HW3

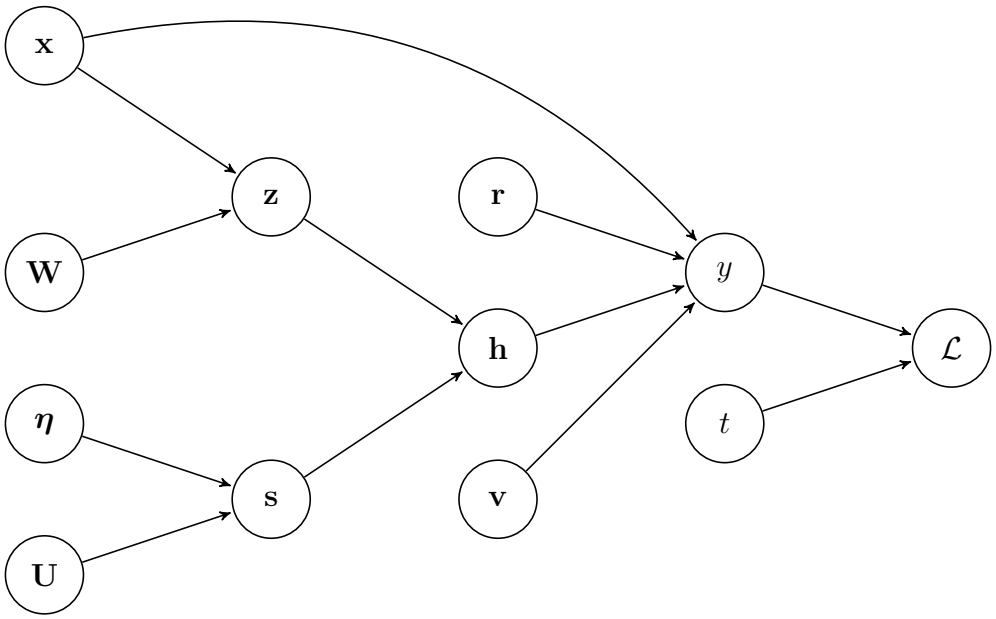
Chao Lin Wang

July 2024

1 Backprop.

(a)

The following is the computation graph of the given architecture.



(b)

The computation is as follows:

$$\begin{aligned}
\frac{\mathrm{d}\sigma(x)}{\mathrm{d}x} &= \frac{\mathrm{d}}{\mathrm{d}x}\frac{1}{1+e^{-x}} \\
&= \left(-\frac{1}{(1+e^{-x})^2}\right)\left(-e^{-x}\right) \\
&= \frac{1}{(1+e^{-x})}\frac{e^{-x}}{(1+e^{-x})} \\
&= \sigma(x)\frac{(1+e^{-x})-1}{(1+e^{-x})} \\
&= \sigma(x)\left(1+\frac{-1}{(1+e^{-x})}\right) \\
&= \sigma(x)\left(1-\sigma(x)\right).
\end{aligned}$$

(c)

Let use compute the error signals for the parameters along with $\overline{\mathbf{x}}$ and $\overline{\boldsymbol{\eta}}$ using backprop. Note that the convention of gradients being **row vectors** is used here.

