ $\quad\bullet\ w_i' \leftarrow w_i \exp(-\alpha_t t^{(i)} h_t(x^{(i)}))$

W.T.S $\quad \text{err}'_t = \frac{1}{2}$

NOTE:

$E^c = \{i : h_t(x^{(i)}) = t^{(i)}\} = 0 \quad \Rightarrow \quad E^c : h_t(x_i)t_i = (1)(1) \text{ or } (-1)(-1) = 1$

$E = \{i : h_t(x^{(i)}) \neq t^{(i)}\} = 1 \qquad\quad E : h_t(x_i)t_i = (-1)(1) = -1$

$$\text{err}_t = \frac{\sum_{i \in E} w_i}{\sum_{i=1}^{n} w_i}$$

$$\text{err}'_t = \frac{\sum_{i=1}^{n} w_i' \mathbb{I}\{h_t(x^{(i)}) \neq t^{(i)}\}}{\sum_{i=1}^{n} w_i'} = \frac{\sum_{i \in E} w_i \exp(-\alpha_t(-1))}{\sum_{i \in E} w_i \exp(-\alpha_t(-1)) + \sum_{i \in E^c} w_i \exp(-\alpha_t(1))} = \frac{\sum_{i \in E} w_i \exp(\alpha_t)}{\sum_{i \in E} w_i \exp(\alpha_t) + \sum_{i \in E^c} w_i \exp(-\alpha_t)}$$

$$= \frac{\sum w_i \exp(\alpha_t)}{\sum w_i \exp(\alpha_t) + \sum w_i(-\alpha_t)} = \frac{\sum w_i \exp(\alpha_t)}{\sum w_i \exp(\alpha t) + \sum_{i \in E^c} w_i} = \frac{\sum_{i \in E} w_i}{\sum_{i \in E} \frac{\exp(\alpha t)}{\exp(\alpha t)} + \sum_{i \in E^c} w_i \frac{\exp(-\alpha_t)}{\exp(\alpha t)}} = \frac{\sum_{i \in E} w_i}{\sum_{i \in E} w_i + \sum_{i \in E^c} w_i \exp(-2\alpha)}$$

$$\frac{\sum_{i \in E} w_i}{\sum_{i=1}^{N} w_i} = \text{err}t = \frac{\sum_{i \in E} w_i}{\sum_{i \in E} w_i + \sum_{i \in E^c} w_i} = \text{err}t$$

$$\therefore \sum_{i \in E} w_i = \text{err}t \cdot \left(\sum_{i \in E} w_i + \sum_{i \in E^c} w_i\right)$$

$$\therefore \frac{\sum_{i \in E} w_i}{\text{err}t} = \sum_{i \in E} w_i + \sum_{i \in E^c} w_i$$

$$\therefore \frac{\sum_{i \in E} w_i}{\text{err}_t} - \sum_{i \in E} w_i = \sum_{i \in E^c} w_i$$

USING THIS:

$$\frac{\sum w_i \exp(\alpha t)}{\sum w_i \exp(\alpha t) + \sum_{i \in E^c} w_i} = \frac{\sum_{i \in E} w_i}{\sum_{i \in E} w_i + \exp(-2\alpha) \cdot \left(\frac{\sum_{i \in E} w_i}{\text{err}t} - \sum_{i \in E} w_i\right)} = \frac{\cancel{\sum w_i}}{\sum_{i \in E} \cancel{w_i} + \left(\frac{\text{err}t}{1-\text{err}t}\right) \frac{\cancel{\sum_{i \in E} w_i}}{\text{err}t} - \left(\frac{\text{err}t}{1-\text{err}t}\right) \cancel{\sum_{i \in E} w_i}}$$

$$= \frac{1}{1 + \left(\frac{\text{err}t}{1-\text{err}t}\right) \cdot \frac{1}{\text{err}t} - \left(\frac{\text{err}t}{1-\text{err}t}\right)} = \frac{1}{1 + \frac{1}{1-\text{err}t} - \frac{\text{err}t}{1-\text{err}t}} = \frac{1}{\frac{1-\text{err}t + 1 - \text{err}t}{1-\text{err}t}}$$

$$= \frac{1}{\left(\frac{2(1-\cancel{\text{err}t})}{1-\cancel{\text{err}t}}\right)} = \boxed{\frac{1}{2}}$$

err'_t being $1/2$ means that for the $t+1$-th iteration, this will be the highest possible weighted error. This is the best error rate at t.

b)

b) 0-1 loss: $\mathbb{I}(h(x^{(n)}) \neq t^{(n)}] = \frac{1}{2}(1 - h(x^{(n)}) \cdot t^{(n)})$

