Trunk 1979

May 19, 2018

Consider a two class classification prblem where class 0 is distributed N(μ_0, I) and class 1 is distributed N(μ_1, I). Let $\mu_0 = \langle 1, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{3}}, ..., \rangle$ and $\mu_1 = -\mu_0$ The likelihood functions for classes 0 and 1 can be written:

$$\lambda_0(X) = \frac{1}{\sqrt{|2\pi\Sigma_0|}} e^{\frac{-1}{2}(X-\hat{\mu})^T \Sigma_0 (X-\hat{\mu})}$$

$$\lambda_1(X) = \frac{1}{\sqrt{|2\pi\Sigma_1|}} e^{\frac{-1}{2}(X+\hat{\mu})^T \Sigma_1 (X+\hat{\mu})}$$

The decision rule of the Bayes Plugin Classifier is written:

$$\underset{i}{\text{amax}} \langle \lambda_0(X), \lambda_1(X) \rangle$$

$$\underset{i}{\text{amax}} \langle \frac{1}{\sqrt{|2\pi\Sigma_0|}} e^{\frac{-1}{2}(X-\hat{\mu})^T \Sigma_0 (X-\hat{\mu})}, \frac{1}{\sqrt{|2\pi\Sigma_1|}} e^{\frac{-1}{2}(X+\hat{\mu})^T \Sigma_1 (X+\hat{\mu})} \rangle$$

Since $\Sigma_0 = \Sigma_1 = \vec{I}$:

$$\underset{i}{\text{amax}} \langle e^{\frac{-1}{2}(X-\hat{\mu})^T \Sigma_0 (X-\hat{\mu})}, e^{\frac{-1}{2}(X+\hat{\mu})^T \Sigma_1 (X+\hat{\mu})} \rangle$$

Since ln is a monotonically increasing function, it can be applied to the maximization without changing the result:

$$\underset{i}{\text{amax}} \langle \frac{-1}{2}(X-\hat{\mu})^T \Sigma_0 (X-\hat{\mu}), \frac{-1}{2}(X+\hat{\mu})^T \Sigma_1 (X+\hat{\mu}) \rangle$$

Multiplying by a negative constant changes the maximization to a minimization:

$$\underset{i}{\text{amin}} \langle (X-\hat{\mu})^T \Sigma_0 (X-\hat{\mu}), (X+\hat{\mu})^T \Sigma_1 (X+\hat{\mu}) \rangle$$

Since the covariances are given as the identity matrix:

$$\underset{i}{\text{amin}} \langle (X-\hat{\mu})^T (X-\hat{\mu}), (X+\hat{\mu})^T (X+\hat{\mu}) \rangle$$

$$\underset{i}{\text{amin}} \langle X^T X - X^T \hat{\mu} - \hat{\mu}^T X + \hat{\mu}^T \hat{\mu}, X^T X + X^T \hat{\mu} + \hat{\mu}^T X + \hat{\mu}^T \hat{\mu} \rangle$$

Subtracting constants does not change the minimization:

$$\underset{i}{\text{amin}} \langle -X^T \hat{\mu} - \hat{\mu}^T X, X^T \hat{\mu} + \hat{\mu}^T X \rangle$$

Since $\hat{\mu}^T X$ is a scalar, $\hat{\mu}^T X = X^T \hat{\mu}$:

$$\underset{i}{\text{amin}} \langle -2X^T \hat{\mu}, 2X^T \hat{\mu} \rangle$$

$$\underset{i}{\text{amin}} \langle -X^T \hat{\mu}, X^T \hat{\mu} \rangle$$

From this form, it is easy to see that the decision rule of the classifier is:

$$g(X) = \begin{cases} 0 & X^T \hat{\mu} \geq 0 \\ 1 & X^T \hat{\mu} < 0 \end{cases}$$

Furthermore, it is easy to see that the decision rule of the optimal classifier is:

$$g^*(X) = \begin{cases} 0 & X^T \mu \geq 0 \\ 1 & X^T \mu < 0 \end{cases}$$

The assignment of the equals sign in this rule is arbitrary, since there is 0 probability mass at the point 0.

