e) $p(\theta_{jc}|x,\pi,c) \propto p(\theta_{jc})p(x,c|\theta_{jc},\pi) = p(\theta_{jc})\prod_{i=1}^{n} p(x^{(i)},c^{(i)}|\theta_{jc},\pi) = p(\theta_{jc})\prod_{i=1}^{n}\left[p(c^{(i)}|\pi)\prod_{j=1}^{d} p(x_j^{(i)}|c^{(i)},\theta_{jc})\right]$

$\propto \theta_{jc}^{\alpha-1}(1-\theta_{jc})^{\beta-1}\prod_{i=1}^{n}\left[p(c^{(i)}|\pi)\prod_{j=1}^{d} p(x_j^{(i)}|c^{(i)},\theta_{jc})\right]$ (note $p(x,c|\theta_{jc},\pi)$ is from part a).

$$\log(p(\theta_{jc}|x,\pi,c)) \propto (\alpha-1)\log\theta_{jc} + (\beta-1)\log(1-\theta_{jc}) + \sum_{i=1}^{n}\left[\log p(c^{(i)}|\pi) + \sum_{j=1}^{d}\log p(x_j^{(i)}|c^{(i)},\theta_{jc})\right]$$

$$= (\alpha-1)\log\theta_{jc} + (\beta-1)\log(1-\theta_{jc}) + \sum_{i=1}^{n}\left[\log \pi_{c^{(i)}} + \sum_{j=1}^{d}\log\left(\theta_{jc}^{x_j^{(i)}}(1-\theta_{jc})^{(1-x_j^{(i)})}\right)\right]$$

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

$$= (\alpha-1)\log\theta_{jc} + (\beta-1)\log(1-\theta_{jc}) + \sum_{i=1}^{n}\log\pi_c^{(i)} + \sum_{j=1}^{d}\left[N_{\text{pixel } j=1}\log\theta_{jc} + N_{\text{pixel } j=0}\log(1-\theta_{jc})\right]$$

where $N_{\text{pixel } j=1}$ is the # of pixels where j=1 for images in class c, and similarly for $N_{\text{pixel } j=0}$.

Set the derivative to 0, so

$$0 \overset{set}{=} \frac{d\log(p(\theta_{jc}|x,\pi,c))}{d\theta_{jc}} = \frac{\alpha-1}{\hat{\theta}_{jc}} - \frac{\beta-1}{1-\hat{\theta}_{jc}} + 0 + \frac{N_{\text{pixel } j=1}}{\hat{\theta}_{jc}} - \frac{N_{\text{pixel } j=0}}{1-\hat{\theta}_{jc}}$$

$$= \frac{N_{\text{pixel } j=1}+\alpha-1}{\hat{\theta}_{jc}} - \frac{N_{\text{pixel } j=0}+\beta-1}{1-\hat{\theta}_{jc}} = N_{\text{pixel } j=1}+\alpha-1-\hat{\theta}_{jc}(N_{\text{pixel } j=1}+\alpha-1)-\hat{\theta}_{jc}(N_{\text{pixel } j=0}+\beta-1)$$

$$\Rightarrow \hat{\theta}_{jc} = \frac{N_{\text{pixel } j=1}+\alpha-1}{N_{\text{pixel } j=1}+N_{\text{pixel } j=0}+\alpha+\beta-2}$$ which for $\alpha=3, \beta=3$ equals $$\frac{N_{\text{pixel } j=1}+2}{N_{\text{pixel } j=1}+N_{\text{pixel } j=0}+4}.$$

Note that the MLE estimator for θ_{jc} from part a can also be expressed as

$$\hat{\theta}_{jc} = \frac{N_{\text{pixel } j=1}}{N_{\text{pixel } j=1}+N_{\text{pixel } j=0}}$$ which is similar to the MAP estimator but without $\alpha-1$ in the numerator

and $\alpha+\beta-2$ in the denominator, where α and β are pseudo-counts.

Parts f) and g) on last page.

h) An advantage is that it is efficient; for instance, during training it only requires one pass through the data, and during testing applying Baye's rule can be cheap due to the model structure.
A disadvantage is that it may be less accurate in practice since its independence assumption is strong or naive; in this case, the probability of a pixel being 1 often affects the probability of neighboring pixels.