W.T.S $w_i' \leftarrow w_i \exp(-\alpha_t t^{(i)} h_t(x^{(i)})) \propto w_i' \leftarrow w_i \exp(2\alpha_t \mathbb{I}\{h_t(x^{(i)}) \neq t^{(i)}\})$

$$w_i' \leftarrow w_i \exp(-\alpha_t t^{(i)} h_t(x^{(i)}))$$

$$= w_i \exp(-\alpha_t \{\mathbb{I} h_t(x^{(i)}) = t^{(i)}\}) + w_i \exp(-\alpha_t \{\mathbb{I} h_t(x^{(i)}) \neq t^{(i)}\})$$

$$= w_i \exp(-\alpha t \cdot 1) + w_i \exp(-\alpha t \cdot (-1))$$

$$= w_i \exp(-\alpha t) + w_i \exp(\alpha t)$$

$$= w_i \cdot \exp\left(-\frac{1}{2}\log\left(\frac{1-errt}{errt}\right)\right) + w_i \cdot \exp\left(\frac{1}{2}\log\left(\frac{1-errt}{errt}\right)\right)$$

$$= w_i\left(\frac{\sqrt{errt}}{\sqrt{1-errt}} + \frac{\sqrt{1-errt}}{\sqrt{errt}}\right) = w_i\left(\frac{\sqrt{errt}}{\sqrt{1-errt}} \cdot (1-errt) + \frac{\sqrt{1-errt}}{\sqrt{errt}} \cdot (errt)\right)$$

$$= w_i\left((\sqrt{errt})(\sqrt{1-errt}) + (\sqrt{1-errt})(\sqrt{errt})\right)$$

$$= w_i\left(\sqrt{errt - errt^2} + \sqrt{errt - errt^2}\right.$$

$$= w_i\left(2\sqrt{errt - errt^2}\right) = w_i \, 2\sqrt{errt(1-errt)}$$

The weight update is proportional to the constant factor $2\sqrt{err_t(1-errt}$.

Question 2

QUESTION 2

- $X^{(i)} \in \{0,1\}^{784}$

NAIVE BAYES

dimensions: 784x10

- $P(x,c|\theta,\pi) = p(c|\theta,\pi)\,p(x|c,\theta,\pi) = p(c|\pi)\prod_{j=1}^{784} p(x_j|c,\theta_{jc})$.
- $P(c|\pi) = \pi_c$
- $P(x_j = 1|c,\theta,\pi) = \theta_{jc}$
- LIKELIHOOD $(x_j \in \{0,1\})$: $\quad p(x_j|c,\theta_{jc}) = \theta_{jc}^{x_j}(1-\theta_{jc})^{(1-x_j)}$

$\quad : \; p(x_j = 1|c,\theta_{jc}) = \theta_{jc} \quad , \quad p(x_j = 0|c,\theta_{jc}) = 1-\theta_{jc}$

- $P(t_c = 1|\pi) = P(c|\pi) = \pi_c \qquad P(t|\pi) = \prod_{j=0}^{9}\pi_j^{t_j} \qquad \sum_{i=0}^{9}\pi_i = 1$

a) MLE FOR CLASS CONDITIONAL PROBABILITIES θ AND PRIOR π

① $\hat{\theta}_{MLE}$

$p(x,c|\theta,\pi) = p(c|\theta;\pi)\,p(x|c,\theta,\pi)$

$= p(c|\pi)\prod_{j=1}^{784} p(x_j|c,\theta_{jc})$

$L(\theta) = \prod_{i=1}^{n}\pi_c\prod_{j=1}^{784}\theta_{jc}^{x_j}(1-\theta_{jc})^{1-x_j}$

log-likelihood $\ell(\theta)$:

$$\ell(\theta) = \sum_{i=1}^{n}\Big[\log\pi_c + \sum_{j=1}^{784}(x_j\log\theta_{jc} + (1-x_j)\log(1-\theta_{jc}))\Big]$$