From this, we can begin solving for $\mathcal{L}_d^*$, or the Bayes Optimal loss with data of dimension d. This is expressed:

$$\begin{aligned}\mathcal{L}_d^* &= \mathcal{L}_d(g^*)\\ &= P(g^*(X_i) \neq Y_i)\forall i\\ &= P(g^*(X_i)) = P(g^*(X_i) = 0|Y_i = 1)P(Y_i = 1) + P(g^*(X_i) = 1|Y_i = 0)P(Y_i = 0)\forall i\end{aligned}$$

Since it is given that $\pi_0 = \pi_1 = \frac{1}{2}$:

$$= P(g^*(X_i)) = P(g^*(X_i) = 0|Y_i = 1)\frac{1}{2} + P(g^*(X_i) = 1|Y_i = 0)\frac{1}{2}\forall i$$

Since $f_{X|Y=1}$ is known $N(-\mu, I_d)$, where the covariance is diagonal, and $\mu_i = \frac{1}{\sqrt{i}}$we can represent the distribution of $X^T\mu|Y = 1$ as the sum:

$$\begin{aligned}f_{X^T\mu|Y=1} &= 1X_{1|Y=1} + \frac{1}{\sqrt{2}}X_{2|Y=1} + ... + \frac{1}{\sqrt{d}}X_{d|Y=1}\\ &= \sum_{i=1}^{n}\frac{1}{\sqrt{i}}X_{i|Y=1}\end{aligned}$$

Noting that:

$$\begin{aligned}E\left[\frac{1}{\sqrt{i}}X_{i|Y=1}\right] &= \frac{1}{\sqrt{i}}E\left[X_{i|Y=1}\right] = \frac{1}{\sqrt{i}}\frac{1}{\sqrt{i}} = \frac{1}{i}\\ Var\left[\frac{1}{\sqrt{i}}X_{i|Y=1}\right] &= \frac{1}{i}Var\left[X_{i|Y=1}\right] = \frac{1}{i}1 = \frac{1}{i}\end{aligned}$$

This sum can be written:

$$\begin{aligned}\sum_{i=1}^{n}\frac{1}{\sqrt{i}}X_{i|Y=1} &= \sum_{i=1}^{d}N\left(\frac{1}{i},\frac{1}{i}\right)\\ &= N\left(\sum_{i=1}^{d}\frac{1}{i},\sum_{i=1}^{d}\frac{1}{i}\right)\end{aligned}$$

From this distribution, we can derive the conditionals:

$$\begin{aligned}P(g^*(X) = 0|Y = 1) &= P(X^T\mu \geq 0|Y = 1)\\ &= 1 - P(X^T\mu < 0|Y = 1)\\ &= 1 - \Phi\left(\frac{0 - E[X^T\mu]}{\sqrt{Var[X^T\mu]}}\right)\\ &= 1 - \Phi\left(\frac{\sum_{i=1}^{d}\frac{1}{d}}{\sqrt{\sum_{i=1}^{d}\frac{1}{i}}}\right)\\ &= 1 - \Phi\left(\sqrt{\sum_{i=1}^{d}\frac{1}{i}}\right)\\ &= 1 - \int_{-\infty}^{\sqrt{\sum_{i=1}^{d}\frac{1}{i}}}\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}dz\\ &= \int_{\sqrt{\sum_{i=1}^{d}\frac{1}{i}}}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}dz\end{aligned}$$

Similarly:

$$P(g^*(X)=1|Y=0)=\int_{\sqrt{\sum_{i=1}^{d}\frac{1}{i}}}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}dz$$

Thus:

$$\begin{aligned}
P(g^*(X_i)) &= P(g^*(X_i)=0|Y_i=1)\frac{1}{2}+P(g^*(X_i)=1|Y_i=0)\frac{1}{2}\\
&= P(g^*(X_i)=0|Y_i=1)\frac{1}{2}+P(g^*(X_i)=0|Y_i=1)\frac{1}{2}\\
&= P(g^*(X_i)=0|Y_i=1)\\
&= \int_{\sqrt{\sum_{i=1}^{d}\frac{1}{i}}}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}dz
\end{aligned}$$

Looking now at:

$$\lim_{d\to\infty} L(g^*(X))$$

Since the harmonic series diverges:

$$\lim_{d\to\infty}\sqrt{\sum_{i=1}^{d}\frac{1}{i}}\to\infty$$

Using this:

$$\lim_{d\to\infty}\int_{\sqrt{\sum_{i=1}^{d}\frac{1}{i}}}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}dz\to\int_{\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}dz\to 0$$

Recall:

$$\hat{\mu}_{MLE}=\frac{1}{n}\sum_{i=1}^{n}(1-2Y_i)X_i$$

Looking at the components of one of the vectors in this sum:

$$\langle(1-2Y_i)X_{i1},...,(1-2Y_i)X_{id}\rangle$$

One can see that:

$$(1-2Y_i)X_i\sim N\left(\left[1,\frac{1}{\sqrt{2}},...,\frac{1}{\sqrt{d}}\right]^T,\vec{I}_d\right)$$

Furthermore:

$$\frac{1}{n}\sum_{i=1}^{n}(1-2Y_i)X_i\sim N\left(\left[1,\frac{1}{\sqrt{2}},...,\frac{1}{\sqrt{d}}\right]^T,\frac{1}{n}\vec{I}_d\right)$$

