Exercise 10.1

- from the notes in class: $G_m(x)$ is the classifier that minimises the weighted misclassification rate

GOAL: find that $\hat{\beta} = \frac{1}{2} \log \frac{1 - err^{(m)}}{err^{(m)}}$

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} w_i^{(m)} \exp\left\{ - y_i \beta G_m(x_i) \right\}$$

$$y_i = G_m(x_i) \quad \rightarrow \sum_{i : y_i = G_m(x_i)} w_i^{(m)} e^{-\beta}$$

$$y_i \neq G_m(x_i) \quad \rightarrow \sum_{i : y_i \neq G_m(x_i)} w_i^{(m)} e^{\beta}$$

$$L(\beta) = \sum_{i : y_i = G_m(x_i)} w_i^{(m)} e^{-\beta} + \sum_{i : y_i \neq G_m(x_i)} w_i^{(m)} e^{\beta}$$

$$\frac{\partial L(\beta)}{\partial \beta} = - \sum_{i : y_i = G_m(x_i)} w_i^{(m)} e^{-\beta} + \sum_{i : y_i \neq G_m(x_i)} w_i^{(m)} e^{\beta} = 0$$

multiply both sides by e^{β}

$$- \sum_{i : y_i = G_m(x_i)} w_i^{(m)} + e^{2\beta} \sum_{i : y_i \neq G_m(x_i)} w_i^{(m)} = 0$$

$$\sum_{i : y_i = G_m(x_i)} w_i^{(m)} = 1 - \sum_{i : y_i \neq G_m(x_i)} w_i^{(m)}$$

because $\tilde{\sum_{i}} w_i^{(m)} = 1$ w/o

$$e^{2\beta} \sum_{i : y_i \neq G_m(x_i)} w_i^{(m)} = 1 - \sum_{i : y_i \neq G_m(x_i)} w_i^{(m)}$$

$$e^{2\beta} = \frac{\sum_{i=1}^{N} w_i^{(m)} \Big/ \sum_{i=1}^{N} w_i^{(m)} - \sum_{i : y_i \neq G_m(x_i)} w_i^{(m)} \Big/ \sum_{i=1}^{N} w_i^{(m)}}{\sum_{i : y_i \neq G(x_i)} w_i^{(m)} \Big/ \sum_{i=1}^{N} w_i^{(m)}}$$

$$= \frac{1 - err_m}{err_m} \quad \Rightarrow \quad \hat{\beta} = \frac{1}{2} \log \frac{1 - err_m}{err_m}$$

Exercise 10.2

We need to prove that $\hat{f}^{(x)}_{AdaBoost} = f^*(x) = \underset{f(x)}{\arg\min}\, E_{y|x}\left[e^{-yf(x)}\right] = \frac{1}{2}\log\frac{Pr[y=1|x]}{Pr[y=-1|x]}$

$$\frac{\partial E_{y|x}\left[e^{-yf(x)}\right]}{\partial f(x)} = 0$$

under regularity condition (Lebesgue dominated convergence theorem) derivation and integration can be swapped (Leibniz integration rule)

$$E_{y|x}\left[\frac{\partial e^{-yf(x)}}{\partial f(x)}\right] = 0$$

$$E_{y|x}\left[-y\, e^{-yf(x)}\right] = 0$$

$$y = \begin{cases} -1 & \text{with prob } Pr[y=-1|x] \\ 1 & \text{with prob } Pr[y=1|x] \end{cases}$$

$$-(-1)e^{-(-1)f(x)}Pr[y=-1|x] - 1e^{-1f(x)}Pr[y=1|x] = 0$$

$$e^{f(x)}Pr[y=-1|x] - e^{-f(x)}Pr[y=1|x] = 0$$

multiply both sides by $e^{f(x)}$

$$e^{2f(x)}Pr[y=-1|x] - Pr[y=1|x] = 0$$

$$e^{2f(x)} = \frac{Pr[y=1|x]}{Pr[y=-1|x]}$$

$$\hat{f}^{(x)}_{AdaBoost} = \frac{1}{2}\log\frac{Pr[y=1|x]}{Pr[y=-1|x]} = \frac{1}{2}\log\frac{Pr[y=1|x]}{1-Pr[y=1|x]}$$

Exercise 10.8

Consider a K-class problem

y_{ik} = ith response

$$y_{ik} = \begin{cases} 1 & \text{if obs. } i \text{ is in class } k \\ 0 & \text{otherwise} \end{cases}$$

current model is $f_k(x)$, $k = 1, \dots, K$ with $\sum_{k=1}^{K} f_k(x) = 0$

GOAL: update the model $f_k(x)$ in a region R
by adding a constant γ_k, $f_k(x) + \gamma_k$, with $\gamma_K = 0$

(a) write down the multinomial log-likelihood, compute score (first derivative) and observed information (second derivative, or Hessian)

(b) using the Hessian from point (a), and starting from $\underline{\gamma_k = 0}$, $\forall k$, show that the one-step approximate Newton update is

$$\gamma_k^1 = \frac{\sum_{i \in R} (y_i - p_{ik})}{\sum_{i \in R} p_{ik}(1 - p_{ik})}, \quad k = 1, \dots, K-1$$

with $p_{ik} = \exp\{f_k(x_i)\} / \exp\{\sum_{\ell=1}^{K} f_\ell(x_i)\}$

(c) Show that $\hat{\gamma}_k = \frac{K-1}{K}\left(\gamma_k^1 - \frac{1}{K}\sum_{k=1}^{K} \gamma_k^1\right) = 0$

(a) $L_i(\gamma_k) = \underbrace{\frac{N}{y_{i1}! \cdot \dots \cdot y_{iK}!}}_{\text{constant}} \prod_{k=1}^{K} p_{ik}(\gamma_k)^{y_{ik}}$

$$\ell_i(\gamma_k) = \sum_{k=1}^{K} y_{ik} \log p_{ik}(\gamma_k)$$

$$= \sum_{k=1}^{K} y_{ik} \log\left(\exp\{f_k(x_i) + \gamma_k\} \Big/ \sum_{\ell=1}^{K} \exp\{f_\ell(x_i) + \gamma_\ell\}\right)$$

$$= \sum_{k=1}^{K} y_{ik}\left(f_k(x_i) + \gamma_k\right) - \underbrace{\sum_{k=1}^{K} y_{ik}}_{=1} \log \sum_{\ell=1}^{K} \exp\{f_\ell(x_i) + \gamma_\ell\}$$

$$= \sum_{k=1}^{K} y_{ik}\left(f_k(x_i) + \gamma_k\right) - \log \sum_{\ell=1}^{K} \exp\{f_\ell(x_i) + \gamma_\ell\}$$

$$\ell(\gamma_k) = \sum_{i=1}^{N} \ell_i(\gamma_k) = \sum_{i \in R}\left(\sum_{k=1}^{K} y_{ik}\left(f_k(x_i) + \gamma_k\right) - \log \sum_{\ell=1}^{K} \exp\{f_\ell(x_i) + \gamma_\ell\}\right)$$

$$\ell(\gamma_k) = \sum_{i=1}^{N} \ell_i(\gamma_k) = \sum_{i \in R} \left(\sum_{k=1}^{K} y_{ik} \left(f_k(x_i) + \gamma_k \right) - \log \sum_{\ell=1}^{K} \exp\{ f_\ell(x_i) + \gamma_\ell \} \right)$$

$$\ell_{\gamma_k}(\gamma_k) = \frac{\partial \ell(\gamma_k)}{\partial \gamma_k} = \sum_{i \in R} \left(y_{ik} - \frac{\exp\{ f_k(x_i) + \gamma_k \}}{\sum_{\ell=1}^{K} \exp\{ f_\ell(x_i) + \gamma_\ell \}} \right)$$

$$= \sum_{i \in R} \left(y_{ik} - p_{ik} \right)$$

$$\ell_{\gamma_k \gamma_k}(\gamma_k) = \frac{\partial^2 \ell(\gamma_k)}{\partial \gamma_k^2} = - \sum_{i \in R} \frac{\exp\{ f_k(x_i) + \gamma_k \} \sum_{\ell=1}^{K} \exp\{ f_\ell(x_i) + \gamma_\ell \} - \exp\{ f_k(x_i) + \gamma_k \}^2}{\left(\sum_{\ell=1}^{K} \exp\{ f_\ell(x_i) + \gamma_\ell \} \right)^2}$$

$$= - \sum_{i \in R} \left(\frac{\exp\{ f_k(x_i) + \gamma_k \}}{\sum_{\ell=1}^{K} \exp\{ f_\ell(x_i) + \gamma_\ell \}} - \frac{\exp\{ f_k(x_i) + \gamma_k \}^2}{\left(\sum_{\ell=1}^{K} \exp\{ f_\ell(x_i) + \gamma_k \} \right)^2} \right)$$

$$= - \sum_{i \in R} \left(p_{ik} - p_{ik}^2 \right)$$

$$= - \sum_{i \in} p_{ik} \left(1 - p_{ik} \right)$$

$$\ell_{\gamma_k \gamma_m}(\gamma_k) = + \sum_{i \in R} \frac{\exp\{ f_k(x_i) + \gamma_k + f_m(x_i) + \gamma_m \}}{\left(\sum_{\ell=1}^{K} \exp\{ f_\ell(x_i) + \gamma_\ell \} \right)^2}$$

Newton-step

$$\beta_{m+1} = \beta_m - \frac{f'(\beta_m)}{f''(\beta_m)}$$

(b) A Newton step has form

$$\gamma_k^1 = \gamma_k^0 - \ell_{\gamma_k \gamma_k}(\gamma_k^0)^{-1} \ell_{\gamma_k}(\gamma_k^0)$$

$$0 - \frac{\ell_{\gamma_k}(0)}{\ell_{\gamma_k \gamma_k}(0)}$$

$$\ell_{\gamma_k}(0) = \sum_{i \in R} \left(y_{ik} - p_{ik} \right)$$

$$\ell_{\gamma_k \gamma_k}(0) = - \sum_{i \in R} p_{ik} \left(1 - p_{ik} \right)$$

$$\gamma_k^1 = \frac{\sum_{i \in R} \left(y_{ik} - p_{ik} \right)}{\sum_{i \in R} p_{ik} \left(1 - p_{ik} \right)}$$

with $p_{ik} = \frac{\exp\{ f_k(x) \}}{\sum_{\ell=1}^{K} \exp\{ f_\ell(x) \}}$ under $\gamma_k = 0$ and $\gamma_\ell = 0$

(c) The update is

$$\gamma_k^1 = \frac{\sum_{i \in R} (y_{ik} - p_{ik})}{\sum_{i \in R} p_{ik}(1 - p_{ik})}$$

how I mean $\sum_{k=1}^{K} \gamma_k = 0$

By the properties of the mean $\sum_{i=1}^{N} (x_i - \bar{x}) = 0$

$$\gamma_k^1 - \frac{1}{K} \sum_{e=1}^{K} \gamma_e^1 = 0$$

Moreover $\frac{k-1}{K}$ is a constant

$$\frac{k-1}{K} \left(\gamma_k^1 - \frac{1}{K} \sum_{e=1}^{K} \gamma_e^1 \right) = 0$$

(k-1 because $\gamma_K = 0$ by construction)