$\hat{\theta}_{MLE} = \max_{\theta\in[0,1]} \ell(\theta)$

$= \min_{\theta\in[0,1]} -\ell(\theta)$

$\therefore$ to maximize $\ell(\theta)$:

$$\frac{\partial\ell}{\partial\theta} = \frac{\partial}{\partial\theta}\sum_{i=1}^{n}\Big[\log\pi_c + \sum_{j=1}^{784}(x_j\log\theta_{jc} + (1-x_j)\log(1-\theta_{jc}))\Big]$$

$$= 0 + \sum_{i=1}^{n}\sum_{j=1}^{784}\frac{x_j}{\theta_{jc}} - \sum_{i=1}^{n}\sum_{j=1}^{784}\frac{(1-x_j)}{(1-\theta_{jc})}$$

$$0 = \sum_{i=1}^{n}\mathbb{I}(c^{(i)} = c)\,(x_j(1-\theta_{jc}) - (1-x_j)\theta_{jc})$$

$$0 = \sum_{i=1}^{n}\mathbb{I}(c^{(i)} = c)(x_j - \theta_{jc}x_j - \theta_{jc} + \theta_{jc}x_j)$$

$$0 = \sum_{i=1}^{n}\mathbb{I}(c^{(i)} = c)(x_j - \theta_{jc})$$

$$\sum_{i=1}^{n}\mathbb{I}(c^{(i)} = c)x_j = \sum_{i=1}^{n}(c^{(i)} = c)\theta_{jc}$$

$$\boxed{\hat{\theta}_{MLE} = \frac{\sum_i \mathbb{I}(c^{(i)} = c)\cdot x_j = 1)}{\sum_i \mathbb{I}(c^{(i)} = c)}}$$

for $j = 0, 1, \dots, 8$

$i = 1, \dots, 784$

② $\hat{\pi}_{MLE}$

$P(t^{(i)}|\pi) = \prod_{j=1}^{9}\pi_j^{t_j} \quad , \quad \sum_{i=0}^{9}\varepsilon_i = 1$

FINDING LOG LIKELIHOOD:

$$L(\theta) = \sum_{i=1}^{n}\log\prod_{j=0}^{9}\pi_j^{t_j}$$

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

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

TAKING THE DERIVATIVE OF $\ell(\theta)$ AND SETTING π TO 0:

$$0 = \sum_{i=1}^{n}\Big[\sum_{j=0}^{8}\frac{t_j^{(i)}}{\pi_j} - \sum_{j=0}^{8}\frac{t_9^{(i)}}{1-\sum_{j=0}^{8}\pi_j}\Big] \qquad 1-\sum_{j=0}^{8}\pi_j = \pi_9$$

$$0 = \sum_{i}^{n}\Big[\sum_{j=0}^{8}\frac{t_j^{(i)}}{\pi_j} - \sum_{j=0}^{8}\frac{t_9^{(i)}}{\pi_9}\Big]$$

$$\sum_{i=1}^{n}\sum_{j=0}^{8}\frac{t_j^{(i)}}{\pi_j} = \sum_{i=1}^{n}\sum_{j=0}^{8}\frac{t_9^{(i)}}{\pi_9}$$

$$\sum_{i=1}^{n}\sum_{j=0}^{8}\frac{\pi_j}{\pi_9} = \sum_{i=1}^{n}\sum_{j=0}^{8}\frac{t_j^{(i)}}{t_9^{(i)}}$$

$$\hat{\pi}_j = \hat{\pi}_0 + \hat{\pi}_1 + \dots + \hat{\pi}_8 = 1$$

$$\therefore \frac{\hat{\pi}_1}{\hat{\pi}_9} + \dots + \frac{\hat{\pi}_9}{\hat{\pi}_9} = \frac{1}{\pi_9}$$

$$\therefore \hat{\pi}_9 = \frac{\sum_{i=1}^{n} t_9^{(i)}}{\sum_{j=0}^{9} t_j^{(i)}}$$

$$\therefore \hat{\pi}_j = \frac{\sum_{i=1}^{n} t_j^{(i)}}{\sum_{j=0}^{9} t_j^{(i)}} = \frac{\sum_{i=1}^{n} t_j^{(i)}}{N}$$

$$\boxed{\therefore \hat{\pi}_j = \frac{1}{N}\sum_{i=1}^{n} t_j^{(i)} \quad \text{for } j = 0, \dots, 8}$$

b) log-likelihood $p(t|X,\theta,\pi)$ for single training example:

$$p(t|X,\theta,\pi) = \frac{P(X,c|\theta,\pi)}{P(X|\theta,\pi)}$$

$$\Longrightarrow = \log\left(\frac{P(c|\pi)\prod_{j=1}^{784} p(x_j|c,\theta_{jc})}{\sum_{c=0}^{9}\pi_c\prod_{j=1}^{784}\theta_{jc}^{x_j}(1-\theta_{jc})^{1-x_j}}\right)$$

$$= \log\left(\frac{\pi_c\prod_{j=1}^{784}\theta_{jc}^{x_j}(1-\theta_{jc})^{(1-x_j)}}{\sum_{c=0}^{9}\pi_c\prod_{j=1}^{784}\theta_{jc}^{x_j}(1-\theta_{jc})^{1-x_j}}\right)$$

