Elements of Statistical Learning
Ejercicios Capítulo 10

Ejercicio 10.1.

La función $ae^{\beta} + be^{-\beta}$ es estrictamente convexa $\forall a, b > 0$. Entonces, basta con derivar con respecto a beta e igualar a cero para hallar el valor de β que minimiza la expresion.

$$f(\beta) = (e^{\beta} - e^{-\beta}) \sum_{i=1}^{N} w_i^{(m)} I(y_i \neq G(x_i)) + e^{-\beta} \sum_{i=1}^{N} w_i^{(m)}$$

$$\frac{\partial f}{\partial \beta} = (e^{\beta} + e^{-\beta}) \sum_{i=1}^{N} w_i^{(m)} I(y_i \neq G(x_i)) - e^{-\beta} \sum_{i=1}^{N} w_i^{(m)} = 0$$

$$(e^{\beta} + e^{-\beta}) \sum_{i=1}^{N} w_i^{(m)} I(y_i \neq G(x_i)) = e^{-\beta} \sum_{i=1}^{N} w_i^{(m)}$$

$$(e^{2\beta} + 1) \sum_{i=1}^{N} w_i^{(m)} I(y_i \neq G(x_i)) = \sum_{i=1}^{N} w_i^{(m)}$$

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

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

Si definimos err_m como $\frac{\sum_{i=1}^{N} w_i^{(m)} I(y_i \neq G(x_i)}{\sum_{i=1}^{N} w_i^{(m)}}$

$$2\beta = log(\frac{1 - err_m}{err_m})$$

$$\beta = \frac{1}{2} log(\frac{1 - err_m}{err_m})$$

Ejercicio 10.2.

Demostrar que $f^*(x) = argmin E_{Y|x}(e^{-Yf(x)}) = \frac{1}{2} log(\frac{Pr(Y=1|x)}{Pr(Y=-1|x)})$

$$E_{Y|x}(e^{-Yf(x)}) = e^{-f(x)} Pr(Y=1|x) + e^{f(x)} Pr(Y=-1|x)$$
$$\frac{\partial E_{Y|x}(e^{-Yf(x)})}{\partial f(x)} = -e^{-f(x)} Pr(Y=1|x) + e^{f(x)} Pr(Y=-1|x) = 0$$
$$e^{-f(x)} Pr(Y=1|x) = e^{f(x)} Pr(Y=-1|x)$$
$$e^{2f(x)} = \frac{Pr(Y=1|x)}{Pr(Y=-1|x)}$$
$$2f(x) = ln(\frac{Pr(Y=1|x)}{Pr(Y=-1|x)})$$
$$f(x) = \frac{1}{2} ln(\frac{Pr(Y=1|x)}{Pr(Y=-1|x)})$$

Ejercicio 10.3.

Marginal Average: $f_S(X_S) = E_{X_C} f(X_S, X_C)$. Si la distribución marginal de X_C es $\phi(X_C)$, entonces $E_{X_C} f(X_S, X_C) = \int f(X_S, X_C)\phi(X_C) dX_C$
Conditional Expectation: $\widetilde{f}_S(X_S) = E(f(X_S, X_C)|X_S) = \int f(X_S, X_C) p(X_C/X_S) dX_C$

Si $f(X) = h_1(X_S) + h_2(X_C)$

Marginal Average: $\int [h_1(X_S) + h_2(X_C)]\phi(X_C) dX_C$

$$\int h_1(X_S)\phi(X_C) dX_C + \int h_2(X_C)\phi(X_C) dX_C$$
$$h_1(X_S) \int \phi(X_C) dX_C + \int h_2(X_C)\phi(X_C) dX_C$$
$$h_1(X_S) + E_{X_C}(h_2(X_C))$$

Donde: $\int \phi(X_C) dX_C = 1$

Conditional Expectation: $E(f(X_S, X_C)|X_S) = E([h_1(X_S) + h_2(X_C)]|X_S)$
$$E(h_1(X_S)|X_S) + E(h_2(X_C)|X_S)$$
$$h_1(X_S) + E(h_2(X_C)|X_S)$$

Si $f(X) = h_1(X_S) h_2(X_C)$

Marginal Average: $\int [h_1(X_S) h_2(X_C)]\phi(X_C) dX_C$

$$h_1(X_S) \int h_2(X_C)\phi(X_C) dX_C$$
$$h_1(X_S) E_{X_C}(h_2(X_C))$$

Conditional Expectation: $E(f(X_S, X_C)|X_S) = E([h_1(X_S) h_2(X_C)]|X_S)$
$$h_1(X_S) E(h_2(X_C)|X_S)$$

En los *conditional expectation*, $E(h_2(X_C)|X_S)$ es una función de X_S, mientras que en los *marginal average*, $E_{X_C}(h_2(X_C))$ es una constante

Ejercicio 10.5 (a).

Population Minimizer of $E_{Y|x}(e^{-Yf(x)})$) subject to $\sum_{k=1}^{K} f_k = 0$
La variable target de cada observación será un vector con K entradas, donde K es el número de clases totales (van de 1 a K). La entrada i del vector toma el valor de 1 si es que la observación pertenece a la clase i. El resto de observaciones del vector Y toman el valor de $-\frac{1}{K-1}$

$$f^*(x) = argmin E_{Y|x}(e^{-\frac{1}{2}Y^T f}) \ s.t. \sum_{k=1}^{K} f_k = 0$$

$$E_{Y|x}(e^{-\frac{1}{2}Y^T f}) = Pr(Y=1|x)e^{f_1 - \frac{1}{1-K}(f_2+f_3+...+f_k)} + ... + Pr(Y=K|x)e^{-\frac{1}{1-K}(f_1+f_2+...+f_{k-1})+f_k}$$

Utilizando la restricción, se puede simplificar a:

$$E_{Y|x}(e^{-\frac{1}{2}Y^T f}) = \sum_{k=1}^{K} Pr(Y=k|x)e^{-\frac{f_k}{K-1}}$$

$$f^*(x) = argmin \sum_{k=1}^{K} Pr(Y=k|x)e^{-\frac{f_k}{K-1}} \text{ s.t. } \sum_{k=1}^{K} f_k = 0$$

Lagrangian FOC:

$$\frac{\partial L}{\partial f_k} = Pr(Y=k|x)e^{-\frac{f_k}{K-1}}(-\frac{1}{K-1}) + \lambda = 0$$

$$\frac{\partial L}{\partial \lambda} = \sum_{k=1}^{K} f_k = 0$$

De la primera condición de primer orden se puede obtener que:

$$Pr(Y=k|x)e^{-\frac{f_k}{K-1}} = (K-1)\lambda$$

$$e^{-\frac{f_k}{K-1}} = \frac{(K-1)\lambda}{Pr(Y=k|x)}$$

$$-\frac{f_k}{K-1} = ln(\frac{(K-1)\lambda}{Pr(Y=k|x)})$$

$$f_k^* = -(K-1)ln(\frac{(K-1)\lambda}{Pr(Y=k|x)})$$

$$\sum_{k=1}^{K} f_k = \sum_{k=1}^{K} -(K-1)ln(\frac{(K-1)\lambda}{Pr(Y=k|x)}) = -(K-1)\sum_{k=1}^{K} ln(\frac{(K-1)\lambda}{Pr(Y=k|x)}) = 0$$

$$\sum_{k=1}^{K} ln(\frac{(K-1)\lambda}{Pr(Y=k|x)}) = ln(\prod_{k=1}^{K} \frac{(K-1)\lambda}{Pr(Y=k|x)}) = 0$$

$$\prod_{k=1}^{K} \frac{(K-1)\lambda}{Pr(Y=k|x)} = 1$$

$$(K-1)^K \lambda^K = \prod_{k=1}^{K} Pr(Y=k|x)$$

$$\lambda = \frac{1}{K-1} \prod_{k=1}^{K} Pr(Y=k|x)^{\frac{1}{K}}$$

De las condiciones de primer orden podemos obtener una expresión para $Pr(Y = k|x)$

$$Pr(Y = k|x)e^{-\frac{f_k}{K-1}}(-\frac{1}{K-1}) + \lambda = 0$$

$$Pr(Y = k|x)e^{-\frac{f_k}{K-1}} = (K-1)\lambda$$

$$Pr(Y = k|x) = (K-1)\lambda e^{\frac{f_k}{K-1}}$$

Reemplazando los valores óptimos de λ y f_k

$$Pr(Y = k|x) = (K-1)(\frac{1}{K-1}\prod_{i=1}^{K} Pr(Y = k|x)^{\frac{1}{K}})e^{\frac{f_k^*}{K-1}}$$

$$Pr(Y = k|x) = (\prod_{i=1}^{K} Pr(Y = k|x)^{\frac{1}{K}})e^{\frac{f_k^*}{K-1}}$$

Aplicando sumatorias a ambos lados y tomando en cuenta que las probabilidades de 1 a K suman 1:

$$1 = (\prod_{i=1}^{K} Pr(Y = k|x)^{\frac{1}{K}})\sum_{k=1}^{K} e^{\frac{f_k^*}{K-1}}$$

$$\prod_{i=1}^{K} Pr(Y = k|x)^{\frac{1}{K}} = (\sum_{k=1}^{K} e^{\frac{f_k^*}{K-1}})^{-1}$$

Entonces, se concluye que:

$$Pr(Y = k|x) = \frac{e^{\frac{f_k^*}{K-1}}}{\sum_{k=1}^{K} e^{\frac{f_k^*}{K-1}}}$$

Ejercicio 10.6 (a).

El error estándar de una distribución binomial (en este caso el éxito es el error), se define como:

$$\sigma = \sqrt{\frac{p(1-p)}{n}}$$
$$\sigma(GBM) = \sqrt{\frac{0.045 * 0.955}{1536}} = 0.52\%$$
$$\sigma(GAM) = \sqrt{\frac{0.055 * 0.9945}{1536}} = 0.5967\%$$

Ejercicio 10.6 (b).

McNemar test (Agresti, 1996):

$$z = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}}$$
$$z = \frac{33 - 18}{\sqrt{33 + 18}} = 2.1$$
$$2 \text{ sided} p - value = 2 * (1 - P(z < 2.1)) = 0.0358$$

Ejercicio 10.7.

$$\hat{\gamma}_{jm} = argmin \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma_{jm})$$

Como en la expresión se asume una función de pérdida exponencial:

$$\hat{\gamma}_{jm} = argmin \sum_{x_i \in R_{jm}} e^{-y_i(f_{m-1}(x_i)}$$
$$\hat{\gamma}_{jm} = argmin \sum_{x_i \in R_{jm}} w_i^{(m)} e^{-y_i(\gamma_{jm})}, \text{ donde } e^{-y_i(f_{m-1}(x_i)+\gamma_{jm})} = w_i^{(m)}$$

CPO w.r.t. γ_{jm}

$$\sum_{x_i \in R_{jm}} w_i^{(m)} e^{-y_i(\gamma_{jm})}(-y_i) = 0$$
$$\sum_{x_i \in R_{jm}, y_i = 1} w_i^{(m)} e^{-\gamma_{jm}}(-1) + \sum_{x_i \in R_{jm}, y_i = -1} w_i^{(m)} e^{\gamma_{jm}} = 0$$
$$\sum_{x_i \in R_{jm}, y_i = 1} w_i^{(m)} e^{-\gamma_{jm}} = \sum_{x_i \in R_{jm}, y_i = -1} w_i^{(m)} e^{\gamma_{jm}} = 0$$
$$\sum_{x_i \in R_{jm}, y_i = 1} w_i^{(m)} e^{-\gamma_{jm}} = \sum_{x_i \in R_{jm}, y_i = -1} w_i^{(m)} e^{\gamma_{jm}}$$
$$\sum_{x_i \in R_{jm}, y_i = 1} w_i^{(m)} = e^{2\gamma_{jm}} \sum_{x_i \in R_{jm}, y_i = -1} w_i^{(m)}$$
$$\frac{\sum_{x_i \in R_{jm}, y_i = 1} w_i^{(m)}}{\sum_{x_i \in R_{jm}, y_i = -1} w_i^{(m)}} = e^{2\gamma_{jm}}$$
$$\frac{1}{2} ln(\frac{\sum_{x_i \in R_{jm}, y_i = 1} w_i^{(m)}}{\sum_{x_i \in R_{jm}, y_i = -1} w_i^{(m)}}) = \gamma_{jm}$$

Ejercicio 10.8 (a).

Según la página 349 del libro, la extensión de la función deviance a un contexto con K-clases es: $-\sum_{k=1}^{K} I(y = G_k) f_k(x) + log(\sum_{k=1}^{K} e^{f_k(x)})$. Entonces la función de log verosimilitud para este problema es:

$$L = -\sum_{k=1}^{K} I(y = G_k)[f_k(x) + \gamma_k] + log(\sum_{k=1}^{K} e^{f_k(x)+\gamma_k})$$

$$\text{Primera derivada: } \frac{\partial L}{\partial \gamma_k} = -I(y = G_k) + \frac{e^{f_k(x)+\gamma_k}}{\sum_{k=1}^{K} e^{f_k(x)+\gamma_k}}$$

$$\text{Segunda derivada: } \frac{\partial^2 L}{\partial \gamma_k^2} = \frac{e^{f_k(x)+\gamma_k} \sum_{k=1}^{K} e^{f_k(x)+\gamma_k} - e^{f_k(x)+\gamma_k} e^{f_k(x)+\gamma_k}}{(\sum_{k=1}^{K} e^{f_k(x)+\gamma_k})^2}$$

$$\frac{\partial^2 L}{\partial \gamma_k^2} = \frac{e^{f_k(x)+\gamma_k}}{\sum_{k=1}^{K} e^{f_k(x)+\gamma_k}} - \frac{e^{2(f_k(x)+\gamma_k)}}{(\sum_{k=1}^{K} e^{f_k(x)+\gamma_k})^2}$$

Ejercicio 10.8 (b).

La segunda derivada de halada en la parte (a) representa la diagonal de la matriz Hessiana. De acuerdo al método de Newton, tenemos que:$\gamma^{t+1} = \gamma^t - \frac{\frac{\partial L}{\partial \gamma_k}}{\frac{\partial^2 L}{\partial \gamma_k^2}}$

$$\gamma^1 = \gamma^0 - \frac{-I(y = G_k) + \frac{e^{f_k(x)}}{\sum_{k=1}^{K} e^{f_k(x)}}}{\frac{e^{f_k(x)}}{\sum_{k=1}^{K} e^{f_k(x)}} + \frac{e^{2(f_k(x))}}{(\sum_{k=1}^{K} e^{f_k(x)})^2}}$$

$$0 - \frac{-I(y = G_k) + \frac{e^{f_k(x)}}{\sum_{k=1}^{K} e^{f_k(x)}}}{\frac{e^{f_k(x)}}{\sum_{k=1}^{K} e^{f_k(x)}} + \frac{e^{2(f_k(x))}}{(\sum_{k=1}^{K} e^{f_k(x)})^2}}$$

$$\frac{I(y = G_k) - p_{ik}}{p_{ik} - p_{ik}^2}$$

$$\frac{y_{ik} - p_{ik}}{p_{ik}(1 - p_{ik})}$$

Como la función de pérdida de la parte 1 se calculó solo sobre la región R, tenemos que:

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

Ejercicio 10.8 (c).

Si a las realizaciones de una vairable se les desvía de su media, su suma será cero, entonces:

$$\sum_{i=1}^{K}(\gamma_k^1 - \frac{1}{K}\sum_{l=1}^{K}\gamma_l^1) = 0$$

La igualdad se mantiene si se le multiplica una constante igual a $\frac{K-1}{K}$

$$\sum_{i=1}^{K}\frac{K-1}{K}(\gamma_k^1 - \frac{1}{K}\sum_{l=1}^{K}\gamma_l^1) = 0$$

Ejercicio 10.9.

Hay pocas diferencias entre el algoritmo 10.3 y 10.4. Utilizando la lógica del algoritmo 10.3, se debe estimar el negativo del gradiente. Según el ejercicio 10.8, este es:

$$-(-I(y = G_k) + \frac{e^{f_k(x)+\gamma_k}}{\sum_{k=1}^{K} e^{f_k(x)+\gamma_k}}) = y_{ik} - \frac{e^{f_k(x)+\gamma_k}}{\sum_{k=1}^{K} e^{f_k(x)+\gamma_k}} = y_{ik} - p_k(x)$$

Luego, se calcula el γ que minimiza la función de pérdida, considerando que las probabilidades deben ser positivas. Por ese motivo, se usa el valor absoluto de r_{ikm}

Ejercicio 10.10.

Según (10.22), la función de multinomial deviance es $-\sum_{k=1}^{K} I(y = G_k) f_k(x) + log(\sum_{k=1}^{K} e^{f_i(x)})$

Como se indica en (10.38), cada árbol se estima con el negativo del gradiente $I(y = G_k) - p_k(x)$

En un problema de clasificación con dos clases, las probabilidades cumplen la condición: $p_1(x) + p_2(x) = 1$

El árbol T_1 se construye sobre el gradiente $I(y = G_1) - p_1(x)$

$I(y = G_1) - (1 - p_2(x)) = I(y = G_1) - 1 + p_2(x) = -(1 - I(y = G_1)) + p_2(x) = -I(y = G_2) + p_2(x)$

$-(I(y = G_2) - p_2(x))$

Entonces, basta con conocer estimar el árbol sobre el gradiente 1 para saber el del gradiente dos, pues es el mismo pero que con el signo cambiado

Ejercicio 10.12.

Mostrar que $E[f(X_1, X_2)|X_2] = \rho X_2$

Si $f(X_1, X_2) = X_1$ entonces, queda demostrar que $E[X_1|X_2] = \rho X_2$

$E[X_1 X_2] = \rho \rightarrow E[E[X_1 X_2|X2]] = \rho \rightarrow E[X_2 E[X_1|X2]] = \rho$

$E[X1|X2]$ es una función de X2, si consideramos que $E[X1|X2] = aX_2 + b$

$E[X_2(aX_2 + b)] = \rho \rightarrow aE[X_2^2] + bE[X_2] = \rho$

Dado que X_2 tiene varianza 1, $E[X_2^2] = 1$ y media cero

$a = \rho$

$b = 2$ porque $E[E(X_1|X_2)] = E[X_1] = 0$