CSC311 Homework 3

Eric Zhu

2/11/2020

Contents

Question 1:

Part a:

Proof:

We want to prove that $err'_t = \frac{1}{2}$.

First we will rewrite err'_t as:

$$\begin{aligned}
err'_t &= \frac{\sum_{i=1}^{N} w'_i \mathbb{I}\{h_t(\mathbf{x}^{(i)}) \neq t\}}{\sum_{i=1}^{N} w'_i} \\
&= \frac{\sum_{i \in E} w'_i}{\sum_{i=1}^{N} w'_i} && \text{(By hint 2)} \\
&= \frac{\sum_{i \in E} w_i \exp(-\alpha_t t^{(i)} h_t(\mathbf{x}^{(i)}))}{\sum_{i=1}^{N} w'_i} && \text{(By the definition of } w'_i\text{)} \\
&= \frac{\sum_{i \in E} w_i \exp(-(0.5 \cdot log\frac{1-err_t}{err_t}) t^{(i)} h_t(\mathbf{x}^{(i)}))}{\sum_{i=1}^{N} w'_i} && \text{(By the definition of } \alpha_t\text{)}
\end{aligned}$$

From $\frac{\sum_{i \in E} w_i \exp(-(0.5 \cdot log\frac{1-err_t}{err_t}) t^{(i)} h_t(\mathbf{x}^{(i)}))}{\sum_{i=1}^{N} w'_i}$, we realize that $t^{(i)} h_t(\mathbf{x}^{(i)})$ is necessarily -1 for all indicies we consider in E because for all elements in E, we either have that $t^{(i)} = 1$ and $h_t(\mathbf{x}^{(i)}) = -1$ or $t^{(i)} = -1$ and $h_t(\mathbf{x}^{(i)}) = 1$ by definition of set E. To further explain, E is the set of indicies such that $\mathbb{I}\{h_t(\mathbf{x}^{(i)}) \neq t\} = 1$, so $h_t(\mathbf{x}^{(i)}) \neq t$, and both t and $h_t(\mathbf{x}^{(i)})$ are either $-1, 1$ by definition. We will call this **fact 1**. It follows that:

$$\frac{\sum_{i \in E} w_i \exp(-0.5 \cdot log\frac{1-err_t}{err_t} \cdot -1)}{\sum_{i=1}^{N} w'_i} \iff \frac{\sum_{i \in E} w_i \exp((0.5 \cdot log\frac{1-err_t}{err_t}) \cdot 1)}{\sum_{i \in E^c} w'_i + \sum_{i \in E} w'_i}$$

We rewrite $\sum_{i \in E^c} w'_i$ as:

$$\begin{aligned}
\sum_{i \in E^c} w'_i &= \sum_{i \in E^c} w_i \exp(-\alpha_t t^{(i)} h_t(\mathbf{x}^{(i)})) \\
&= \sum_{i \in E^c} \exp(-\alpha_t) w_i && \text{(By "Note 1" below)} \\
&= \exp(-\alpha_t) \cdot \sum_{i \in E^c} w_i \\
&= \exp(-\alpha_t) \cdot (\frac{\sum_{i \in E} w_i}{err_t} - \sum_{i \in E} w_i) && \text{(Rewriting hint 2)}
\end{aligned}$$

Note 1: We know that for all $i \in E^c$, we have that $t^{(i)} h_t(\mathbf{x}^{(i)}) = 1$ by the definition of E^c (recall that E^c is the set of indicies such that $\mathbb{I}\{h_t(\mathbf{x}^{(i)}) \neq t\} = 0$). In other words, E^c contains all the indicies where $h_t = t$.