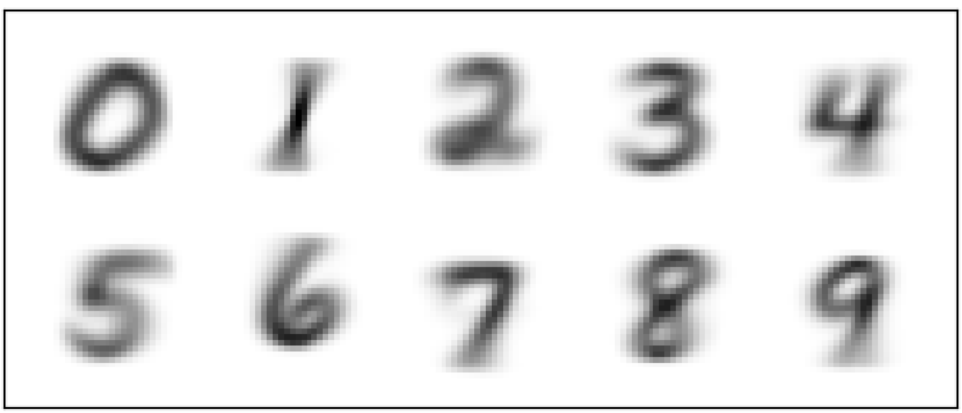
$$= \left(\log\pi_c + \sum_{j=1}^{784}\left(x_j\log\theta_{jc} + (1-x_j)\log(1-\theta_{jc})\right)\right) - \sum_{c=0}^{9}\left(\log(\pi_c)\sum_{j=1}^{784}\left(x_j\log(\theta_{jc}) + (1-x_j)\log(1-\theta_{jc})\right)\right)$$

c)

When trying to report the log-likelihood for a single training example, there is a runtime error, where log is divided zero in the function and average log-likelihood is being returned as nan. This could because of the number of zeros in theta. Since were our finding the log-likelihood of only one training example, there is only one entry that has 1 in the labels, and so log(0) is being calculated int the process, which could be causing the error.

d)

MLE estimator $\hat{\theta}$ as 10 separate greyscale images, one for each class.



e) MAP ESTIMATOR FOR θ USING $Beta(3,3)$ PRIOR ON EACH θ_{jc}.

$$P(\theta \mid x, c, \pi) \propto p(\theta)\, p(x, c \mid \theta, \pi)$$

$$\theta \sim Beta(3,3)$$

$$P(\theta \mid x, c, \pi) \propto p(\theta)\, p(x, c \mid \theta, \pi)$$

$$\ell(\theta) = \log\left[\theta^{3-1}(1-\theta)^{3-1} \cdot p(c \mid \pi) \prod_{j=1}^{784} p(x_j \mid c, \theta_{jc})\right]$$

$$= \log\left[\theta^2 (1-\theta)^2 \cdot \sum_{i=1}^{n}\left[\pi_c \prod_{j=1}^{784} \theta_{jc}^{x_j} (1-\theta)_{jc}^{1-x_j}\right]\right]$$

$$= 2\log\theta + 2\log(1-\theta) + \sum_{i=1}^{n} \log\left(\pi_c \prod_{j=1}^{784} \theta_{jc}^{x_j} (1-\theta_{jc})^{1-x_j}\right)$$

$$= 2\log\theta + 2\log(1-\theta) + \sum_{i=1}^{n}\left[\log\pi_c + \sum_{j=1}^{784} x_j \log(\theta_{jc}) + (1-\theta_{jc})\log(1-x_j)\right]$$

$$= 2\log\theta + 2\log(1-\theta) + \sum_{i=1}^{n}\left[\log\pi_c + \sum_{c=0}^{9}\sum_{j=1}^{784} (c^{(i)} = k)\left(\log(\theta_{jc}) + (1-\theta_{jc})\log(1-\theta_{cj})\right)\right]$$

$$\frac{\partial \ell}{\partial \theta} = \frac{2}{\theta_{jc}} - \frac{2}{1-\theta_{jc}} + \sum_{i=1}^{n} \mathbb{1}(c^{(i)} = c)\left(\frac{x_j}{\theta_{jc}} - \frac{1-x_j}{1-\theta_{jc}}\right)$$