$$\begin{aligned}
\overline{\mathcal{L}} &\equiv 1 \\
\overline{y} = \overline{\mathcal{L}}\,\frac{\partial \mathcal{L}}{\partial y} &= \frac{t-y}{y(1-y)} \\
\boxed{\overline{\mathbf{r}}} = \overline{y}\,\frac{\partial y}{\partial \mathbf{r}} &= \boxed{\overline{y}\;\sigma(\mathbf{v}^\top\mathbf{h}+\mathbf{r}^\top\mathbf{x})(1-\sigma(\mathbf{v}^\top\mathbf{h}+\mathbf{r}^\top\mathbf{x}))\mathbf{x}^\top} \\
\overline{\mathbf{h}} = \overline{y}\,\frac{\partial y}{\partial \mathbf{h}} &= \overline{y}\;\sigma(\mathbf{v}^\top\mathbf{h}+\mathbf{r}^\top\mathbf{x})(1-\sigma(\mathbf{v}^\top\mathbf{h}+\mathbf{r}^\top\mathbf{x}))\mathbf{v}^\top \\
\boxed{\overline{\mathbf{v}}} = \overline{y}\,\frac{\partial y}{\partial \mathbf{v}} &= \boxed{\overline{y}\;\sigma(\mathbf{v}^\top\mathbf{h}+\mathbf{r}^\top\mathbf{x})(1-\sigma(\mathbf{v}^\top\mathbf{h}+\mathbf{r}^\top\mathbf{x}))\mathbf{h}^\top} \\
\overline{\mathbf{z}} = \overline{\mathbf{h}}\,\frac{\partial \mathbf{h}}{\partial \mathbf{z}} &= \overline{\mathbf{h}}\;\mathrm{diag}\,(\mathbf{s}) \\
\overline{\mathbf{s}} = \overline{\mathbf{h}}\,\frac{\partial \mathbf{h}}{\partial \mathbf{s}} &= \overline{\mathbf{h}}\;\mathrm{diag}\,(\mathbf{z}) \\
\overline{\mathbf{x}} = \overline{\mathbf{z}}\,\frac{\partial \mathbf{z}}{\partial \mathbf{x}} + \overline{y}\,\frac{\partial y}{\partial \mathbf{x}} &= \boxed{\overline{\mathbf{z}}\;\mathbf{W} + \overline{y}\;\sigma(\mathbf{v}^\top\mathbf{h}+\mathbf{r}^\top\mathbf{x})(1-\sigma(\mathbf{v}^\top\mathbf{h}+\mathbf{r}^\top\mathbf{x}))\mathbf{r}^\top} \\
\overline{\mathbf{W}} = \overline{\mathbf{z}}\,\frac{\partial \mathbf{z}}{\partial \mathbf{W}} &= \boxed{\overline{\mathbf{z}}(\mathbf{1}\mathbf{x}^\top)} \\
\overline{\boldsymbol{\eta}} = \overline{\mathbf{s}}\,\frac{\partial \mathbf{s}}{\partial \boldsymbol{\eta}} &= \boxed{\overline{\mathbf{s}}\;\mathbf{U}} \\
\overline{\mathbf{U}} = \overline{\mathbf{s}}\,\frac{\partial \mathbf{s}}{\partial \mathbf{U}} &= \boxed{\overline{\mathbf{s}}(\mathbf{1}\eta^\top)}
\end{aligned}$$

2 Fitting a Naïve Bayes Model.

(a)

Let us follow the procedure for the *maximum likelihood estimator* (MLE).
First, we find the likelihood. Using the independence of the N random samples, we get:

$$L(\mathbf{x}, c|\boldsymbol{\theta}, \boldsymbol{\pi}) = \prod_{i=1}^{N} p(\mathbf{x}^{(i)}, c^{(i)}|\boldsymbol{\theta}, \boldsymbol{\pi}).$$

Further, by applying the Naïve Bayes assumption, we have:

$$\begin{aligned} L(\mathbf{x}, c|\boldsymbol{\theta}, \boldsymbol{\pi}) &= \prod_{i=1}^{N} p(c^{(i)}|\boldsymbol{\theta}, \boldsymbol{\pi}) p(\mathbf{x}^{(i)}|c^{(i)}, \boldsymbol{\theta}, \boldsymbol{\pi}) \\ &= \prod_{i=1}^{N} \left(p(c^{(i)}|\boldsymbol{\pi}) \prod_{j=1}^{784} p(x_j^{(i)}|c^{(i)}, \theta_{jc}) \right). \end{aligned}$$

Second, we can apply the log to get the log-likelihood:

$$\begin{aligned} l(\mathbf{x}, c|\boldsymbol{\theta}, \boldsymbol{\pi}) &= \sum_{i=1}^{N} \log \left(p(c^{(i)}|\boldsymbol{\pi}) \prod_{j=1}^{784} p(x_j^{(i)}|c^{(i)}, \theta_{jc}) \right) \\ &= \sum_{i=1}^{N} \left(\log p(c^{(i)}|\boldsymbol{\pi}) + \sum_{j=1}^{784} \log p(x_j^{(i)}|c^{(i)}, \theta_{jc}) \right) \\ &= \underbrace{\sum_{i=1}^{N} \log p(c^{(i)}|\boldsymbol{\pi})}_{\substack{\text{Log-likelihood} \\ \text{of labels}}} + \sum_{j=1}^{784} \underbrace{\sum_{i=1}^{N} \log p(x_j^{(i)}|c^{(i)}, \theta_{jc})}_{\substack{\text{Log-likelihood} \\ \text{for feature } x_j}}. \end{aligned} \tag{3}$$

Finally, we differentiate the log-likelihood to optimize the parameters $\boldsymbol{\theta} = \left[\theta_{jc}\right]$ and $\boldsymbol{\pi}$.
Note that only the first term of eq. (3)[1] depends on $\boldsymbol{\pi}$ and only the second term of eq. (3) depends on $\boldsymbol{\theta}$.