Next, taking $\sum_{i \in E^c} w'_i = \exp(-\alpha_t) \cdot (\frac{\sum_{i \in E} w_i}{err_t} - \sum_{i \in E} w_i)$, we have that:

$$
\begin{aligned}
\frac{\sum_{i\in E} w_i \exp((0.5 \cdot log\frac{1-err_t}{err_t}))}{\sum_{i\in E^c} w'_i + \sum_{i\in E} w'_i} &= \frac{\sum_{i\in E} w_i \exp((0.5 \cdot log\frac{1-err_t}{err_t}))}{\exp(-\alpha_t)\cdot(\frac{\sum_{i\in E} w_i}{err_t} - \sum_{i\in E} w_i) + \sum_{i\in E} w'_i} \\
&= \frac{\sum_{i\in E} w_i \exp(\alpha_t)}{\exp(-\alpha_t)\cdot(\frac{\sum_{i\in E} w_i}{err_t} - \sum_{i\in E} w_i) + \sum_{i\in E} w'_i} \\
&= \frac{\sum_{i\in E} w_i \exp(2\cdot\alpha_t)}{\frac{\sum_{i\in E} w_i}{err_t} - \sum_{i\in E} w_i + \exp(\alpha_t)\sum_{i\in E} w_i \exp(-\alpha_t t^{(i)} h_t(\mathbf{x}^{(i)}))} \\
&= \frac{\sum_{i\in E} w_i \exp(2\cdot\alpha_t)}{\frac{\sum_{i\in E} w_i}{err_t} - \sum_{i\in E} w_i + \exp(\alpha_t)\sum_{i\in E} w_i \exp(\alpha_t \cdot 1)} && \text{(By \textbf{fact 1})} \\
&= \frac{\sum_{i\in E} w_i \exp(2\cdot\alpha_t)}{\frac{\sum_{i\in E} w_i}{err_t} - \sum_{i\in E} w_i + \exp(\alpha_t)\sum_{i\in E} w_i \exp(\alpha_t)} \\
&= \frac{\exp(2\cdot\alpha_t)}{\frac{1}{err_t} - \frac{err_t}{err_t} + \exp(2\cdot\alpha_t)} && \text{(Factorization of } \sum_{i\in E} w_i\text{)} \\
&= \frac{\exp(2\cdot\alpha_t)}{\exp(2\cdot\alpha_t) + \exp(2\cdot\alpha_t)} && \text{(By the definition of } \alpha_t\text{)} \\
&= \frac{\exp(2\cdot\alpha_t)}{2\exp(2\cdot\alpha_t)} \\
&= \frac{1}{2}
\end{aligned}
$$

We have therefore shown what we wanted to prove.

■

Part b:

Proof:

To begin, note that t, α_t is some fixed quantity from the problem, along with $\mathbf{x}^{(i)}$, which is some training example, and $t^{(i)}$ is the associated target.

We want to prove that the two expressions for w'_i are proportional up to a constant factor, i.e., we wish to prove the statement that $\exists k \in \mathbb{R}, k \cdot w_i \cdot exp(-\alpha_t t^{(i)} h_t(\mathbf{x}^{(i)})) = w_i \cdot exp(2\alpha_t \mathbb{I}\{h_t(\mathbf{x}^{(i)})) \neq t^{(i)}\})$.

Thus, let $k = e^{\alpha_t}$. It follows that $k \cdot w_i \cdot exp(-\alpha_t t^{(i)} h_t(\mathbf{x}^{(i)}))$ is exactly $w_i \cdot e^{\alpha_t - \alpha_t t^{(i)} h_t(\mathbf{x}^{(i)})} \iff w_i \cdot exp\{\alpha_t - \alpha_t t^{(i)} h_t(\mathbf{x}^{(i)})\}$.

Note that:

$$
w_i \cdot exp\{\alpha_t - \alpha_t t^{(i)} h_t(\mathbf{x}^{(i)})\} = w_i \cdot exp\{\alpha_t(1 - t^{(i)} h_t(\mathbf{x}^{(i)}))\} \iff w_i \cdot exp\{2\alpha_t(\frac{1}{2}(1 - t^{(i)} h_t(\mathbf{x}^{(i)}))\}
$$

With the rewritten expression, $w_i \cdot exp\{2\alpha_t(\frac{1}{2}(1 - t^{(i)} h_t(\mathbf{x}^{(i)}))\}$, we notice that $\frac{1}{2}(1 - t^{(i)} h_t(\mathbf{x}^{(i)}))$ is equivalent to the 0-1 loss ($\mathbb{I}\{h_t(\mathbf{x}^{(i)}) \neq t^{(i)}\}$). Substituting in $\mathbb{I}\{h_t(\mathbf{x}^{(i)})) \neq t^{(i)}\}$, we arrive at $w_i \cdot exp(2\alpha_t \mathbb{I}\{h_t(\mathbf{x}^{(i)})) \neq t^{(i)}\})$, which is exactly what we wanted to show.

