Chapter 8. Logistic regression

8.3.
a.

$$\frac{d}{da}\left(\frac{1}{1+e^{-a}}\right) = -\frac{-e^{-a}}{(1+e^{-a})^2} = \frac{1}{1+e^{-a}} \cdot \frac{e^{-a}}{1+e^{-a}} = \sigma(a)(1-\sigma(a)).$$

b.

$$\mu_i = \frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}_i}} \Rightarrow \log \mu_i = -\log(1+e^{-\mathbf{w}^T\mathbf{x}_i})$$

$$\log(1-\mu_i) = -\mathbf{w}^T\mathbf{x}_i - \log(1+e^{-\mathbf{w}^T\mathbf{x}_i})$$

$$\Rightarrow f(\mathbf{w}) = \sum_i((1-y_i)\mathbf{w}^T\mathbf{x}_i + \log(1+e^{-\mathbf{w}^T\mathbf{x}_i})) = \sum_i(-y_i\mathbf{w}^T\mathbf{x}_i + \log(e^{\mathbf{w}^T\mathbf{x}_i}+1))$$

$$\Rightarrow \mathbf{g}(\mathbf{w}) = \sum_i(-y_i\mathbf{x}_i + \frac{\mathbf{x}_i e^{\mathbf{w}^T\mathbf{x}_i}}{1+e^{\mathbf{w}^T\mathbf{x}_i}}) = \sum_i(-y_i\mathbf{x}_i + \frac{\mathbf{x}_i}{1+e^{-\mathbf{w}^T\mathbf{x}_i}})$$

$$= \sum_i(\mu_i - y_i)\mathbf{x}_i = \mathbf{X}^T(\boldsymbol{\mu} - \mathbf{y}).$$

c. For nonzero $\mathbf{z} \in \mathbb{R}^n$, since $\mathbf{X}$ is full rank and $0 < \mu_i < 1$ for all i,

$$\mathbf{z}^T\mathbf{H}\mathbf{z} = \mathbf{z}^T\mathbf{X}^T\mathbf{S}\mathbf{X}\mathbf{z} = (\mathbf{X}\mathbf{z})^T\mathbf{S}(\mathbf{X}\mathbf{z}) = (\sqrt{\mathbf{S}}\mathbf{X}\mathbf{z})^T(\sqrt{\mathbf{S}}\mathbf{X}\mathbf{z}) > 0,$$

where $\sqrt{\mathbf{S}} = \mathrm{diag}(\sqrt{\mu_1(1-\mu_1)}, \cdots, \sqrt{\mu_n(1-\mu_n)})$.

8.4.
a.

$$\mu_{ik} = \frac{e^{\eta_{ik}}}{\sum_c e^{\eta_{ic}}} = e^{\eta_{ik} - \log(\sum_c e^{\eta_{ic}})}$$

$$\frac{\partial}{\partial \eta_{ij}}\mu_{ik} = \mu_{ik}\frac{\partial}{\partial \eta_{ij}}(\eta_{ik} - \log(\sum_c e^{\eta_{ic}})) = \mu_{ik}(\delta_{kj} - \frac{1}{\sum_c e^{\eta_{ic}}}\frac{\partial}{\partial \eta_{ij}}(\sum_c e^{\eta_{ic}}))$$

$$= \mu_{ik}(\delta_{kj} - \frac{1}{\sum_c e^{\eta_{ic}}}(\sum_c e^{\eta_{ic}}\delta_{cj})) = \mu_{ik}(\delta_{kj} - \frac{e^{\eta_{ij}}}{\sum_c e^{\eta_{ic}}}) = \mu_{ik}(\delta_{kj} - \mu_{ij}).$$

b.

$$l = \sum_i \sum_k (y_{ik}\log(\mu_{ik})) = \sum_i \sum_k (y_{ik}(\eta_{ik} - \log(\sum_c e^{\eta_{ic}})))$$

$$\Rightarrow \nabla_{\mathbf{w}_c} l = \sum_i (y_{ic}(1-\mu_{ic})\mathbf{x}_i + \sum_{c' \neq c} y_{ic'}(-\mu_{ic'})\mathbf{x}_i)$$

$$= \sum_i (y_{ic}(1-\mu_{ic})\mathbf{x}_i + (1-y_{ic})(-\mu_{ic})\mathbf{x}_i) = \sum_i ((y_{ic} - \mu_{ic})\mathbf{x}_i).$$

c.

$$\mathbf{H}_{c,c'} = \nabla_{\mathbf{w}_{c'}}(\nabla_{\mathbf{w}_c} l)^T = \nabla_{\mathbf{w}_{c'}}(\sum_i ((y_{ic} - \mu_{ic})\mathbf{x}_i^T))$$

$$= \sum_i (\nabla_{\mathbf{w}_{c'}}((y_{ic} - \mu_{ic})\mathbf{x}_i^T)) = -\sum_i \nabla_{\mathbf{w}_{c'}} \mu_{ic}\mathbf{x}_i^T = -\sum_i (\mu_{ic}(\delta_{c,c'} - \mu_{i,c'})\mathbf{x}_i\mathbf{x}_i^T).$$

8.5.

$$\nabla_{\mathbf{w}_c} l = \sum_i (y_{ic} - \mu_{ic})\mathbf{x}_i - \frac{\partial}{\partial \mathbf{w}_c}(\lambda \sum_{c'} \|\mathbf{w}_{c'}\|_2^2) = \sum_i (y_{ic} - \mu_{ic})\mathbf{x}_i - 2\lambda \mathbf{w}_c$$

We have

$$\sum_c \nabla_{\hat{\mathbf{w}}_c} l = \mathbf{0}$$

$$\Rightarrow \sum_i [\sum_c (y_{ic} - \mu_{ic})]\mathbf{x}_i - 2\lambda \sum_c \hat{\mathbf{w}}_c = \sum_i [\sum_c y_{ic} - \sum_c \mu_{ic}]\mathbf{x}_i - 2\lambda \sum_c \hat{\mathbf{w}}_c$$

$$= \sum_i (1-1)\mathbf{x}_i - 2\lambda \sum_c \hat{\mathbf{w}}_c = \mathbf{0}$$

$\Rightarrow \sum_c \hat{w}_{cj} = 0$ for all j.

8.6.

a. False.

To show this, we could show that $\mathbf{H}_{\text{neg}}$ is positive definite (and therefore the regularized loss is strictly convex).

Since

$$\mathbf{H} = \sum_i (\mu_i(1-\mu_i)\mathbf{x}_i\mathbf{x}_i^T) + 2\lambda$$

is positive definite, $J(\mathbf{w})$ has a unique global optimum.

b. False.
l_2 regularization tends to penalize larger weights more heavily, but this does not need to imply sparsity of the global optimum.

c. True.
If $\lambda = 0$, the optimum can be a step function, similar to a logistic function where $\|\mathbf{w}\| \to \infty$.

d. False.
If λ increases, bias for both train set and test set increase, therefore log likelihood decreases.

e. False.
If λ increases, bias for both train set and test set increase, therefore log likelihood decreases.

8.7.
a. A possible decision boundary would be a straight line that completely separates two sets.

b. A possible decision boundary would be a straight line that passes the origin, making impossible to separates two sets completely.

c. A possible decision boundary would be a straight line parallel to the x-axis.

d. A possible decision boundary would be a straight line parallel to the y-axis.