$$0 = \frac{2}{\theta_{jc}} - \frac{2}{1-\theta_{jc}} + \sum_{i=1}^{n} \mathbb{1}(c^{(i)} = c)\frac{x_j}{\theta_{jc}} - \sum_{i=1}^{n} \mathbb{1}(c^{(i)} = c)\frac{1-x_j}{1-\theta_{jc}}$$

$$0 = 2 - 2\theta_{jc} - 2\theta_{jc} + \sum_{i=1}^{n} \mathbb{1}(c^{(i)} = c)(x_j - x_j\theta_{jc} - \theta_{jc} + x_j\theta_{jc})$$

$$0 = 2 - 4\theta_{jc} + \sum_{i=1}^{n} \mathbb{1}(c^{(i)} = c)(x_j) - \sum_{i=1}^{n} \mathbb{1}(c^{(i)} = c)\theta_{jc}$$

$$4\theta_{jc} + \sum_{i=1}^{n} \mathbb{1}(c^{(i)} = c)\theta_{jc} = 2 + \sum_{i=1}^{n} \mathbb{1}(c^{(i)} = c)x_j$$

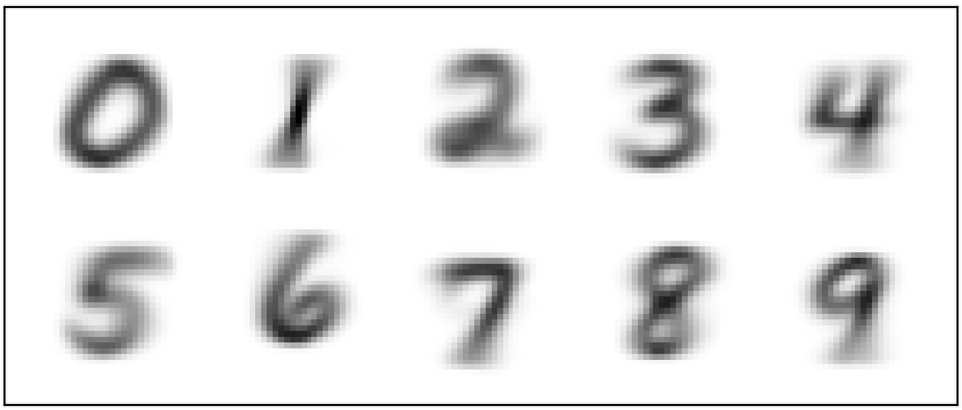
$$\hat{\theta}_{MAP} = \frac{2 + \sum_{i=1}^{n} \mathbb{1}(c^{(i)} = c)\, x_j}{4 + \sum_{i=1}^{n} \mathbb{1}(c^{(i)} = c)}$$

$$j = 1, \ldots, 784$$

$$i = 1, \ldots, n$$

f)

Average log-likelihood for MAP = -3.3570631378602847 Training accuracy for MAP = 0.8352166666666667
Test accuracy for MAP = 0.816



Question 3

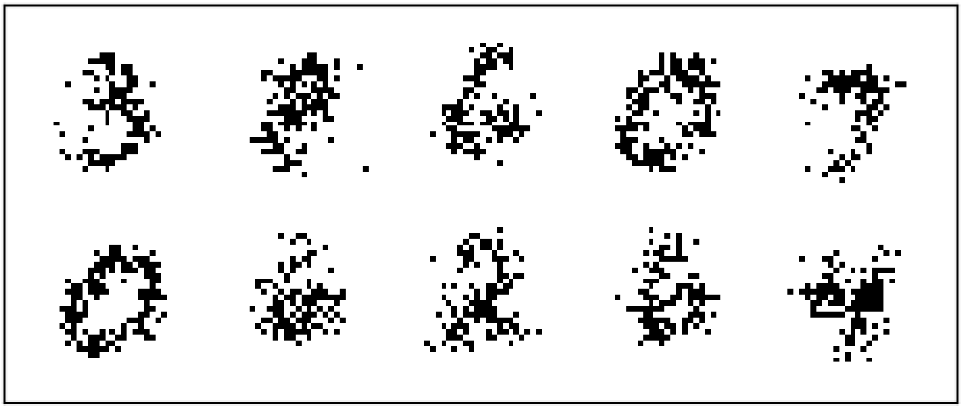
a) True. We assume x_i and x_j are independent on the condition of c.

b) False.

p(x_i, x_j) $= \Sigma p(x_i, x_j|c) = \Sigma p(x_i, x_j|c) = \Sigma p(x_i|c)p(x_j|c)$.

x_i and x_j are dependent when marginalized, $p(x_i, x_j) \neq p(x_i)(x_j)$.

c)



c: [3 8 6 0 7 0 6 2 6 4]