■

Question 2:

Part a:

To begin, we will define our log-likelihood function using the joint probability distribution given in the problem:

$$
\begin{aligned}
l(\theta) = \sum_{i=1}^{N} log(p(\mathbf{x}, c|\theta, \pi)) &= \sum_{i=1}^{N} (log(p(c|\pi)) + log(\prod_{j=1}^{784} p(x_j^{(i)}|c, \theta_{jc}))) \\
&= \sum_{i=1}^{N} (log(p(c|\pi))) + \sum_{i=1}^{N} (log(\prod_{j=1}^{784} p(x_j^{(i)}|c, \theta_{jc}))) \\
&= \sum_{i=1}^{N} (log(p(c|\pi))) + \sum_{i=1}^{N} (\sum_{j=1}^{784} log(p(x_j^{(i)}|c, \theta_{jc}))) \\
&= \sum_{i=1}^{N} (log(p(c|\pi))) + \sum_{i=1}^{N} (\sum_{j=1}^{784} log(\theta_{jc}^{x_j^{(i)}} (1-\theta_{jc})^{1-x_j^{(i)}})) \\
&= \sum_{i=1}^{N} (log(p(c|\pi))) + \sum_{i=1}^{N} (\sum_{j=1}^{784} log(\theta_{jc}^{x_j^{(i)}}) + log((1-\theta_{jc})^{1-x_j^{(i)}}))
\end{aligned}
$$

Now with our log-likelihood, $l(\theta)$ (we want to find the likelihoods of both θ, π separately), we need to derive $\hat{\theta_{jc}}$ and $\hat{\pi_j}$. We will first derive $\hat{\theta_{jc}}$ on the next page.

Finding $\hat{\theta_{jc}}$:

We begin with the concept of MLE in this context, i.e., we wish to find $\frac{\partial l}{\partial \theta_{jc}} = 0$, which will be our maximum. Recall from above that $l(\theta) = \sum_{i=1}^{N}(log(p(c|\pi))) + \sum_{i=1}^{N}(\sum_{j=1}^{784} log(\theta_{jc}^{x_j}) + log((1-\theta_{jc})^{1-x_j}))$. It follows that our partial derivative is:

$$\begin{aligned}
\frac{\partial l}{\partial \theta_{jc}} l(\theta) &= \frac{\partial l}{\partial \theta_{jc}}(\sum_{i=1}^{N}(log(p(c|\pi))) + \sum_{i=1}^{N}(\sum_{j=1}^{784} log(\theta_{jc}^{x_j^{(i)}}) + log((1-\theta_{jc})^{1-x_j^{(i)}}))) = 0 \\
&= \frac{\partial l}{\partial \theta_{jc}}(\sum_{i=1}^{N}(log(p(c|\pi))) + \sum_{i=1}^{N}(log(\theta_{1c}^{x_1^{(i)}}) + log((1-\theta_{1c})^{1-x_1^{(i)}})) \\
&+ log(\theta_{2c}^{x_2^{(i)}}) + log((1-\theta_{2c})^{2-x_2^{(i)}})) \\
&+ log(\theta_{3c}^{x_3^{(i)}}) + log((1-\theta_{3c})^{1-x_3^{(i)}})) \\
&+ ... + log(\theta_{jc}^{x_j^{(i)}}) + log((1-\theta_{jc})^{1-x_j^{(i)}})) \\
&+ ... + log(\theta_{784c}^{x_{784}^{(i)}}) + log((1-\theta_{784c})^{1-x_{784}^{(i)}}))) = 0 \\
&= 0 + \sum_{i=1}^{N} x^{(i)} log(\theta_{jc}) + (1-x_j^{(i)}) log(1-\theta_{jc}) = 0 \\
&= \sum_{i=1}^{N}(\frac{x_j^{(i)}}{\theta_{jc}} - \frac{1-x_j^{(i)}}{1-\theta_{jc}}) = 0 \\
&= \sum_{i=1}^{N} \mathbb{I}(c^{(i)} = t)(x_j^{(i)}(1-\theta_{jc}) - (\theta_{jc} - x_j^{(i)}\theta_{jc})) = 0 \qquad \text{(By **"Note 2"**)}
\end{aligned}$$

