CS6140 Assignment 3

1.4 ①

$$\sigma(a) = \frac{1}{1+e^{-w^T x}}$$

$$\frac{\partial \sigma(x)}{\partial x} = \frac{d\sigma(x)}{dx}\left(\left(1+e^{-w^T x}\right)^{-1}\right)$$

$$= \frac{\partial \sigma(x)}{\partial x}\left(-\frac{1}{\left(1+e^{-w^T x}\right)^2}\right)\left(-w^T e^{-w^T x}\right)$$

$$\boxed{\frac{\partial(\sigma(x))}{\partial(x)} = \frac{w^T e^{-w^T x}}{\left(1+e^{-w^T x}\right)^2}}$$

② $P(y=1|x,w) = \sigma(w^T x) = \frac{1}{1+e^{-w^T x}}$

$P(y=-1|x,w) = \sigma(w^T x) \stackrel{?}{=} \frac{1}{1+e^{w^T x}}$

$$P(y=-1|x,w) = 1 - P(y=1,x,w)$$

$$= 1 - \frac{1}{1+e^{-w^T x}}$$

$$= \frac{\cancel{1}+e^{-w^T x} - \cancel{1}}{1+e^{-w^T x}}$$

$$= \frac{e^{-w^T x}}{1+e^{-w^T x}} \times \frac{1/e^{-w^T x}}{1/e^{-w^T x}}$$

$$= \frac{1}{\frac{1}{e^{-w^T x}} + 1} = \frac{1}{1+e^{w^T x}} \checkmark$$

$$\boxed{\therefore P(y=\pm 1|x,w) = \sigma(w^T x) = \frac{1}{1+e^{-y w^T x}}}$$

③ $L_{log} = \sum_{i=1}^{N} \log\left(1+e^{-y_i w^T x_i}\right)$

$$L_{log} = \sum_{i=1}^{N} y_i \log\left(\sigma(w^T x_i)\right) + (1-y_i)\log\left(1-\sigma(w^T x_i)\right)$$

$$= \sum_{i=1}^{N} y_i\left(\log\left(\frac{1}{1+e^{-y_i w^T x_i}}\right)\right) + \log\left(\frac{1}{1+e^{-y_i w^T x_i}}\right) - y_i\left(\log\left(1-\frac{1}{1+e^{-y_i w^T x_i}}\right)\right)$$

from part 2

$$= \sum_{i=1}^{N} \cancel{-y_i\left(\log\left(1+e^{-y_i w^T x_i}\right)\right)} + \log\left(1+e^{-y_i w^T x_i}\right) \cancel{+ y_i\left(\log\left(1+e^{-y_i w^T x_i}\right)\right)}$$

$$\boxed{L_{log} = \sum_{i=1}^{N} \log\left(1+e^{-y_i w^T x_i}\right)} \checkmark$$

if (y_i) & $(w^T x_i)$ have the same sign, the error recorded (log loss) is <u>small</u> and close to 0, causing little correction during gradient descent. If their signs are different, error (log loss) is high, and a larger correction is made to the weight creating a model during future iterations with less log loss.

2.6

1. $$p(\theta; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}$$

Bernoulli likelihood:

$$L(\pi, \theta | D) = \prod_{i=1}^{N} \prod_{k=1}^{c} \prod_{j=1}^{|V|} \left(\pi_k \, \theta_{jk}^{ij} \left(1 - \theta_{jk}^{x_{ij}}\right) \right)^{y_{ik}}$$

$$L(\theta) \times p(\theta) = \underbrace{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(B)} \theta^{\alpha-1} (1-\theta)^{\beta-1} \times \left(\prod_{i=1}^{N} \prod_{k=1}^{c} \prod_{j=1}^{|V|} \pi_k \theta_{jk}^{ij} \left(1 - \theta_{jk}^{k_{ij}}\right) \right)}$$

$$\frac{\partial (L(\theta) \times p(\theta))}{\partial \theta} = \underbrace{0 = \frac{\partial}{\partial \theta} \left(\quad '' \quad \right)}$$

I have no idea how to solve this...
but solving for θ, will give us the MAP

2.6 ②

$$P(\theta) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \prod_{j=1}^{K} \theta_j^{\alpha_j - 1}$$

$$L(\theta) = \prod_{i=1}^{N} \prod_{k=1}^{C} \prod_{j=1}^{|V|} \left(\pi_k \, \theta_{jk}^{ij} \left(1 - \theta_{jk}^{x_{ij}}\right)\right)^{y_{ik}}$$

$$\frac{\partial (L(\theta) \times P(\theta))}{\partial \theta} = \underbrace{\frac{d}{d\theta} \left(\prod_{i=1}^{N} \prod_{j=1}^{C} \prod_{k=1}^{|V|} \left(\pi_k \theta_{jk}^{ij} \left(1 - \theta_{jk}^{x_{ij}}\right)\right)^{y_{ik}} \times \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \prod_{j=1}^{K} \theta_j^{\alpha_j - 1} \right) = 0}_{\text{Solve for } \theta \text{ for MAP estimate.}}$$

(2.6)

(3) The only thing that differs between MLE and MAP estimates is the inclusion of the prior, $P(\theta)$ in MAP. This means that the MAP is now weighted with some prior.

If the prior is uniform, NOT a beta or Dirichlet distribution, then we could ignore the constant and $\theta_{MAP} = \theta_{MLE}$.