Finding $\hat{\theta}_{jc}$: Let us optimize an arbitrary θ_{jc} (i.e. pick any $j \in \{1, \ldots, 784\}$ and any $c \in \{0, \ldots, 9\}$). Let us take the derivative of the log-likelihood w.r.t. the θ_{jc}.

[1] Equation (1) and (2) are from the hw3.pdf handout.

$$
\begin{aligned}
\frac{\partial l(\mathbf{x}, c|\boldsymbol{\theta}, \boldsymbol{\pi})}{\partial \theta_{jc}} &= \frac{\partial}{\partial \theta_{jc}} \sum_{j'=1}^{784} \sum_{i=1}^{N} \log p(x_{j'}^{(i)}|c^{(i)}, \theta_{j'c}) \\
&= \sum_{j'=1}^{784} \delta_{jj'} \frac{\partial}{\partial \theta_{jc}} \sum_{i=1}^{N} \log \prod_{c'=0}^{9} \left(\theta_{j'c'}^{x_{j'}^{(i)}} (1-\theta_{j'c'})^{1-x_{j'}^{(i)}} \right)^{t_{c'}^{(i)}} \qquad \text{(by Equation (1))} \\
&= \sum_{i=1}^{N} \frac{\partial}{\partial \theta_{jc}} \sum_{c'=0}^{9} t_{c'}^{(i)} \left(x_j^{(i)} \log \theta_{jc'} + (1-x_j^{(i)}) \log(1-\theta_{jc'}) \right) \\
&= \sum_{i=1}^{N} \sum_{c'=0}^{9} \delta_{cc'} \frac{\partial}{\partial \theta_{jc}} t_{c'}^{(i)} \left(x_j^{(i)} \log \theta_{jc'} + (1-x_j^{(i)}) \log(1-\theta_{jc'}) \right) \\
&= \sum_{i=1}^{N} t_c^{(i)} \left(\frac{x_j^{(i)}}{\theta_{jc}} - \frac{1-x_j^{(i)}}{1-\theta_{jc}} \right) \\
&= \frac{\sum_{i=1}^{N} t_c^{(i)} x_j^{(i)}}{\theta_{jc}} - \frac{\sum_{i=1}^{N} t_c^{(i)} \left(1-x_j^{(i)}\right)}{1-\theta_{jc}}
\end{aligned}
$$

Now we set the derivative to 0 and solve for $\hat{\theta}_{jc}$.

$$
\begin{aligned}
0 \overset{\text{Set}}{=} \left. \frac{\partial l(\mathbf{x}, c|\boldsymbol{\theta}, \boldsymbol{\pi})}{\partial \theta_{jc}} \right|_{\theta_{jc}=\hat{\theta}_{jc}} &= \frac{\sum_{i=1}^{N} t_c^{(i)} x_j^{(i)}}{\hat{\theta}_{jc}} - \frac{\sum_{i=1}^{N} t_c^{(i)} \left(1-x_j^{(i)}\right)}{1-\hat{\theta}_{jc}} \\
0 &= (1-\hat{\theta}_{jc}) \sum_{i=1}^{N} t_c^{(i)} x_j^{(i)} - \hat{\theta}_{jc} \sum_{i=1}^{N} t_c^{(i)} \left(1-x_j^{(i)}\right) \\
\hat{\theta}_{jc} \sum_{i=1}^{N} t_c^{(i)} x_j^{(i)} + \hat{\theta}_{jc} \sum_{i=1}^{N} t_c^{(i)} \left(1-x_j^{(i)}\right) &= \sum_{i=1}^{N} t_c^{(i)} x_j^{(i)} \\
\hat{\theta}_{jc} \left[\sum_{i=1}^{N} t_c^{(i)} x_j^{(i)} + t_c^{(i)} \left(1-x_j^{(i)}\right) \right] &= \sum_{i=1}^{N} t_c^{(i)} x_j^{(i)} \\
\hat{\theta}_{jc} &= \frac{\sum_{i=1}^{N} t_c^{(i)} x_j^{(i)}}{\sum_{i=1}^{N} t_c^{(i)}} = \frac{\sum_{i=1}^{N} \mathbb{1}[x_j^{(i)} = 1 \ \& \ c^{(i)} = c]}{\sum_{i=1}^{N} \mathbb{1}[c^{(i)} = c]} \\
\hat{\theta}_{jc} &= \boxed{\frac{\text{\# feature } j \text{ appears in class } c}{\text{\# class } c \text{ appears in data}}}
\end{aligned}
$$

We can confirm that this is indeed the optimal parameter if we notice that the second derivative of the log-likelihood w.r.t. the θ_{jc} is always negative(i.e. $\frac{\partial^2 l}{(\partial \theta_{jc})^2} < 0$).[2]

[2]The optimal parameters $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ can be vectorized to make use of matrix operations. See more in part (c) and *naive_bayes.py*.

