8.1

(a) $$\frac{\partial\, \sigma(a)}{\partial a} = -\frac{1}{(1+e^{-a})^2}\cdot -e^{-a}$$

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

$$= \sigma(a)\cdot(1-\sigma(a))$$

(b) $$1-\sigma(a) = \frac{e^{-a}}{1+e^{-a}} = \frac{1}{1+e^{a}} = \sigma(-a)$$

(c) $$y = \sigma(a) = \frac{1}{1+e^{-a}}$$

$$y + y e^{-a} = 1$$

$$e^{-a} = \frac{1-y}{y}$$

$$-a = \log(1-y) - \log y$$

$$a = \log y - \log(1-y)$$

$$\sigma^{-1}(a) = \log\frac{a}{1-a}$$

8.2

(a) $$E(w) = \sum_i -\{y_i \log\pi_i + (1-y_i)\cdot\log(1-\pi_i)\}$$

$$\nabla E(w) = \sum_i -\left\{\frac{y_i}{\pi_i}\cdot\pi_i(1-\pi_i)\cdot x_i - \frac{1-y_i}{1-\pi_i}\cdot\pi_i(1-\pi_i)\cdot x_i\right\}$$

$$= \sum_i -\{y_i(1-\pi_i)x_i - (1-y_i)\pi_i\cdot x_i\}$$

$$= \sum_i -\{y_i x_i - \pi_i x_i\}$$

$$= X\cdot(\pi-y)$$

(b) $$\nabla^2 E(w) = \frac{\partial}{w}\frac{\partial}{w^T}[E(w)]$$

$$= \frac{\partial}{\partial w}(\pi-y)^T\cdot X^T$$

(consider $\frac{\partial}{\partial w}\pi^T$

$= \left[\frac{\partial}{\partial w_1}\right]$

$$\begin{bmatrix} \vdots \\ \frac{\partial}{\partial w_n} \end{bmatrix} [\pi_1, \ldots, \pi_n] = [\pi_1(1-\pi_1)x_1, \ldots, \pi_n(1-\pi_n)x_n]$$

$$= XR$$

$$\nabla^2 E(w) = XRX^T$$

(c) since $\pi_i \in (0,1)$

Hence $R > 0$

(d) $w^{(new)} = w^{(old)} - [\nabla^2 E(w)]^{-1} \nabla E(w)$

$$w^{(new)} = w^{(old)} - (XRX^T)^{-1} X \cdot (\pi - y)$$

$$= (XRX^T)^{-1} XR(X^T w^{(old)} - R^{-1}(\pi - y))$$

8.3 (a) $\nabla E(w) = 0$

$\Rightarrow (\pi - y) X = 0$

$\Rightarrow \pi - y = 0$

Hence when $y_i = 1 \Rightarrow \sigma(w^T x) \to 1$

$w^T x \to \infty$

$y_i = 0 \Rightarrow \sigma(w^T x) \to 0$

$w^T x \to -\infty$

(b) If x is slightly across the margin

$p(y|x) \to 0/1$

It leaves no space for error/uncertainty.

8.4

(a) $\hat{E}(w) = E(w) - \log p(w)$

$= E(w) - \left(\log\left(\frac{1}{(2\pi)^{\frac{d}{2}} \cdot |\Sigma|^{\frac{1}{2}}}\right) - \frac{1}{2} w^T \Gamma w\right)$

$= E(w) + \frac{1}{2} w^T \Gamma w$

(b) $\nabla \hat{E}(w) = \nabla E(w) + \Gamma w$

$= X(\pi - y) + \Gamma w$

(c) $\nabla^2 \hat{E}(w) = \nabla^2 E(w) + \nabla^2 \frac{1}{2} w^T \Gamma w$

$= XRX^T + \frac{\partial}{\partial w} \frac{\partial}{\partial w^T} \left(\frac{1}{2} w^T \Gamma w\right)$

$= XRX^T + \frac{\partial}{\partial w} w^T \Gamma^T$

$= XRX^T + \Gamma^T = XRX^T + \Gamma$

(d) $w^{(new)} = w^{(old)} - (\nabla^2 \hat{E}(w))^{-1} \cdot \nabla \hat{E}(w)$

$= w^{(old)} - (XRX^T + \Gamma)^{-1} \cdot (X(\pi - y) + \Gamma w^{(old)})$

$= (XRX^T + \Gamma)^{-1} (XRX^T + \Gamma) w^{(old)} - (XRX^T + \Gamma)^{-1} (X(\pi - y) + \Gamma w^{(old)})$

$= (XRX^T + \Gamma)^{-1} (XRX^T w^{(old)} - X(\pi - y))$

$= (XRX^T + \Gamma)^{-1} \cdot XR \cdot z$

(f) $\hat{E}(w) = E(w) + \frac{1}{2} w^T \Gamma w$

if $\Gamma = \Gamma_2$, the last term, i.e. $\tilde{b}$

will not be constraint.