Thus, we have $\sum_{i=1}^{N} \mathbb{I}(c^{(i)} = t)(x_j^{(i)}(1-\theta_{jc}) - (\theta_{jc} - x_j^{(i)}\theta_{jc})) = 0$. Solving for θ_{jc} to find the MLE, we get:

$$\hat{\theta_{jc}} = \frac{\sum_{i=1}^{N} \mathbb{I}(c^{(i)} = t) x_j^{(i)}}{\sum_{i=1}^{N} \mathbb{I}(c^{(i)} = t)}$$

Note 2: We add the identity function because we wish to only consider the terms where the class label for the particular pixel is the correct target, i.e, $c^{(i)} = t$. Note that t is a label from 1-of-10 encoded class labels given by the handout.

Finding $\hat{\pi_j}$:

Again we wish to find $\frac{\partial l}{\partial \pi_j} = 0$. However, note that we are given $\sum_{j=0}^{9} \pi_j = 1$, which means we may consider this a constrained optimization problem. We therefore, wish to optimize $l(\theta) + \lambda \cdot \sum_{j=0}^{9} \pi_j$:

$$\frac{\partial(l(\theta) + \lambda \cdot \sum_{j=0}^{9} \pi_j)}{\partial \pi_j} = 0$$

Since the log-likelihood function is a linear combination of Bernoulli log-likelihood of labels and the Bernoulli log-likelihood for the features $\mathbf{x}$, we get that as a result of differentiation we are left with:

$$\frac{\partial(l(\theta) + \lambda \cdot \sum_{j=0}^{9} \pi_j)}{\partial \pi_j} = \frac{\partial(\sum_{i}^{n} log(p(c|\pi)) + \lambda \cdot \sum_{j=0}^{9} \pi_j)}{\partial \pi_j}$$

In other words, we have:

$$\frac{\partial(\sum_i^n log(p(c|\pi)) + \lambda \cdot \sum_{j=0}^9 \pi_j)}{\partial \pi_j} = \frac{\partial(\sum_i^n log(p(\mathbf{t}|\pi)) + \lambda \cdot \sum_{j=0}^9 \pi_j)}{\partial \pi_j}$$

Applying the logarithm rules:

$$\frac{\partial(\sum_i^n log(p(\mathbf{t}|\pi)) + \lambda \cdot \sum_{j=0}^9 \pi_j)}{\partial \pi_j} = \frac{\partial(\sum_i^n log(\prod_{j=1}^9 \pi_j^{t_j}) + \lambda \cdot \sum_{j=0}^9 \pi_j)}{\partial \pi_j}$$

$$\frac{\partial(\sum_i^n \sum_{j=1}^9 log(\pi_j^{t_j}) + \lambda \cdot \sum_{j=0}^9 \pi_j)}{\partial \pi_j}$$

Finally,

$$\frac{\partial(\sum_i^n \sum_{j=1}^9 log(\pi_j^{t_j}) + \lambda \cdot \sum_{j=0}^9 \pi_j)}{\partial \pi_j} = \sum_i^N \frac{t_j^{(i)}}{\pi_j} + \lambda = 0$$

Solving for π_j, we get $\pi_j = -\frac{\sum_{i=1}^N t_j^{(i)}}{\lambda}$. Since $t_j^{(i)}$ is only 1 if and only if image i belongs to target category t_j (and else 0), we can write:

$$\begin{aligned}
\sum_{j=0}^9 \pi_j &= \sum_{j=1}^9 (-\frac{\sum_{i=1}^N t_j^{(i)}}{\lambda}) \\
&= -\frac{\sum_{i=1}^N t_1^{(i)}}{\lambda} - \frac{\sum_{i=1}^N t_2^{(i)}}{\lambda} ... - \frac{\sum_{i=1}^N t_N^{(i)}}{\lambda} \\
&= \frac{-\sum_{i=1}^N t_1^{(i)} - \sum_{i=1}^N t_2^{(i)} ... - \sum_{i=1}^N t_N^{(i)}}{\lambda} \\
&= 1
\end{aligned}$$