3. a) Note that $p(y=0|x,\theta) = 1 - \frac{1}{1+\exp(-x^T\theta)}$, so

$$L(\theta|x,y) = \prod_{i=1}^{N}\left[\left(\frac{1}{1+\exp(-x^{(i)T}\theta)}\right)^{y_i}\left(1-\frac{1}{1+\exp(-x^{(i)T}\theta)}\right)^{1-y_i}\right]$$

$$\log L(\theta|x,y) = \sum_{i=1}^{N}\log\left[\left(\frac{1}{1+\exp(-x^{(i)T}\theta)}\right)^{y_i}\left(1-\frac{1}{1+\exp(-x^{(i)T}\theta)}\right)^{1-y_i}\right]$$

$$= \sum_{i=1}^{N}\left[y_i\log\left(\frac{1}{1+\exp(-x^{(i)T}\theta)}\right) + (1-y_i)\log\left(\frac{\exp(-x^{(i)T}\theta)}{1+\exp(-x^{(i)T}\theta)}\right)\right]$$

$$= \sum_{i=1}^{N}\left[y_i\left(\log\left(\frac{1}{1+\exp(-x^{(i)T}\theta)}\right) - \log\left(\frac{\exp(-x^{(i)T}\theta)}{1+\exp(-x^{(i)T}\theta)}\right)\right) + \log\left(\frac{\exp(-x^{(i)T}\theta)}{1+\exp(-x^{(i)T}\theta)}\right)\right]$$

$$= \sum_{i=1}^{N}\left[y_i x_i^{(i)T}\theta + \log\left(\frac{\exp(-x^{(i)T}\theta)}{1+\exp(-x^{(i)T}\theta)}\right)\right]$$

I would optimize this using gradient descent or another iterative method like the Newton-Raphson method since there is no closed form solution to this maximization problem.

b) Note $p(\theta) = (2\pi)^{-P/2}|\sigma^2 I|^{-1/2}\exp\left(-\frac{1}{2}x^T(\sigma^2 I)^{-1}x\right) = (2\pi)^{-P/2}|\sigma^2 I|^{-1/2}\exp\left(-\frac{\|x\|^2}{2\sigma^2}\right)$

$$p(D|\theta) = L(\theta|x,y) = \prod_{i=1}^{N}\left[\left(\frac{1}{1+\exp(-x^{(i)T}\theta)}\right)^{y_i}\left(\frac{\exp(-x^{(i)T}\theta)}{1+\exp(-x^{(i)T}\theta)}\right)^{1-y_i}\right]$$

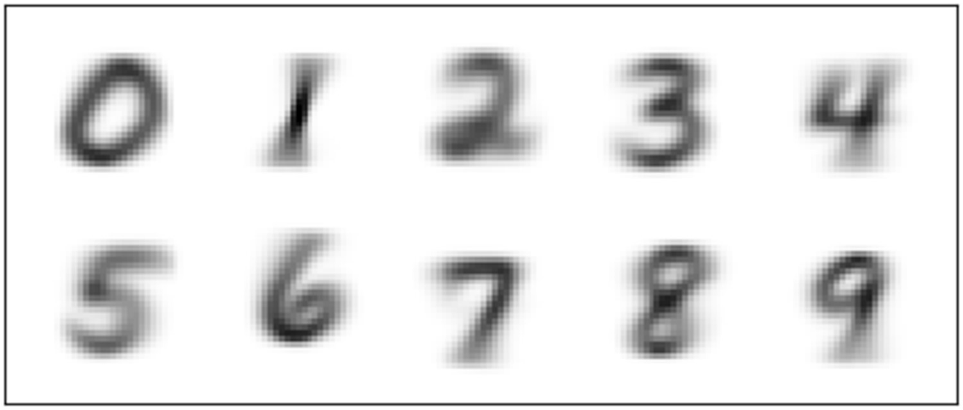
$p(\theta|D) \propto p(\theta)\,p(D|\theta)$ so

$\log p(\theta|x,y) \propto \log p(\theta) + \log p(D|\theta)$

$$= -\frac{P}{2}\log 2\pi - \frac{1}{2}\log|\sigma^2 I| - \frac{1}{2\sigma^2}\|x\|^2 + \sum_{i=1}^{N}\left[y_i x_i^{(i)T}\theta + \log\left(\frac{\exp(-x^{(i)T}\theta)}{1+\exp(-x^{(i)T}\theta)}\right)\right] \quad \text{(from part a)}$$

Question 2.

(d) The plot of the MLE estimator is as follows:



(f)

```
Average log-likelihood for MAP is  -3.3570631378602918
Training accuracy for MAP is  0.8352166666666667
Test accuracy for MAP is  0.816
```

(g) The plot of the MAP estimator is as follows: