Statistics for Machine Learning

Problem 1:

Suppose that $\mathbf{X} \in \mathbb{R}^{n \times m}$ with $n \geq m$ and $\mathbf{t} \in \mathbb{R}^n$ and that $\mathbf{t}|(\mathbf{X}, \mathbf{w}) \sim \mathcal{N}(Xw, \sigma^2 \mathbf{I})$

(a) Show that the maximum likelihood estimate $\hat{w}$ of w is given by $\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t}$.

(b) Find the distribution of $\hat{\mathbf{w}}$, its expectation and covariance matrix.

(c) Now suppose we place a normal prior on $\mathbf{w}|\mathbf{X}$, i.e., $\mathbf{w} \sim \mathcal{N}(0, \tau^2\mathbf{I})$. Show that the MAP estimate of $\mathbf{w}$ is given by $\hat{\mathbf{w}}_{MAP} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{t}$ where $\lambda = \sigma^2/\tau^2$

Proof.

(a) We first note that since the components of $\mathbf{t}$ normally distributed and are a linear combination of $\mathbf{w}$, they are jointly normally distributed. Furthermore, since the covariance matrix is diagonal, each component is pairwise uncorrelated. This is sufficient to show that each component t_i is independent of t_j for $i \neq j$.

Let $\mathbf{X}_i$ be the i^{th} row of $\mathbf{X}$. We can treat t_i as one of n i.i.d. normal random vectors having mean $\mathbf{X}_i\mathbf{w}$ and variance σ^2. Thus, their joint density is:

$$\begin{aligned}
L(t_1, ..., t_n|\mathbf{X}\mathbf{w}, \sigma^2) &= \prod_{i=1}^{n} p(t_i|X_i w, \sigma^2) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(t_i - X_i w)^2}{2\sigma^2}\right) \\
&= \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^n} \exp\left(-\sum_{i=1}^{n} \frac{(t_i - X_i w)^2}{2\sigma^2}\right) \\
&= \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^n} \exp\left(-\frac{||\mathbf{t} - \mathbf{X}\mathbf{w}||^2}{2\sigma^2}\right)
\end{aligned}$$

Now, $\hat{\mathbf{w}}$ is defined as the value of $\mathbf{w}$ that maximises. To get that, we get the log-likelihood function from the equation above (which will give us the same value of $\hat{\mathbf{w}}$ by monotonicity of the logarithm):

$$\ln\left(\frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^n} \exp\left(-\frac{||\mathbf{t} - \mathbf{X}\mathbf{w}||^2}{2\sigma^2}\right)\right) = -\frac{n\ln(2\pi\sigma^2)}{2} - \frac{||\mathbf{t} - \mathbf{X}\mathbf{w}||^2}{2\sigma^2}$$

We know the maximum value occurs at a point where the derivative/gradient w.r.t $\mathbf{w}$ is 0. Thus, doing so we get:

$$
\begin{aligned}
0 &= \nabla_{\mathbf{w}} \left(-\frac{n \ln(2\pi\sigma^2)}{2} - \frac{||\mathbf{t} - \mathbf{X}\mathbf{w}||^2}{2\sigma^2} \right) \\
&= \frac{1}{2\sigma^2} \cdot \nabla_{\mathbf{w}} ||\mathbf{t} - \mathbf{X}\mathbf{w}||^2 \\
&= \frac{1}{2\sigma^2} \cdot \nabla_{\mathbf{w}} \left(||\mathbf{t}||^2 - 2\mathbf{t}^T\mathbf{X}\mathbf{w} + \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} \right) \\
&= \frac{-2}{2\sigma^2} (\mathbf{X}^T\mathbf{t} - \mathbf{X}^T\mathbf{X}\mathbf{w})
\end{aligned}
$$

If $\mathbf{X}^T\mathbf{X}$ is invertible, we can rearrange the above expression to get the $arg\, max$

$$\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t}$$

(b) From $\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t}$, we can see that $\hat{\mathbf{w}}$ is a linear transformation of of $\mathbf{t}$ which is normally distributed. Thus, $\hat{\mathbf{w}}$ is also normally distributed.

The expectation of $\hat{\mathbf{w}}$:

$$
\begin{aligned}
\mathbb{E}(\hat{\mathbf{w}}) &= \mathbb{E}((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t}) \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{E}(\mathbf{t}) \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\mathbf{w} && \text{(since } t \sim \mathcal{N}(\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I})) \\
&= \mathbf{w} && \text{(by cancellation)}
\end{aligned}
$$

Since each component of $\hat{\mathbf{w}}$ is independent from each other (based on the independence of t), the non-diagonal entries of $Cov(\hat{\mathbf{w}})$ are 0.

The diagonal entries, are defined as the individual variances of each component.

$$
\begin{aligned}
\mathrm{Cov}(\hat{\mathbf{w}})_{ii} &= \mathrm{Var}(\hat{\mathbf{w}}_i) \\
&= \mathrm{Var}((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t}_i) \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathrm{Var}(\mathbf{t}_i)((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T && (\mathrm{Var}(AY) = A\mathrm{Var}(\mathrm{Y})A^T) \\
&= \mathrm{Var}(\mathbf{t}_i)(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1} && ((X^{-1})^T = (X^T)^{-1}) \\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}_{ii}
\end{aligned}
$$

Thus, $\mathrm{Cov}(\hat{\mathbf{w}})_{ij} = \begin{cases} 0 & \text{for } i \neq j \\ \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}_{ii} & \text{for } i = j \end{cases}$

(c) Since $p(\mathbf{w}|\mathbf{X},\mathbf{t}) \propto p(\mathbf{t}|\mathbf{X},\mathbf{w})p(\mathbf{w}|\mathbf{X})$, we can take the logarithm of the right-hand side to find the $arg\, max$.

$$
\begin{aligned}
\ln(p(\mathbf{w}|\mathbf{X},\mathbf{t})) \propto \ln(p(\mathbf{t}|\mathbf{X},\mathbf{w}) \cdot p(\mathbf{w}|\mathbf{X})) \\
&= -\frac{n\ln(2\pi\tau^2\sigma^2)}{2} - \frac{||\mathbf{t}-\mathbf{X}\mathbf{w}||^2}{2\sigma^2} - \frac{||\mathbf{w}-0||^2}{2\tau^2} && \text{(calculation from part (a))} \\
&\propto -\frac{||\mathbf{t}-\mathbf{X}\mathbf{w}||^2}{2\sigma^2} - \frac{||\mathbf{w}||^2}{2\tau^2} && \text{(removing constant terms)}
\end{aligned}
$$

Taking the derivative with respect to $\mathbf{w}$ and setting it to zero:

$$
\begin{aligned}
0 &= \nabla_{\mathbf{w}}\left(-\frac{||\mathbf{t}-\mathbf{X}\mathbf{w}||^2}{2\sigma^2} - \frac{||\mathbf{w}||^2}{2\tau^2}\right) \\
&= \frac{2\mathbf{X}^T(\mathbf{t}-\mathbf{X}\mathbf{w})}{2\sigma^2} - \frac{2\mathbf{w}}{2\tau^2} && \text{Since } \frac{d||A||^2}{dA} = \frac{dA^TA}{dA} = 2A \\
&= \mathbf{X}^T\mathbf{t} - \mathbf{X}^T\mathbf{X}\mathbf{w} - \frac{\sigma^2}{\tau^2}\mathbf{w} && \text{(scale by } \sigma^2\text{)} \\
\Rightarrow \mathbf{X}^T\mathbf{t} &= \mathbf{w}(\mathbf{X}^T\mathbf{X} + \frac{\sigma^2}{\tau^2}\mathbf{I})
\end{aligned}
$$

Thus, if $(\mathbf{X}^T\mathbf{X} + \frac{\sigma^2}{\tau^2}\mathbf{I})$ is invertible, rearranging the above expression gets us

$$\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^t\mathbf{t}$$

where $\lambda = \dfrac{\sigma^2}{\tau^2}$ as desired. $\square$

Problem 2: Suppose you have a D-dimensional data vector $\mathbf{x}$ and an associated class variable $t \in \{0, 1\}$ which is a Bernoulli random variable. Assume that the dimensions of $\mathbf{x}$ are conditionally independent given t and that the conditional distribution of each x_i is Gaussian.

(a) Use Bayes' Rule to show that $p(t = 1|\mathbf{x})$ takes the form of the logistic function

$$\sigma(\mathbf{w}^T\mathbf{x} + b) = \frac{1}{1 + \exp - \sum_{i=1}^{D} w_i x_i - b}$$

(b) Suppose you have a training set $\mathcal{D} = \{(\mathbf{x}^{(1)}, t^{(1)}), ..., (\mathbf{x}^{(N)}, t^{(N)})\}$. Derive an expression for $L(\mathbf{w}, b)$, the negative log-likelihood under the i.i.d. assumption. Then derive expressions of the derivatives with respsect to the model parameters.

(c) Now treat the $\mathbf{x}^{(i)}$'s as deterministic and assume a Gaussian prior is placed on each element $\mathbf{w}$ such that $p(w_i) = \mathcal{N}(w_i|0, 1, \lambda)$ and a flat prior on b such that $p(b) = 1$. Show that the negative logarithm of this prosterior takes the form

$$L_{post}(\mathbf{w}, b) = L(\mathbf{w}, b) + \frac{\lambda}{2}\sum_{i=1}^{D} w_i^2 + C$$

Proof.

(a)

$$\begin{aligned}
p(t = 1|\mathbf{x}) &= \frac{p(\mathbf{x}|t = 1)p(t = 1)}{p(\mathbf{x})} && \text{(Bayes' Rule)} \\
&= \frac{p(\mathbf{x}|t = 1)p(t = 1)}{p(\mathbf{x}|t = 1)p(t = 1) + p(\mathbf{x}|t = 0)p(t = 0)} && \text{(Law of Total Probability)} \\
&= \frac{1}{1 + \dfrac{p(\mathbf{x}|t = 0)p(t = 0)}{p(\mathbf{x}|t = 1)p(t = 1)}} && \text{(Factor out numerator)}
\end{aligned}$$

Now we simplify $\dfrac{p(\mathbf{x}|t = 0)p(t = 0)}{p(\mathbf{x}|t = 1)p(t = 1)}$. Note that $p(\mathbf{x}|t) = \prod_{i=1}^{D} p(x_i|t)$ since x_i are independent given t.

$$
\begin{aligned}
\frac{p(\mathbf{x}|t=0)p(t=0)}{p(\mathbf{x}|t=1)p(t=1)} &= \exp\left(\ln\left(\frac{p(\mathbf{x}|t=0)p(t=0)}{p(\mathbf{x}|t=1)p(t=1)}\right)\right) \qquad \text{(since probabilities} \geq 0\text{)} \\
&= \exp\left(\ln\frac{p(t=0)}{p(t=1)} + \ln\frac{p(\mathbf{x}|t=0)}{p(\mathbf{x}|t=1)}\right) \\
&= \exp\left(\ln\frac{1-\alpha}{\alpha} + \ln\prod_{i=1}^{D}\frac{p(x_i|t=0)}{p(x_i|t=1)}\right) \\
&= \exp\left(\ln\frac{1-\alpha}{\alpha} + \sum_{i=1}^{D}\ln\frac{p(x_i|t=0)}{p(x_i|t=1)}\right)
\end{aligned}
$$

Now we simplify $\ln\frac{p(x_i|t=0)}{p(x_i|t=1)}$. Since $x_i \sim \mathcal{N}(\mu_{it}, \sigma_i^2)$:

$$
\begin{aligned}
\ln\frac{p(x_i|t=0)}{p(x_i|t=1)} &= \ln\frac{\exp\left(\frac{-(x_i-\mu_{i0})^2}{2\sigma_i^2}\right)}{\exp\left(\frac{-(x_i-\mu_{i1})^2}{2\sigma_i^2}\right)} \\
&= \ln\exp\left(\frac{(x_i-\mu_{i1})^2-(x_i-\mu_{i0})^2}{2\sigma_i^2}\right) \\
&= \frac{x_i^2-2\mu_{i1}x_i+\mu_{i1}^2-x_i^2+2\mu_{i0}x_i-\mu_{i0}^2}{2\sigma^2} \\
&= \frac{2(\mu_{i0}-\mu_{i1})x_i+(\mu_{i1}^2-\mu_{i0}^2)}{2\sigma^2} \\
&= \frac{\mu_{i0}-\mu_{i1}}{\sigma_i^2}x_i + \frac{\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2}
\end{aligned}
$$