Note that there are at most N images, and since each image has a label and $t_j^{(i)}$, we know $\frac{-\sum_{i=1}^N t_1^{(i)} - \sum_{i=1}^N t_2^{(i)} ... - \sum_{i=1}^N t_N^{(i)}}{\lambda} = \frac{-N}{\lambda}$. To further explain, we know that our numerator is exactly N since each image belongs to a class and as such each image will get assigned 1 exactly once.Adding up N 1's, we get N. Thus, we find that our lagrange multiplier is $-N$. Finally, we get that:

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

Part b:

To start we define our likelihood as $p(\mathbf{t}|\mathbf{x}, \theta, \pi)$, and so it follows that our log-likelihood is $log(p(\mathbf{t}|\mathbf{x}, \theta, \pi)) \iff log(\frac{p(c|\pi)p(\mathbf{x}|\mathbf{t}, \theta, \pi)}{\sum_{c=0}^9 p(c|\pi)p(\mathbf{x}|\mathbf{t}, \theta, \pi)})$.

We can use the equations provided in the homework to rewrite the log likelihood as:

$$log(\frac{p(c|\pi)p(\mathbf{x}|\mathbf{t}, \theta, \pi)}{\sum_{c=0}^9 p(c|\pi)p(\mathbf{x}|\mathbf{t}, \theta, \pi)}) \iff log(\frac{p(c|\pi)\prod_{j=1}^{784} \theta_{jc}^{x_j}(1-\theta_{jc})^{1-x_j})}{\sum_{c=0}^9 (p(c|\pi)\prod_{j=1}^{784} \theta_{jc}^{x_j}(1-\theta_{jc})^{1-x_j})})$$

Then we apply the log rules to obtain:

$$log(\frac{p(c|\pi)\prod_{j=1}^{784}\theta_{jc}^{x_j}(1-\theta_{jc})^{1-x_j})}{\sum_{c=0}^{9}(p(c|\pi)\prod_{j=1}^{784}\theta_{jc}^{x_j}(1-\theta_{jc})^{1-x_j})}) \iff log(p(c|\pi)\prod_{j=1}^{784}\theta_{jc}^{x_j}(1-\theta_{jc})^{1-x_j}))-log(\sum_{c=0}^{9}(p(c|\pi)\prod_{j=1}^{784}\theta_{jc}^{x_j}(1-\theta_{jc})^{1-x_j}))$$

Further simplifying we get:

$$log(p(c|\pi)\prod_{j=1}^{784}\theta_{jc}^{x_j}(1-\theta_{jc})^{1-x_j})) - log(\sum_{c=0}^{9}p(c|\pi)\prod_{j=1}^{784}\theta_{jc}^{x_j}(1-\theta_{jc})^{1-x_j}))$$

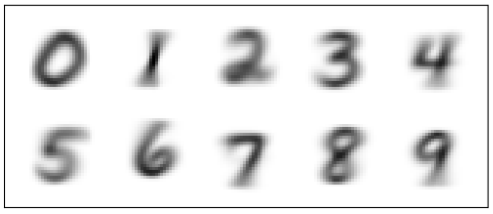
$$\iff$$

$$log(p(c|\pi)) + \sum_{j=1}^{784}(x_j log(\theta_{jc}) + (1-x_j)log((1-\theta_{jc}))) - log(\sum_{c=0}^{9}p(c|\pi)\prod_{j=1}^{784}\theta_{jc}^{x_j}(1-\theta_{jc})^{1-x_j}))$$

Part c:

Average log-likelihood for MLE is nan due to "divide by zero encountered in log `log_likelihood_cat = np.dot(image, np.log(theta)) + np.dot((1-image),np.log(1-theta))`". From this error, we see that we have division by zero due to some θ_{jc} being 0.

Part d:



Part e:

To start, we define our posterior distribution as $p(\theta|c,\pi,\mathbf{x})$, which is proportional to $p(\theta)p(x,c|\theta,\pi)$. Note that $p(x,c|\theta,\pi)$ is our likelihood function from **part a** and $p(\theta) \sim$ Beta(3,3). To derive the MAP estimator for θ, we need to solve $\frac{\partial p(\theta|c,\pi,\mathbf{x})}{\partial\theta} = \frac{\partial log(p(\theta|c,\pi,\mathbf{x}))}{\partial\theta} = 0$. We begin the derivation by defining the log-likelihood ($l(\theta)$):

Deriving the likelihood function $l(\theta)$:

$$\begin{aligned}
l(\theta) = log(p(\theta|c,\pi,\mathbf{x})) &= \sum_{i=1}^{N} log(p(\theta)(\sum_{i=1}^{N} p(\mathbf{x}^{(i)}, c|\theta,\pi))) \\
&= log(p(\theta)) + \sum_{i=1}^{N} log(p(\mathbf{x}^{(i)}, c|\theta,\pi)) \\
&= log(p(\theta) + \sum_{i=1}^{N} log(p(c|\pi) \prod_{j=1}^{784} \theta_{jc}^{x_j^{(i)}} (1-\theta_{jc})^{1-x_j^{(i)}}))) \\
&= log(\theta_{jc}^2(1-\theta_{jc})^2) && \text{(Beta PDF)} \\
&+ \sum_{i=1}^{N}(log(p(c|\pi)) + \sum_{j=1}^{784}(x_j log(\theta_{jc}) \\
&+ (1-x_j)log((1-\theta_{jc})))) \\
&= log(\theta_{jc}^2(1-\theta_{jc})^2) && \text{(Definition of } p(c|\pi)) \\
&+ \sum_{i=1}^{N}(log(\pi_c) + \sum_{j=1}^{784}(x_j log(\theta_{jc}) + (1-x_j)log((1-\theta_{jc}))))
\end{aligned}$$

Deriving the MAP estimator $\hat{\theta_{MAP}}$:

We begin by taking the derivative of $l(\theta)$:

$$\begin{aligned}
\frac{\partial l(\theta)}{\partial \theta} &= \frac{\partial}{\partial \theta_{jc}}(log(\theta_{jc}^2(1-\theta_{jc})^2) + \sum_{i=1}^{N}(log(\pi_c) + \sum_{j=1}^{784}(x_j log(\theta_{jc}) + (1-x_j)log(1-\theta_{jc})))) \\
&= \frac{\partial}{\partial \theta_{jc}}(\log(\theta_{jc}^2(1-\theta_{jc})^2)) + \sum_{i=1}^{N} \mathbb{I}(c^{(i)} = t)(\frac{x_j^{(i)}}{\theta_{jc}} - \frac{1-x_j^{(i)}}{1-\theta_{jc}}) && \text{(By part 2a and \textbf{"Note 3"})} \\
&= \frac{\partial}{\partial \theta_{jc}}((2log(\theta_{jc}) + 2log(1-\theta_{jc}))) + \sum_{i=1}^{N} \mathbb{I}(c^{(i)} = t)(\frac{x_j^{(i)}}{\theta_{jc}} - \frac{1-x_j^{(i)}}{1-\theta_{jc}}) \\
&= 2(\frac{1}{\theta_{jc}} - \frac{1}{(1-\theta_{jc})}) + \sum_{i=1}^{N} \mathbb{I}(c^{(i)} = t)(\frac{x_j^{(i)}}{\theta_{jc}} - \frac{1-x_j^{(i)}}{1-\theta_{jc}}) && \text{(By \textbf{Note 4})} \\
&= 2(\frac{(1-\theta_{jc})}{\theta_{jc}(1-\theta_{jc})} - \frac{\theta_{jc}}{(1-\theta_{jc})\theta_{jc})}) + \sum_{i=1}^{N} \mathbb{I}(c^{(i)} = t)(\frac{x_j^{(i)}}{\theta_{jc}} - \frac{1-x_j^{(i)}}{1-\theta_{jc}}) = 0 \\
&= 2(\frac{(1-\theta_{jc})}{\theta_{jc}(1-\theta_{jc})} - \frac{\theta_{jc}}{(1-\theta_{jc})\theta_{jc})}) + \sum_{i=1}^{N} \mathbb{I}(c^{(i)} = t)(\frac{x_j^{(i)}}{\theta_{jc}} - \frac{1-x_j^{(i)}}{1-\theta_{jc}}) = 0
\end{aligned}$$

Rearranging, we get this equality:

$$-2\frac{(1-2\theta_{jc})}{\theta_{jc}(1-\theta_{jc})} = \sum_{i=1}^{N} \mathbb{I}(c^{(i)} = t)(\frac{x_j^{(i)}}{\theta_{jc}} - \frac{1-x_j^{(i)}}{1-\theta_{jc}})$$

Cancelling out the denominators on both sides we get:

$$-2\frac{(1-2\theta_{jc})}{\theta_{jc}(1-\theta_{jc})} = \sum_{i=1}^{N}\mathbb{I}(c^{(i)}=t)(x_j^{(i)}(1-\theta_{jc}) - (\theta_{jc} - x_j^{(i)}\theta_{jc})) \iff (-2+4\theta_{jc}) = \sum_{i=1}^{N}\mathbb{I}(c^{(i)}=t)(x_j^{(i)} - \theta_{jc})$$

It follows that:

$$-2+4\theta_{jc} = \sum_{i=1}^{N}\mathbb{I}(c^{(i)}=t)(x_j^{(i)} - \theta_{jc}) \iff -2+4\theta_{jc} + \sum_{i=1}^{N}\mathbb{I}(c^{(i)}=t)(\theta_{jc}) = \sum_{i=1}^{N}\mathbb{I}(c^{(i)}=t)(x_j^{(i)})$$

Next,

$$-2+4\theta_{jc}+\sum_{i=1}^{N}\mathbb{I}(c^{(i)}=t)(\theta_{jc}) = \sum_{i=1}^{N}\mathbb{I}(c^{(i)}=t)(x_j^{(i)}) \iff -2+4\theta_{jc}+\theta_{jc}\sum_{i=1}^{N}\mathbb{I}(c^{(i)}=t) = \sum_{i=1}^{N}\mathbb{I}(c^{(i)}=t)(x_j^{(i)})$$

Now solving for θ_{jc} by moving -2 to the right hand side of $-2+4\theta_{jc}+\theta_{jc}\sum_{i=1}^{N}\mathbb{I}(c^{(i)}=t) = \sum_{i=1}^{N}\mathbb{I}(c^{(i)}=t)(x_j^{(i)})$ and factoring out θ_{jc}, we obtain the MAP estimator for θ_{jc} ($\hat{\theta_{MAP}}$):

$$\hat{\theta_{MAP}} = \theta_{jc} = \frac{\sum_{i=1}^{N}\mathbb{I}(c^{(i)}=t)(x_j^{(i)}) + 2}{\sum_{i=1}^{N}\mathbb{I}(c^{(i)}=t) + 4}$$

Note 3: Using the logic of **note 2** from part 2a, we add the identity function before factoring out the denominator $\sum_{i=1}^{N}\mathbb{I}(c^{(i)}=t)(\frac{x_j^{(i)}}{\theta_{jc}} - \frac{1-x_j^{(i)}}{1-\theta_{jc}})$ because unlike MLE we also have the derivative of the log of the prior to consider in this case ($\frac{\partial}{\partial\theta_{jc}}(log(\theta_{jc}^2(1-\theta_{jc})^2)$.

Note 4: We are able to obtain $2(\frac{1}{\theta_{jc}} - \frac{1}{(1-\theta_{jc})})$ from $\frac{\partial}{\partial\theta_{jc}}((2log(\theta_{jc}) + 2log(1-\theta_{jc})))$ because the partial derivative of $2log(1-\theta_{jc})$ is negative by the chain rule. I did not write it in the derivation for better notional clarity.

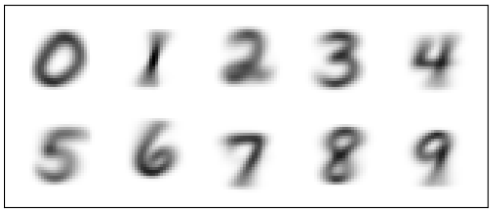
Part f:

Average log-likelihood for MAP is -3.3570631378601683

Training accuracy for MAP is 0.8352166666666667

Test accuracy for MAP is 0.816

Part g:



Question 3:

Part a:

True. By the assumptions of naive Bayes, we assume that the pixels x_i and x_j are conditionally independent. Here our condition is c.

Part b:

False. Having the pixels be marginally independent of class would mean that knowing the class value, e.g., 0 or 1, would have no effect on our probability of the pixels. Given that the context of the class is the digit and digits do indeed have unique shapes, the pixels should be marginally dependent.

Part c: