2) $Pr[\text{type} = t, \text{length} = l] := \begin{cases} \frac{3}{4} \frac{3^l e^{-3}}{l!} & \text{if } t = 1 \text{ \& } l = 0, 1, 2 \ldots \\ \frac{1}{4} \frac{6^l e^{-6}}{l!} & \text{if } t = 0 \text{ \& } l = 0, 1, 2 \ldots \end{cases}$

$t = 1 \Rightarrow$ Salmon & $t = 0 \Rightarrow$ Sea bass

(a) $Pr[\text{type} = 1] = \sum Pr[\text{type} = 1, \text{length} = l]$

$$= \sum \frac{3}{4} \frac{3^l e^{-3}}{l!}$$

$$= \frac{3e^{-3}}{4} \sum \frac{3^l}{l!} \quad \text{where } l = 0, 1, 2 \ldots$$

& $\sum_{n=0}^{\infty} \frac{x^n}{n!} = e^x$

$$\therefore \frac{3^l}{l!} = e^3$$

$$\therefore Pr[\text{type} = 1] = \frac{3}{4} e^{-3} \cdot e^3$$

$$= \frac{3}{4} \,//$$

(b) Compute Likelihood function $Pr[\text{length} = l \mid \text{type} = 1]$
$\forall\, l = 0, 1, 2 \ldots$

According to Baye's rule,

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

$$\Rightarrow Pr[\text{length} = l \mid \text{type} = 1] = \frac{Pr[\text{type} = 1 \mid \text{length} = l]\, Pr[\text{length} = l]}{Pr[\text{type} = 1]}$$

& $$Pr[A \mid B] = \frac{Pr[A, B]}{Pr[B]}$$

$$\Rightarrow Pr[\text{length} = l \mid \text{type} = 1] = \frac{\dfrac{Pr[\text{type} = 1, \text{length} = l]}{\cancel{Pr[\text{length} = l]}} \cdot \cancel{Pr[\text{length} = l]}}{Pr[\text{type} = 1]}$$

$$= \frac{Pr[\text{type} = 1, \text{length} = l]}{Pr[\text{type} = 1]}$$

$$\Pr[\text{type}=1, \text{length}=l] = \frac{3}{4}\,\frac{3^{l}e^{-3}}{l!}$$

$$\& \quad \Pr[\text{type}=1] = \frac{3}{4}$$

$$\Rightarrow \quad \frac{\frac{3}{4}\,\frac{3^{l}e^{-3}}{l!}}{\frac{3}{4}} = \frac{3^{l}e^{-3}}{l!} \quad //$$

(c) Find optimal threshold that minimizes expected Zero-one loss for threshold function $1[\text{length} \leq \tau]$.

Zero-one loss function is given by,

$$L(\hat{t}, t) = \begin{cases} 1 & \text{if } \hat{t} \neq t \\ 0 & \text{otherwise} \end{cases}$$

Have to reduce the empherical risk $\mathbb{E}_{L,t}[(\ldots)]$

$$= \mathbb{E}[L(1(l \leq \tau), t) \mid t=1] P[t=1] + \mathbb{E}[L(1(l < \tau), t) \mid t=0] P[t=0]$$

$$= \mathbb{E}[1\{1[l \leq \tau] = 0\} \mid t=1] P[t=1] + \mathbb{E}[1\{1[l \leq \tau] = 1\} \mid t=0] P[t=0]$$

$$= \mathbb{E}[1[l > \tau] \mid t=1] P[t=1] + \mathbb{E}[1[l \leq \tau] \mid t=0] P[t=0]$$

$$= P[l > \tau \mid t=1] P[t=1] + P[l \leq \tau \mid t=0] P[t=0]$$

$$= [1 - F_1(\tau)] P[t=1] + F_0(\tau) P[t=0]$$

$P[t=0] = 1 - P[t=1] = 1/4$.

$$= \left[\left(1 - \frac{3}{4}\frac{3^r e^{-3}}{r!}\right)\frac{3}{4} + \left(\frac{1}{4}\frac{3^r e^{-3}}{r!}\right)\frac{1}{4}\right] //$$

(d) $l(0,1) = 1$, costs 1 for labeling salmon as sea bass

& $l(1,0) = 3$, costs 3 for labeling sea bass as salmon.

$l(1,1) = l(0,0) = 0$, since correct predictions

In order to compute the optimal threshold that minimized the expected loss, we compute the minimum possible overall risk $R(\alpha | x)$.

$$\Rightarrow R(type=1 | length=l) = l(1,1)P(type=1 | length=l) + l(1,0)P(type=0 | length=l)$$

$$\Rightarrow R(type=1 | length=l) = l(1,0)P(type=0 | length=l)$$
$[\because l(1,1)=0]$

$$\& \ R(type=0 | length=l) = l(0,1)P(type=1 | length=l) + l(0,0)P(type=0 | length=l)$$

$$\Rightarrow R(type=0 | length=l) = l(0,1)P(type=1 | length=l)$$
$[\because l(0,0)=0]$

So we have,

$$R(type=1 | length=l) = l(1,0)P(type=0 | length=l)$$

$$\& \ R(type=0 | length=l) = l(0,1)P(type=1 | length=l)$$

Bayes Decision Rule => optimal

$$l(1,0)\,P(\text{type}=0 \mid \text{length}=l) \underset{>\ \text{bass}}{\overset{<\ \text{salmon}}{?}} l(0,1)\,P(\text{type}=1 \mid \text{length}=l)$$

So, need to solve,

$$\frac{l(1,0)}{l(0,1)} < \frac{P(\text{type}=1 \mid \text{length}=l)}{P(\text{type}=0 \mid \text{length}=l)}$$

Substituting values for all,

$$\frac{3}{1} < \frac{\frac{3}{4}\,\frac{3^l e^{-3}}{l!}}{\frac{1}{4}\,\frac{6^l e^{-6}}{l!}}$$

$$1 < \frac{3^l e^{-3}}{6^l e^{-6}}$$

$$1 < \frac{3^l e^{3}}{2^l \times 3^l}$$

$$2^l < e^3$$

Taking log on both sides,

$$\log(2^{l}) < \log(e^{3})$$

$$l < 1.3028$$

$\therefore$ $l < 1.3028$ is the optimal threshold that minimized the expected less.

3) $Pr[\text{type}=1] = 1/2$

Likelihood function $Pr[\text{length}=l \mid \text{type}=1] = P_1(l)$

$Pr[\text{length}=l \mid \text{type}=0] = P_0(l)$

$\forall l \in N_0 = \{0, 1, \dots\}$

(a) Find optimal classifier $f: N_0 \to \{0,1\}$ that minimize expected zero-one loss.

The optimal classifier that minimizes expected zero-one loss is given by, $f(x) := \arg\max p(y|x)$

$\Rightarrow Pr[\text{type}=t \mid \text{length}=l]$

from Baye's equation, we have

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

& distribution is Continuous, $f(x) = \arg\max_y p(x|y)p(y)$

$\therefore Pr[\text{type}=t \mid \text{length}=l] = Pr[\text{length}=l \mid \text{type}=t]\, pr[t]$

$\Rightarrow Pr[\text{length}=l \mid \text{type}=1]\, pr[\text{type}=1] + pr[\text{length}=l \mid \text{type}=0]\, pr[\text{type}=0]$

$$= P_1(l)\cdot\frac{1}{2} + P_0(l)\cdot\frac{1}{2}$$

$$= \frac{1}{2}\left[p_0(l) + p_1(l)\right] \quad \forall\, l \in N_0$$

$\therefore$ This can be written as,

$$\Rightarrow \sum_{l \in N_0} \frac{1}{2}\left[p_0(l) + p_1(l)\right]$$

$$= \frac{1}{2} \sum_{l \in N_0} \left[p_0(l) + p_1(l)\right] \,//$$

which is the optimal classifier in terms of p_0 & p_1.

(b) $$d_{TV}(p_0, p_1) := \frac{1}{2} \sum_{l \in N_0} |p_0(l) - p_1(l)|$$

Show that :- expected optimal zero-one loss $= \frac{1}{2}\left(1 - d_{TV}(p_0, p_1)\right)$

$$d_{TV}(p_0, p_1) = \frac{1}{2}\left(p_0(l) - p_1(l)\right)$$

Substituting value of $p_0(l)$ & $p_1(l)$

$$= \frac{1}{2}\left(Pr[length = l \mid type = 0] - Pr[length = l \mid type = 1]\right)$$

$$= \frac{1}{2} Pr[length = l \mid type = 0] - \frac{1}{2} Pr[length = l \mid type = 1]$$

Wkt, $Pr[type = 1] = Pr[type = 0] = 1/2$

$\therefore$ The equation becomes,

$$\Rightarrow Pr[length = l \mid type = 0]\, Pr[type = 0] - Pr[length = l \mid type = 1]\, Pr[type = 1]$$

$$\Rightarrow d_{TV}(P_0, P_1) = Pr[type = 0 \mid length = l] - Pr[type = 1 \mid length = l]$$

Substituting d_{TV} in optimal expected zero-one loss we get,

$$\Rightarrow \frac{1}{2}\left(1 - d_{TV}(P_0, P_1)\right)$$

$$= \frac{1}{2}\left(1 - \left[Pr[type = 0 \mid length = l] - Pr[type = 1 \mid length = l]\right]\right)$$

$$= \frac{1}{2}\left(1 - Pr[type = 0 \mid length = l] + Pr[type = 1 \mid length = l]\right)$$

$$\Rightarrow 1 - Pr[type = 0 \mid length = l] = Pr[type = 1 \mid length = l]$$

$$\therefore = \frac{1}{2}\left[2 \cdot Pr[type = 1 \mid length = l]\right]$$

$$= Pr[type = 1 \mid length = l]$$

which is the optimal expected zero-one loss

5. Training Data -> feature $x_1, \ldots x_N \in R^d$

outcomes -> $y_1 \ldots y_N \in R$

$$L(\omega) := \frac{1}{2} \sum_{i=1}^{N} r_i (\omega^T x_i - y_t)^2$$

where $r_1 \ldots r_N > 0$.

(a) Show that $L(\omega) = (X\omega - y)^T R (X\omega - y)$.

Expanding the loss equation we have,

$$L(\omega) = \frac{1}{2} \left[r_1 (\omega^T x_1 - y_t)^2 + r_2 (\omega^T x_2 - y_t)^2 \ldots + r_N (\omega^T x_N - y_t)^2 \right]$$

we know that, By using design Matrix, the above equation can be represented as,

$$= \frac{1}{2} \left[\begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_N \end{bmatrix} \left(\begin{bmatrix} x_1^T \omega \\ x_2^T \omega \\ \vdots \\ x_N^T \omega \end{bmatrix} - \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \right)^2 \right]$$

This equation is nothing but,

$$= \frac{1}{2} R (X\omega - y)^2$$

1/2 value can be ignored because its just a constant & eliminating it won't change the difference between the predictions & true labels.

$$\Rightarrow \quad L(w) = R\,(Xw - y)^2$$

Furthermore,

$$= R\left((Xw)^2 - 2(Xw \cdot y) + (y)^2 \right)$$

$$= R\left(w^T X^T X w - w^T X^T y - w^T X^T y + y^T y \right)$$

$$= R\left(w^T X^T (Xw - y) - y\,(w^T X^T - y^T) \right)$$

$$= R\left(w^T X^T (Xw - y) - y\,(Xw - y)^T \right)$$

$$= R\left(Xw - y \right)^T (Xw - y)$$

$$\Rightarrow \quad L(w) = (Xw - y)^T R\,(Xw - y)$$ //

(b) when all r_i's are 1,

then $L(w^*) = \min_w (L(w))$

when $w^* = (X^T X)^{-1} X^T y$.

To compute: New w^* that minimizes $L(w)$ in closed form as a function of $X, R,$ & y.

from solution of (a), we know that,

$$L(w) = (Xw - y)^T R (Xw - y)$$

$$= \left((Xw)^T R - y^T R\right)(Xw - y)$$

$$= w^T X^T R X w - Y^T R X w - y^T R X w + y^T R y$$

$$= w^T X^T R X w - 2 y^T R X w + y^T R y \quad -①$$

& $$\frac{\partial (B^T A B)}{\partial B} = AB + A^T B$$

$$\frac{\partial (AB)}{\partial B} = A^T$$

$\therefore$ ① becomes,

$$\frac{\partial(L(w))}{\partial w} = X^TRXw + X^TRXw - 2X^TRY$$

$$= 2X^TRXw - 2X^TRY$$

Setting the derivative to 0,

$$\frac{\partial L(w)}{\partial w} = 0$$

$$\Rightarrow 2X^TRXw - 2X^TRY = 0$$

$$\cancel{2}X^TRXw = X^TRY$$

$$\therefore w^* = (X^TRX)^{-1}X^TRY$$

Maximum Likelihood Estimate :

(c) $$p(y_i | x_i ; \omega) = \frac{1}{\sqrt{2\pi \sigma_i^2}} \exp\left(- \frac{(y_i - \omega^T x_i)^2}{2(\sigma_i)^2} \right)$$

where Mean of $y_i \Rightarrow \omega^T x_i$

Variance $\Rightarrow \sigma_i^2$

Since log function is non-decreasing, we can take log of MLE function.

$$\Rightarrow \log(p(y_i | x_i ; \omega))$$

$$= \underset{\omega}{\operatorname{argmax}} -\frac{1}{2} \left(\frac{y_i - \omega^T x_i}{(\sigma_i)^2} \right)^2 \quad -①$$

Since σ_i^2 is the variance of y_i, we can represent the ~~...~~ Covariance Matrix of observation errors as,

$$\begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

Since Covariance Matrix $= \sigma^2 . I$

Since we have a weighted setting for Linear Regression here, i.e., we have weight for each data point,

we define the reciprocal of each variance σ_i^2 as,

$$\gamma_i = \frac{1}{\sigma_i^2}$$

∴ The Matrix R will be a diagonal Matrix having these Variance for each data point.

$$\Rightarrow R = \begin{bmatrix} 1/\sigma_1^2 & 0 & \cdots & 0 \\ 0 & 1/\sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/\sigma_n^2 \end{bmatrix}$$

∴ The log Likelihood function ① becomes,

$$= \underset{w}{\operatorname{argmax}} \; -\frac{1}{2} (y - w^T X)^2 . R$$

$$= \underset{w}{\operatorname{argmin}} \; (Xw - y)^2 . R$$

$$= \underset{w}{\operatorname{argmin}} \; (Xw - y)^T R \, (Xw - y) \quad \left[\text{from solution (a) \& (b)}\right]$$

∴ Maximizing the Likelihood function reduces to minimizing the Squared Loss.