Substituting all our terms, we get

$$
p(t=1|\mathbf{x}) = \frac{1}{1+\exp\left(\ln\frac{1-\alpha}{\alpha} + \sum_{i=1}^{D}\frac{\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2} + \sum_{i=1}^{D}\frac{(\mu_{i0}-\mu_{i1})}{\sigma_i^2}x_i\right)}
$$

Setting $b = \sum_{i=1}^{D}\frac{\mu_{i0}^2-\mu_{i1}^2}{2\sigma_i^2} - \ln\frac{1-\alpha}{\alpha}$ and $w_i = \frac{\mu_{i1}-\mu_{i0}}{\sigma_i^2}$ we get

$$
p(t=1|\mathbf{x}) = \frac{1}{1+\exp\left(-\sum_{i=1}^{D}w_ix_i - b\right)} = \sigma(\mathbf{w}^T\mathbf{x}+b)
$$

as desired.

(b) Since t is a binary variable, $p(t^{(n)} = 0|\mathbf{x}^{(n)}, \mathbf{w}, b) = 1 - p(t^{(n)} = 1|\mathbf{x}^{(n)}, \mathbf{w}, b) = 1 - \sigma(\mathbf{w}^T\mathbf{x} + b)$. We can conveniently use the exponents t and $1-t$ as an indicator variable. Thus, by the bernoulli distribution:

$$p(t^{(i)}|\mathbf{x}^{(i)}, \mathbf{w}, b) = \sigma(\mathbf{w}^T\mathbf{x} + b)^{t^{(i)}} \cdot \left[1 - \sigma(\mathbf{w}^T\mathbf{x} + b)\right]^{1-t^{(i)}}$$

Since the likelihood is the probability our weights match the training samples:

$$\begin{aligned}
Likelihood(\mathbf{w}, b) &= p(t^{(1)}, ..., t^{(N)}|\mathbf{x}^{(i)}, \mathbf{w}, b) \\
&= \prod_{i=1}^{N} p(t^{(i)}|\mathbf{x}^{(i)}, \mathbf{w}, b) && \text{(i.i.d. assumption)} \\
&= \prod_{i=1}^{N} \sigma(\mathbf{w}^T\mathbf{x} + b)^{t^{(i)}} \cdot \left[1 - \sigma(\mathbf{w}^T\mathbf{x} + b)\right]^{1-t^{(i)}}
\end{aligned}$$

Taking the negative log, we get the negative log-likelihood:

$$\begin{aligned}
L(\mathbf{w}, b) &= -\sum_{i=1}^{N} t^{(i)} \ln\left(\sigma(\mathbf{w}^T\mathbf{x} + b)\right) + (1 - t^{(i)}) \ln\left[1 - \sigma(\mathbf{w}^T\mathbf{x} + b)\right] \\
&= -\sum_{i=1}^{N} \ln\left[1 - \sigma(\mathbf{w}^T\mathbf{x} + b)\right] + t^{(i)} \ln\left(\frac{\sigma(\mathbf{w}^T\mathbf{x} + b)}{1 - \sigma(\mathbf{w}^T\mathbf{x} + b)}\right) \\
&= -\sum_{i=1}^{N} \ln\left[\frac{\exp(-\mathbf{w}^T\mathbf{x} - b)}{1 + \exp(-\mathbf{w}^T\mathbf{x} - b)}\right] + t^{(i)}\sigma^{-1}\sigma(\mathbf{w}^T\mathbf{x} + b) && \text{(logit)} \\
&= \sum_{i=1}^{N} \ln\left[1 + \exp(\mathbf{w}^T\mathbf{x} + b)\right] - t^{(i)}(\mathbf{w}^T\mathbf{x} + b)
\end{aligned}$$

Taking the derivative with respect to w_i:

$$\begin{aligned}
\frac{\partial L(\mathbf{w}, b)}{\partial w_i} &= \sum_{i=1}^{N} \frac{\exp(\mathbf{w}^T\mathbf{x} + b)}{1 + \exp(\mathbf{w}^T\mathbf{x} + b)} x_i - t^{(i)} x_i \\
&= \sum_{i=1}^{N} \frac{1}{1 + \exp(-\mathbf{w}^T\mathbf{x} - b)} x_i - t^{(i)} x_i \\
&= \sum_{i=1}^{N} \left[\sigma(\mathbf{w}^T\mathbf{x} + b) - t^{(i)}\right] x_i
\end{aligned}$$

And since b has no coefficient:

$$\frac{\partial L(\mathbf{w}, b)}{\partial b} = \sum_{i=1}^{N} \sigma(\mathbf{w}^T\mathbf{x} + b) - t^{(i)}$$

(c) By Bayes' Rule:

$$p(\mathbf{w}, b|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w}, b)p(\mathbf{w}, b)}{p(\mathcal{D})}$$

Since $p(\mathcal{D})$ does not depend on the model parameters and $\mathbf{x}^{(i)}$ are deterministic in $\mathcal{D} = \{(\mathbf{x}^{(1)}, t^{(1)}), ..., (\mathbf{x}^{(N)}, t^{(N)})\}$,

$$p(\mathbf{w}, b|\mathcal{D}) \propto p(\mathcal{D}|\mathbf{w}, b)p(\mathbf{w}, b) = p(\mathcal{D}|\mathbf{w}, b)p(\mathbf{w})p(b)$$

To calculate L_{post}, first, we note that since $p(w_i) = \mathcal{N}(w_i|0, 1/\lambda)$

$$\begin{aligned} p(w_1)p(w_2)...p(w_N) &= \sqrt{\frac{\lambda}{2\pi}} \prod_{i=1}^{N} \exp\left(\frac{-\lambda w_i^2}{2}\right) \\ &= \sqrt{\frac{\lambda}{2\pi}} \exp\left(\frac{-\lambda}{2}\sum_{i=1}^{N} w_i^2\right) \end{aligned}$$

And per the definition of $Likelihood(\mathbf{w}, b)$ in part b:

$$\begin{aligned} L_{post}(\mathbf{w}, b) &= -\ln(p(\mathbf{w}, b|t^{(1)}, ..., t^{(1)})) \\ &= -\ln(A \cdot p(\mathbf{w})p(b) \cdot Likelihood(\mathbf{w}, b)) \qquad \text{(for some constant } A) \\ &= -\ln(Likelihood(\mathbf{w}, b)) - \ln(p(w_1) \cdot ... \cdot p(w_N)) - \ln(A) \\ &= L(\mathbf{w}, b) - \ln\left(\sqrt{\frac{\lambda}{2\pi}} \exp\left(\frac{-\lambda}{2}\sum_{i=1}^{N} w_i^2\right)\right) - \ln(A) \\ &= L(\mathbf{w}, b) + \frac{\lambda}{2}\sum_{i=1}^{N} w_i^2 - \frac{1}{2}\ln\left(\frac{A^2\lambda}{2\pi}\right) \\ &= L(\mathbf{w}, b) + \frac{\lambda}{2}\sum_{i=1}^{N} w_i^2 + C \end{aligned}$$

where $C = -\frac{1}{2}\ln\left(\frac{A^2\lambda}{2\pi}\right)$ which is only dependent on λ.

As per calculations of the derivative of $L(\mathbf{w}, b)$, the derivative with respect to w_i and b:

$$\begin{aligned} \frac{\partial L_{post}(\mathbf{w}, b)}{\partial w_i} &= \sum_{i=1}^{N} \left[\sigma(\mathbf{w}^T\mathbf{x} + b) - t^{(i)}\right] x_i + \lambda w_i \\ \frac{\partial L_{post}(\mathbf{w}, b)}{\partial b} &= \sum_{i=1}^{N} \left[\sigma(\mathbf{w}^T\mathbf{x} + b) - t^{(i)}\right] \end{aligned}$$

$\square$

Problem 3: Naïve Bayes.

(a) Derive the maximum likelihood estimator for class-conditional probabilities $\boldsymbol{\theta}$ and the prior $\boldsymbol{\pi}$.

(b) Derive the log-likelihood $\log p(\mathbf{t}|\mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\pi})$.

(c) Derive the Maximum a posteriori Probability (MAP) estimator for the class-conditional pixel probabilities $\boldsymbol{\theta}$, using a Beta(3, 3) prior on each θ_{jc}.

Proof. Let N be the number of training samples and $\ell(x)$ be the log-likihood function of random variable x.

(a) To get $\hat{\boldsymbol{\theta}}_{MLE}$, it suffices to get its components, $\hat{\theta}_{jc}$. By definition:

$$\text{Likelihood of } \theta_{jc} = \prod_{i=1}^{N} p(x_j^{(i)}|c, \theta_{jc})^{t_c^{(i)}}$$

$$\begin{aligned}
\ell(\theta_{jc}) &= \log\left(\prod_{i=1}^{N} p(x_j^{(i)}|c, \theta_{jc})\right) \\
&= \sum_{i=1}^{N} t_c^{(i)} \left(x_j^{(i)} \log(\theta_{jc}) + (1 - x_j^{(i)}) \log(1 - \theta_{jc})\right)
\end{aligned}$$

$$\begin{aligned}
\frac{d\ell(\theta_{jc})}{d\theta_{jc}} &= \sum_{i=1}^{N} t_c^{(i)} \left(\frac{x_j^{(i)}}{\theta_{jc}} - \frac{1 - x_j^{(i)}}{1 - \theta_{jc}}\right) \\
&= \frac{1}{\theta_{jc}} \sum_{i=1}^{N} t_c^{(i)}(x_j^{(i)}) - \frac{1}{1 - \theta_{jc}} \sum_{i=1}^{N} t_c^{(i)}(1 - x_j^{(i)})
\end{aligned}$$

Setting the derivative to 0 and multiplying both sides by $(\theta_{jc})(1 - \theta_{jc})$:

$$\begin{aligned}
\theta_{jc} \sum_{i=1}^{N} t_c^{(i)}(1 - x_j^{(i)}) &= (1 - \theta_{jc}) \sum_{i=1}^{N} t_c^{(i)} x_j^{(i)} \\
\theta_{jc} \sum_{i=1}^{N} t_c^{(i)}(1 - x_j^{(i)} + x_j^{(i)}) &= \sum_{i=1}^{N} t_c^{(i)} x_j^{(i)} \\
\Rightarrow \hat{\theta}_{jc} &= \frac{\sum t_c^{(i)} x_j^{(i)}}{\sum t_c^{(i)}}
\end{aligned}$$

Intuitively, this means that the $\hat{\theta}_{jc}$ is the sum of x_j's among all samples labeled c, divided by all samples labeled c.

Now, to get $\hat{\boldsymbol{\pi}}_{MLE}$, we write down the likelihood:

$$\begin{aligned}\text{Likelihood of } \boldsymbol{\pi} &= \prod_{i=1}^{N} p(\boldsymbol{t}^{(i)}|\boldsymbol{\pi}) \\ &= \prod_{i=1}^{N}\prod_{c=0}^{9} \pi_c^{t_c^{(i)}}\end{aligned}$$

$$\ell(\boldsymbol{\pi}) = \sum_{i=1}^{N}\sum_{c=0}^{9} t_c^{(i)} \log(\pi_c)$$

We have a constraint $\sum_{c=1}^{9} \pi_c = 1$. Let $g(\boldsymbol{\pi}) = 1 - \sum_{c=1}^{9} \pi_c$ By way of Lagrange multipliers, we know that $\nabla \ell(\boldsymbol{\pi}) = \lambda \nabla g(\boldsymbol{\pi})$ for some real λ.

Thus, to solve for the MLE, we just need to maximise the Lagrange function:

$$J(\boldsymbol{\pi}, \lambda) = \ell(\boldsymbol{\pi}) + \lambda g(\boldsymbol{\pi})$$

Taking the derivative of J with respect to π_c and setting it to zero we get:

$$\frac{dJ(\boldsymbol{\pi},\lambda)}{d\pi_c} = \frac{d\ell(\boldsymbol{\pi})}{d\pi_c} + \lambda \frac{dg(\boldsymbol{\pi})}{d\pi_c}$$

$$\sum_{i=1}^{N} \frac{t_c^{(i)}}{\pi_c} + \lambda = 0$$

$$\sum_{i=1}^{N} t_c^{(i)} = -\lambda \pi_c$$

$$\Rightarrow \sum_{c=1}^{9}\sum_{i=1}^{N} t_c^{(i)} = \sum_{c=1}^{9} -\lambda \pi_c \qquad \text{(summing up all derivatives)}$$

$$N = -\lambda \qquad \text{(since } \textstyle\sum_c \pi_c = 1\text{)}$$

$$\Rightarrow \hat{\pi}_c = \frac{\sum_i^N t_c^{(i)}}{N} \qquad \text{(substituting the last equation to the first)}$$

Intuitively, this is the number of samples labeled c, over the total number of samples.

(b) For a single training example, to get the log likelihood of $p(\boldsymbol{t}|\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\pi})$, we just calculate its individual components. By Bayes' Rule and the Law of Total Proability:

$$\begin{aligned}\log p(t_c|\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\pi}) &= \log\left(\frac{p(t_c)\prod_{j=1}^{784} p(x_j|t_c)}{\sum_{c'} p(t_{c'})\prod_{j=1}^{784} p(x_j|t_{c'})}\right) \\ &= \log(\pi_c) + \sum_{j=1}^{784}\Big[x_j \log(\theta_{jc}) + (1-x_j)\log(1-\theta_{jc})\Big] - \log\left(\sum_{c'} p(t_{c'})\prod_{j=1}^{784} p(x_j|t_{c'})\right)\end{aligned}$$

The subtrahend is easy to calculate since it is the log of the sum of the components. We simply raise e to the components, sum them up, and take the logarithm.

(c)

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{MAP} &\propto \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_{i=1}^{N}\prod_{j=1}^{784}\prod_{c=0}^{9} p(x_j^{(i)}|c, \theta_{jc})^{t_c^{(i)}} p(\theta_{jc}) \\ &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{i=1}^{N}\sum_{j=1}^{784}\sum_{c=0}^{9}\left(\log\left(\frac{\theta_{jc}^2(1-\theta_{jc})^2}{B(3,3)}\right) + \log\left(\theta_{jc}^{x_j}(1-\theta_{jc})^{1-x_j}\right)\right)\end{aligned}$$

We can ignore the $B(3,3)$ term as it is constant. Now consider the log-likelihood of a component of $\boldsymbol{\theta}$:

$$\ell(\theta_{jc}) = 2\log(\theta_{jc}) + 2\log(1-\theta_{jc}) + \sum_{i=1}^{N} t_c^{(i)}\left[(x_j)\log(\theta_{jc}) + (1-x_j)\log(1-\theta_{jc})\right]$$

$$\frac{d\ell(\theta_{jc})}{d\theta_{jc}} = \left(\frac{2+\sum_{i=1}^{N} t_c^{(i)} x_j}{\theta_{jc}} - \frac{2+\sum_{i=1}^{N} t_c^{(i)}(1-x_j)}{1-\theta_{jc}}\right)$$

For simplicity, we denote $N_C = \sum_{i=1}^{N} t_c^{(i)}$. Setting the derivative to 0 and multiplying both sides by $\theta_{jc}(1-\theta_{jc})$ yields:

$$\begin{aligned}0 &= (2+N_C x_j)(1-\theta_{jc}) - (2+N_C(1-x_j))\theta_{jc} \\ &= 2 + N_C x_j - 4\theta_{jc} - N_C\theta_{jc} \\ (4-N_C)\theta_{jc} &= 2 + N_C x_j\end{aligned}$$

$$\Rightarrow \hat{\theta}_{jc} = \frac{2+N_C x_j}{4+N_C} = \frac{2+\sum_{i=1}^{N} t_c^{(i)} x_j}{4+\sum_{i=1}^{N} t_c^{(i)}}$$

Thus, with a Beta(3, 3) prior, $\hat{\boldsymbol{\theta}}_{MAP}$ will not have the same calculation error as $\hat{\boldsymbol{\theta}}_{MLE}$ since

$$0 < 2 + \sum_{i=1}^{N} t_c^{(i)} x_j < 4 + \sum_{i=1}^{N} t_c^{(i)}$$

implies that no component of $\hat{\boldsymbol{\theta}}_{MAP}$ is 1 or 0.

$\square$