Consider (b) ⇒ gradient and analysis

8.5

(a) Define N_{mis} as the number of misclassifation points

By definition, $L_{01}(z_i) = \begin{cases} 0, & z_i \geq 0 \\ 1, & z_i < 0 \end{cases}$

gives 0 when z_i is successfuly predicted and 1 when z_i is misclassified, and $\sum_i L_{01}(z_i)$ is the total number of misclassified points

Hence $\min(N_{mis})$

$= \min\{\sum_i (L_{01}(z_i))\}$

(b) $R_{emp}(w)$

$= \sum_{i \in M} -y_i w^T x_i$

$= \sum_{i \in M} -z_i$

where M is set of misclassfied points

Hence

$L_p(z_i) = \begin{cases} 0, & z_i \geq 0 \\ -z_i, & z_i < 0 \end{cases}$

$= \max\{0, -z_i\}$

(c) $R_{emp}(w)$

$= \sum_i (y_i - w^T x_i)^2$

$= \sum_i (y_i^T y_i - 2y_i^T w^T x_i + x_i^T w w^T x_i)$

$= \sum_i (1 - 2z + x_i^T w y_i^T y_i w^T x_i)$

$= \sum_i (1 - 2z + z_i^T z)$

$= \sum_i (z_i - 1)^2$

Hence $L_{LSC}(z_i) = (z_i - 1)^2$

(4) $R_{emp}(w)$

$$= -\sum_i y_i \log \pi_i + (1-y_i)\log(1-\pi_i)$$

$$= -\sum_{i \in I_1} y_i \log \pi_i + \sum_{i \in I_2} (1-y_i)\log(1-\pi_i)$$

where I_1 are those points with $y_i = 1$, class label 1

I_2 are those points with $y_i = 0$, class label -1

$R_{emp}(w)$

$$= -\sum_{i \in I_1} \log \pi_i + \sum_{i \in I_2} \log(1-\pi_i)$$

$$= -\sum_{i \in I_1} \log \sigma(w^T x_i) + \sum_{i \in I_2} \log(1-\sigma((1-y_i)w^T x_i))$$

$$= -\sum_{i \in I_1} \log \sigma(z_i) + \sum_{i \in I_2} \log \sigma(y_i w^T x_i - w^T x_i)$$

$$= -\sum_{i \in I_1} \log \sigma(z_i) + \sum_{i \in I_2} \log \sigma(-w^T x_i)$$

$$= -\left(\sum_{i \in I_1} \log \sigma(z_i) + \sum_{i \in I_2} \log \sigma(z_i)\right)$$

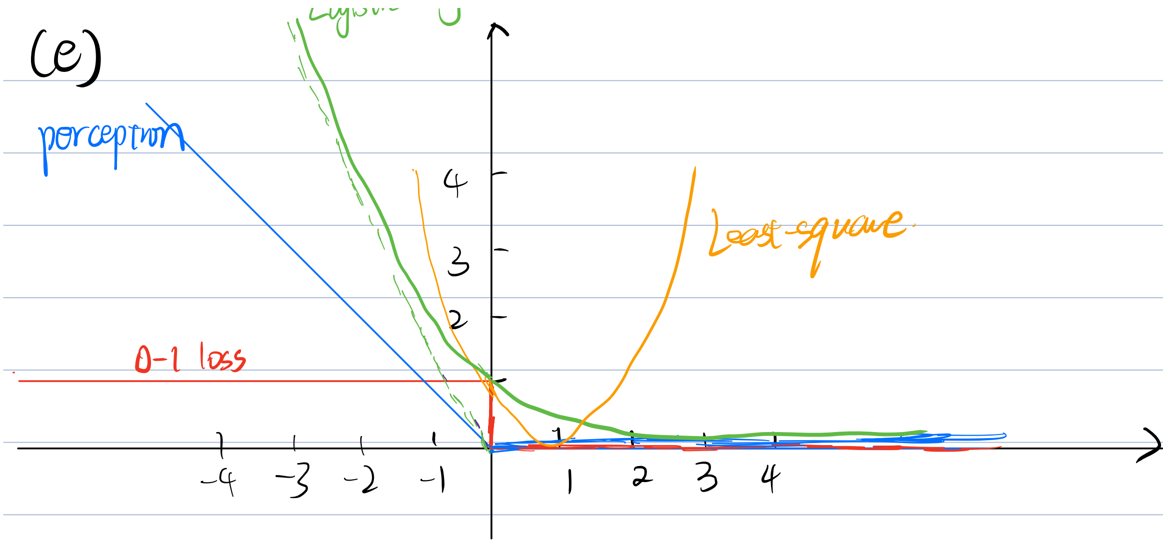
$$= -\sum_i \log \sigma(z_i)$$

$$= \log(1+e^{-z_i})$$

$$\propto \frac{1}{\log 2} \log(1+e^{-z_i})$$

logistic regression

(e)



Intuitively,

① least-square tends to penalize the too correct output which is not desirable.

② 0-1 loss function is difficult to optimize which is not desirable.

③ perception and logistic regression are similar, however, logistic regression has some loss for correctly classified points near the boundary, which tends to push the boundary further away from those points

Hence logistic regression may be better