Finding $\hat{\boldsymbol{\pi}}$: Since one of the π_j's is constrained by the others, let us write $\pi_9 = 1 - \sum_{j=0}^{8} \pi_j$. Now let us optimize an arbitrary π_j where $j \in \{0, \ldots, 8\}$. Let us take the derivative of the log-likelihood w.r.t. the π_j.

$$\begin{aligned}
\frac{\partial l(\mathbf{x}, c|\boldsymbol{\theta}, \boldsymbol{\pi})}{\partial \pi_j} &= \frac{\partial}{\partial \pi_j} \sum_{i=1}^{N} \log p(c^{(i)}|\boldsymbol{\pi}) \\
&= \sum_{i=1}^{N} \frac{\partial}{\partial \pi_j} \log \prod_{j'=0}^{9} \pi_{j'}^{t_{j'}^{(i)}} \\
&= \sum_{i=1}^{N} \frac{\partial}{\partial \pi_j} \sum_{j'=0}^{8} t_{j'}^{(i)} \log \pi_{j'} + t_9^{(i)} \log \pi_9 \\
&= \sum_{i=1}^{N} \sum_{j'=0}^{8} \delta_{jj'} \frac{\partial}{\partial \pi_j} t_{j'}^{(i)} \log \pi_{j'} + \frac{\partial}{\partial \pi_j} t_9^{(i)} \log(1 - \textstyle\sum_{j''=0}^{8} \pi_{j''}) \\
&= \frac{\sum_{i=1}^{N} t_j^{(i)}}{\pi_j} - \frac{\sum_{i=1}^{N} t_9^{(i)}}{\pi_9}
\end{aligned}$$

Now we set the derivative to 0 and solve for $\hat{\pi}_j$.

$$\begin{aligned}
0 \overset{\text{Set}}{=} \left. \frac{\partial l(\mathbf{x}, c|\boldsymbol{\theta}, \boldsymbol{\pi})}{\partial \pi_j} \right|_{\pi_j = \hat{\pi}_j} &= \frac{\sum_{i=1}^{N} t_j^{(i)}}{\hat{\pi}_j} - \frac{\sum_{i=1}^{N} t_9^{(i)}}{\hat{\pi}_9} \\
0 &= \hat{\pi}_9 \sum_{i=1}^{N} t_j^{(i)} - \hat{\pi}_j \sum_{i=1}^{N} t_9^{(i)} \\
\hat{\pi}_j &= \hat{\pi}_9 \frac{\sum_{i=1}^{N} t_j^{(i)}}{\sum_{i=1}^{N} t_9^{(i)}} \qquad \text{for } j = 0, \ldots, 8 \\
\hat{\pi}_9 = 1 - \sum_{j=0}^{8} \hat{\pi}_j &= 1 - \hat{\pi}_9 \sum_{j=0}^{8} \frac{\sum_{i=1}^{N} t_j^{(i)}}{\sum_{i=1}^{N} t_9^{(i)}} \\
\hat{\pi}_9 \left(\sum_{i=1}^{N} t_9^{(i)} + \sum_{j=0}^{8} \sum_{i=1}^{N} t_j^{(i)} \right) &= \sum_{i=1}^{N} t_9^{(i)} \\
\hat{\pi}_9 &= \frac{\sum_{i=1}^{N} t_9^{(i)}}{N} = \boxed{\frac{\#\text{ class } c = 9 \text{ appears in data}}{\text{total } \#\text{ samples}}} \\
\text{for } j \neq 9: \qquad \hat{\pi}_j &= \frac{\sum_{i=1}^{N} t_9^{(i)}}{N} \frac{\sum_{i=1}^{N} t_j^{(i)}}{\sum_{i=1}^{N} t_9^{(i)}} = \boxed{\frac{\#\text{ class } c = j \text{ appears in data}}{\text{total } \#\text{ samples}}}
\end{aligned}$$

Again, we can confirm that $\frac{\partial^2 l}{(\partial \pi_j)^2} < 0$.

(b)

By Bayes' rule and the naïve Bayes assumption, we have the following log-likelihood of class c given a single image $\mathbf{x}^{(i)}$.

$$\begin{aligned} p(c|\mathbf{x}^{(i)}, \boldsymbol{\theta}, \boldsymbol{\pi}) &= \frac{p(c|\boldsymbol{\theta}, \boldsymbol{\pi}) p(\mathbf{x}^{(i)}|c, \boldsymbol{\theta}, \boldsymbol{\pi})}{p(\mathbf{x}^{(i)}|\boldsymbol{\theta}, \boldsymbol{\pi})} \\ &= \frac{p(c|\boldsymbol{\pi}) p(\mathbf{x}^{(i)}|c, \boldsymbol{\theta})}{\sum_{c'} p(c'|\boldsymbol{\pi}) p(\mathbf{x}^{(i)}|c', \boldsymbol{\theta})} \\ &= \frac{p(c|\boldsymbol{\pi}) \prod_{j=1}^{784} p(x_j^{(i)}|c, \boldsymbol{\theta})}{\sum_{c'} p(c'|\boldsymbol{\pi}) \prod_{j=1}^{784} p(x_j^{(i)}|c', \boldsymbol{\theta})} \end{aligned}$$

Now take the log.

$$\begin{aligned} \log p(c|\mathbf{x}^{(i)}, \boldsymbol{\theta}, \boldsymbol{\pi}) &= \log p(c|\boldsymbol{\pi}) + \sum_{j=1}^{784} \log p(x_j^{(i)}|c, \theta_{jc}) - \log \sum_{c'} p(c'|\boldsymbol{\pi}) \prod_{j=1}^{784} p(x_j^{(i)}|c', \theta_{jc}) \\ &= \log \pi_c + \sum_{j=1}^{784} \log \left(\theta_{jc}^{x_j^{(i)}} (1 - \theta_{jc})^{1 - x_j^{(i)}} \right) \\ &\quad - \log \sum_{c'=0}^{9} \pi_{c'} \prod_{j=1}^{784} \theta_{jc}^{x_j^{(i)}} (1 - \theta_{jc})^{1 - x_j^{(i)}} \\ &= \log \pi_c + \sum_{j=1}^{784} \left(x_j^{(i)} \log \theta_{jc} + (1 - x_j^{(i)}) \log(1 - \theta_{jc}) \right) \\ &\quad - \log \sum_{c'=0}^{9} \pi_{c'} \prod_{j=1}^{784} \theta_{jc}^{x_j^{(i)}} (1 - \theta_{jc})^{1 - x_j^{(i)}} \end{aligned} \tag{4}$$

(c)

Fitting the parameters was successful. It did help to vectorize some of the equations in part (a) as matrix operations simplify and optimize the calculations. For example, the (i, j) entry of $\mathbf{X}^\top \mathbf{t}$ counts the how many times the i-th feature appears class $c = j$. This is precisely the numerator of $\hat{\theta}_{jc}$. See more in `naive_bayes.py`.

At first, I used the form of log-likelihood as in eq. (4). However, it is numerically unstable as some values of $\hat{\boldsymbol{\theta}}$ can be 0, so a numerical log would have a hard time dealing with inputs $x \approx 0$ as $\lim_{x->0^+} \log x = -\infty$.

My fix for this is to compute the `numpy.log` of the $\boldsymbol{\theta}$ matrix with the argument: `where=(theta != 0)`. This means the log is only taken for non zero values; zero values are left as zero.

The resulting average log-likelihood for MLE is $\boxed{-3.46}$.

(d)

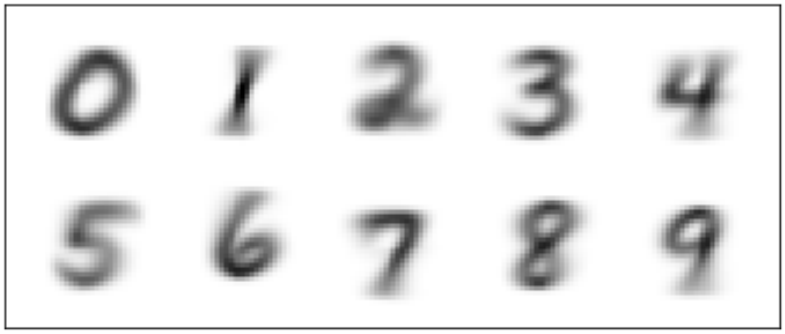


Figure 1: The MLE estimator $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ represented as 10 images, one for each class.

(e)

We are given that the prior distribution of the parameter θ_{jc} follows Beta(α, β). Therefore, we can obtain the posterior distribution of θ_{jc} using Bayes rule. Note that we are not given any prior information for $\boldsymbol{\pi}$, so let us take a uniform prior for each π_c (i.e. the $\hat{\boldsymbol{\pi}}$ here will be the same as the one in MLE).

Finding $\hat{\theta}_{jc}$:

$$
\begin{aligned}
p(\theta_{jc}|\mathbf{x}, c) &= \frac{\overbrace{p(\theta_{jc})}^{\text{prior}}\overbrace{p(\mathbf{x}, c|\theta_{jc})}^{\text{likelihood}}}{p(\mathbf{x}, c)} \\
&\propto p(\theta_{jc})p(\mathbf{x}, c|\theta_{jc}) \\
&= \left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta_{jc}^{\alpha-1}(1-\theta_{jc})^{\beta-1}\right]\left[\prod_{i=1}^{N}\Big(p(c^{(i)}|\boldsymbol{\pi})p(x_j^{(i)}|c^{(i)}, \theta_{jc})\Big)\right] \\
&\propto \left[\theta_{jc}^{\alpha-1}(1-\theta_{jc})^{\beta-1}\right]\left[\prod_{i=1}^{N}\Big(p(c^{(i)}|\boldsymbol{\pi})p(x_j^{(i)}|c^{(i)}, \theta_{jc})\Big)\right]
\end{aligned}
$$

Now optimize the above quantity by taking the log and the derivative.

$$\begin{aligned}
\log p(\theta_{jc}|\mathbf{x}, c) &\propto \log \left\{ \left[\theta_{jc}^{\alpha-1}(1-\theta_{jc})^{\beta-1}\right] \left[\prod_{i=1}^{N} \left(p(c^{(i)}|\boldsymbol{\pi})p(x_j^{(i)}|c^{(i)}, \theta_{jc})\right)\right]\right\} \\
&= (\alpha-1)\log\theta_{jc} + (\beta-1)\log(1-\theta_{jc}) \\
&\quad + \sum_{i=1}^{N} \log p(c^{(i)}|\boldsymbol{\pi}) + \sum_{i=1}^{N} \log \left[\left(\theta_{jc}^{x_j^{(i)}}(1-\theta_{jc})^{1-x_j^{(i)}}\right)^{t_c^{(i)}}\right] \\
&= (\alpha-1)\log\theta_{jc} + (\beta-1)\log(1-\theta_{jc}) \\
&\quad + \log\theta_{jc}\sum_{i=1}^{N} t_c^{(i)}x_j^{(i)} + \log(1-\theta_{jc})\sum_{i=1}^{N} t_c^{(i)}\left(1-x_j^{(i)}\right) + \sum_{i=1}^{N}\log p(c^{(i)}|\boldsymbol{\pi}) \\
&= \log\theta_{jc}\left(\alpha - 1 + \sum_{i=1}^{N} t_c^{(i)}x_j^{(i)}\right) + \log(1-\theta_{jc})\left(\beta - 1 + \sum_{i=1}^{N} t_c^{(i)}\left(1-x_j^{(i)}\right)\right) \\
&\quad + \sum_{i=1}^{N}\log p(c^{(i)}|\boldsymbol{\pi})
\end{aligned}$$

$$\frac{\partial}{\partial\theta_{jc}}\log p(\theta_{jc}|\mathbf{x}, c) = A\left[\frac{\alpha - 1 + \sum_{i=1}^{N} t_c^{(i)}x_j^{(i)}}{\theta_{jc}} - \frac{\beta - 1 + \sum_{i=1}^{N} t_c^{(i)}\left(1-x_j^{(i)}\right)}{1-\theta_{jc}}\right]$$

for some constant $A \neq 0$

$$\begin{aligned}
\frac{\partial}{\partial\theta_{jc}}\log p(\theta_{jc}|\mathbf{x}, c)\Big|_{\theta_{jc}=\hat{\theta}_{jc}} &\overset{\text{Set}}{=} 0 \\
\implies 0 &= \frac{\alpha - 1 + \sum_{i=1}^{N} t_c^{(i)}x_j^{(i)}}{\hat{\theta}_{jc}} - \frac{\beta - 1 + \sum_{i=1}^{N} t_c^{(i)}\left(1-x_j^{(i)}\right)}{1-\hat{\theta}_{jc}} \\
\hat{\theta}_{jc} &= \frac{\alpha - 1 + \sum_{i=1}^{N} t_c^{(i)}x_j^{(i)}}{\alpha - 1 + \sum_{i=1}^{N} t_c^{(i)}x_j^{(i)} + \beta - 1 + \sum_{i=1}^{N} t_c^{(i)}\left(1-x_j^{(i)}\right)} \\
\hat{\theta}_{jc} &= \boxed{\frac{\alpha - 1 + \#\text{ feature j appears in class } c}{\alpha + \beta - 2 + \#\text{ class } c \text{ appears in data}}}.
\end{aligned}$$

Finding $\hat{\boldsymbol{\pi}}$: With a uniform prior (pdf(x) = $\mathbb{1}(0 \leq x \leq 1)$), the posterior is just equal to the likelihood for $\boldsymbol{\pi}$. Therefore, the optimal $\boldsymbol{\pi}$ using MAP is the same as the MLE one:

$$\hat{\pi}_c = \frac{\sum_{i=1}^{N} t_c^{(i)}}{N} = \boxed{\frac{\#\text{ class } c \text{ appears in data}}{\text{total } \#\text{ samples}}}.$$

(f)

The average log-likelihood is $\boxed{-3.36}$.

The training accuracy for MAP is $\boxed{0.835}$.

The test accuracy for MAP is $\boxed{0.816}$.

(g)

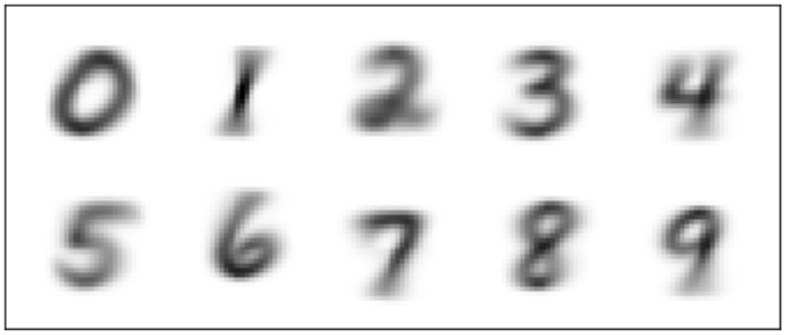


Figure 2: The MAP estimator $\hat{\boldsymbol{\theta}}_{\text{MAP}}$ represented as 10 images, one for each class. Note that the difference between this figure and fig. 1 is hardly perceptible by eye only. They are different numerically however.

(h)

One advantage to the Naïve Bayes approach is that the training process is simple and quick, amounting to computing "pseudo-counts".

In this problem though, it might not be appropriate as the Naïve Bayes assumption requires the probabilities of each feature to be conditionally independent on a given class. However, it is clear that features in this case are pixels of an image, and nearby pixels are likely correlated.

3 Logistic Regression with Gaussian Prior.

(a)

The likelihood is as follows:

$$\begin{aligned} L(\mathbf{x}, y|\boldsymbol{\theta}) &= \prod_{i=1}^{N} p(\mathbf{x}^{(i)}, y^{(i)}|\boldsymbol{\theta}) \\ &= \prod_{i=1}^{N} p(\mathbf{x}^{(i)}|\boldsymbol{\theta}) p(y^{(i)}|\mathbf{x}^{(i)}, \boldsymbol{\theta}) \\ &= \prod_{i=1}^{N} p(\mathbf{x}^{(i)}) \left(\frac{1}{1+\exp(-{\mathbf{x}^{(i)}}^T \boldsymbol{\theta})} \right)^{y^{(i)}} \left(1 - \frac{1}{1+\exp(-{\mathbf{x}^{(i)}}^T \boldsymbol{\theta})} \right)^{(1-y^{(i)})}. \end{aligned}$$

Now take the log for the log-likelihood.

$$\begin{aligned} l(\mathbf{x}, y|\boldsymbol{\theta}) &= \log L(\mathbf{x}, y|\boldsymbol{\theta}) \\ &= \sum_{i=1}^{N} \log p(\mathbf{x}^{(i)}) + \sum_{i=1}^{N} \Big[-y^{(i)} \log \Big(1 + \exp(-{\mathbf{x}^{(i)}}^T \boldsymbol{\theta}) \Big) \\ &\qquad + (1-y^{(i)}) \Big(\log \exp(-{\mathbf{x}^{(i)}}^T \boldsymbol{\theta}) - \log(1+\exp(-{\mathbf{x}^{(i)}}^T \boldsymbol{\theta})) \Big) \Big] \\ &= \boxed{\sum_{i=1}^{N} \log p(\mathbf{x}^{(i)}) + \sum_{i=1}^{N} \big[-{\mathbf{x}^{(i)}}^T \boldsymbol{\theta} \big]} \end{aligned}$$

To optimize for $\hat{\boldsymbol{\theta}}$, one could set the derivative of the log-likelihood $\dfrac{\partial}{\partial \boldsymbol{\theta}} l(\mathbf{x}, y|\boldsymbol{\theta})$ to 0 and solve. Note that the first term above should not depend on $\boldsymbol{\theta}$.

(b)

Given the prior, we know the likelihood for the posterior:

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{x}, y) &\propto p(\boldsymbol{\theta}) L(\mathbf{x}, y|\boldsymbol{\theta}) \\ &\propto \prod_{i=1}^{N} \exp\left(-\frac{1}{2} {\mathbf{x}^{(i)}}^T (\sigma^2 I)^{-1} \mathbf{x}^{(i)} \right) \\ &\quad \times \prod_{i=1}^{N} p(\mathbf{x}^{(i)}) \left(\frac{1}{1+\exp(-{\mathbf{x}^{(i)}}^T \boldsymbol{\theta})} \right)^{y^{(i)}} \left(1 - \frac{1}{1+\exp(-{\mathbf{x}^{(i)}}^T \boldsymbol{\theta})} \right)^{(1-y^{(i)})}. \end{aligned}$$

Now, take the log.

$$\log p(\boldsymbol{\theta}|\mathbf{x}, y) \propto \boxed{-\frac{1}{2} \sum_{i=1}^{N} \left({\mathbf{x}^{(i)}}^T (\sigma^2 I)^{-1} \mathbf{x}^{(i)} \right) + \sum_{i=1}^{N} \log p(\mathbf{x}^{(i)}) + \sum_{i=1}^{N} \big[-{\mathbf{x}^{(i)}}^T \boldsymbol{\theta} \big]}.$$

4 Gaussian Discriminant Analysis.

Pick a arbitrary $k \in \{0, \ldots, 9\}$. The likelihood is as below.

$$\begin{aligned} L(\mathbf{x}, y = k|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) &= \prod_{i=1}^{N} p(y^{(i)} = k) p(\mathbf{x}^{(i)}|y^{(i)} = k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\ &= \prod_{i=1}^{N} \left[\frac{1}{10} (2\pi)^{-d/2} |\boldsymbol{\Sigma}_k|^{-1/2} \exp\left\{ -\frac{1}{2} (\mathbf{x}^{(i)} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}^{(i)} - \boldsymbol{\mu}_k) \right\} \right] \end{aligned}$$

The log-likelihood:

$$\begin{aligned} l(\mathbf{x}, y = k|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = &-N \log(10) - N\frac{d}{2} \log(2\pi) + N \log |\boldsymbol{\Sigma}_k|^{-1/2} \\ &+ \sum_{i=1}^{N} \left[-\frac{1}{2} (\mathbf{x}^{(i)} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}^{(i)} - \boldsymbol{\mu}_k) \right]. \end{aligned}$$

Finding $\hat{\boldsymbol{\mu}}_k$:

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\mu}_k} l(\mathbf{x}, y = k|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \Big|_{\boldsymbol{\mu}_k = \hat{\boldsymbol{\mu}}_k} &= 0 \\ 0 &= -\boldsymbol{\Sigma}_k^{-1} \sum_{i=1}^{N} (\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}}_k) \\ \implies \quad \hat{\boldsymbol{\mu}}_k &= \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}^{(i)} \end{aligned}$$

Finding $\hat{\boldsymbol{\Sigma}}_k$:

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\Sigma}_k} l(\mathbf{x}, y = k|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \Big|_{\boldsymbol{\Sigma}_k = \hat{\boldsymbol{\Sigma}}_k} &= 0 \\ \implies \quad \hat{\boldsymbol{\Sigma}}_k &= \frac{1}{N} \sum_{i=1}^{N} (\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}}_k)^T \\ &= (\mathbf{X} - \mathbf{1}\hat{\boldsymbol{\mu}}_k^T)^T (\mathbf{X} - \mathbf{1}\hat{\boldsymbol{\mu}}_k^T) \end{aligned}$$

(a)

By Bayes' rule, the average log-likelihood is:

$$p(y = k|\mathbf{x}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{p(\mathbf{x}|y = k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) p(\mathbf{x}, y = k|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{k'} p(\mathbf{x}|y = k\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}$$