Using this we can find the expectation and variance of $X_i^T\hat{\mu}_{MLE}|Y_i=0$:

$$\begin{aligned}
E[X_i^T\hat{\mu}_{MLE}|Y_i=0] &= E\left[\sum_{j=1}^{d}X_{ij}\hat{\mu}_{MLE_j}|Y_i=0\right]\\
&= \sum_{j=1}^{d}E\left[X_{ij}\hat{\mu}_{MLE_j}|Y_i=0\right]\\
&= \sum_{j=1}^{d}E[X_{ij}|Y=0]E[\hat{\mu}_{MLE_j}|Y_i=0]\\
&= \sum_{j=1}^{d}\left(\frac{1}{\sqrt{j}}\right)^2\\
&= \sum_{j=1}^{d}\frac{1}{j}
\end{aligned}$$

$$
\begin{aligned}
V[X_i^T\hat{\mu}_{MLE}|Y_i=0] &= V\Bigg[\sum_{j=1}^{d} X_{ij}\hat{\mu}_{MLE_j}|Y_i=0\Bigg] \\
&= \sum_{j=1}^{d} V\Bigg[X_{ij}\hat{\mu}_{MLE_j}|Y_i=0\Bigg] + \sum_{j\neq k} C\Bigg[x_{ij}\hat{\mu}_{MLE_j}, x_{ik}\hat{\mu}_{MLE_k}|Y_i=0\Bigg] \\
&= \sum_{j=1}^{d} V\Bigg[X_{ij}\hat{\mu}_{MLE_j}|Y_i=0\Bigg] \\
&= \sum_{j=1}^{d}\Bigg(E[X_{ij}^2\hat{\mu}_{MLE_j}^2|Y_i=0] - E[X_{ij}\hat{\mu}_{MLE_j}|Y_i=0]^2\Bigg) \\
&= \sum_{j=1}^{d}\Bigg(E[X_{ij}^2|Y_i=0]E[\hat{\mu}_{MLE_j}^2|Y_i=0] - \frac{1}{j^2}\Bigg) \\
&= \sum_{j=1}^{d}\Bigg((V[X_{ij}|Y_i=0]+E[X_{ij}|Y_i=0]^2)(V[\hat{\mu}_{MLE_j}|Y_i=0]+E[\hat{\mu}_{MLE_j}|Y_i=0]^2) - \frac{1}{j^2}\Bigg) \\
&= \sum_{j=1}^{d}\Bigg((1+1/j)(1/n+1/j) - \frac{1}{j^2}\Bigg) \\
&= \sum_{j=1}^{d}\Bigg(\frac{1}{n}+\frac{1}{j}+\frac{1}{jn}\Bigg) \\
&= \sum_{j=1}^{d}\Bigg(\frac{1}{n}+\frac{1}{j}(1+1/n)\Bigg) \\
&= \frac{d}{n} + \Bigg(1+\frac{1}{n}\Bigg)\sum_{i=1}^{d}\frac{1}{j}
\end{aligned}
$$

Since the Lindeberg conditions are satisfied, the asymptotic distribution of $X_i^T\hat{\mu}_{MLE}|Y_i=0$ can be expressed:

$$
X_i^T\hat{\mu}_{MLE}|Y_i=0 \sim N(E[X_i^T\hat{\mu}_{MLE}|Y_i=0], V[X_i^T\hat{\mu}_{MLE}|Y_i=0])
$$

Recalling that the probability of making an error is written:

$$
P(g(X_i)\neq Y_i) = P(g(X_i)=0|Y_i=1)\frac{1}{2} + P(g(X_i)=1|Y_i=0)\frac{1}{2}
$$

By symmetry:

$$
\begin{aligned}
&= P(g(X_i)=1|Y_i=0)\frac{1}{2} + P(g(X_i)=1|Y_i=0)\frac{1}{2} \\
&= P(g(X_i)=1|Y_i=0) \\
&= P(X_i^T\hat{\mu}_{MLE}\geq 0|Y_i=0) \\
&= 1 - P(X_i^T\hat{\mu}_{MLE} < 0|Y_i=0) \\
&= 1 - \Phi\left(\frac{-\sum_{i=1}^{d}\frac{1}{i}}{\sqrt{(1+1/n)\sum_{i=1}^{d}1/i + d/n}}\right) \\
&= \Phi\left(\frac{\sum_{i=1}^{d}\frac{1}{i}}{\sqrt{(1+1/n)\sum_{i=1}^{d}1/i + d/n}}\right) \\
&= \int_{\frac{\sum_{i=1}^{d}\frac{1}{i}}{\sqrt{(1+1/n)\sum_{i=1}^{d}1/i+d/n}}}^{\infty} \frac{1}{\sqrt{2\pi}}e^{\frac{-z^2}{2}}
\end{aligned}
$$

We can examine the limiting behavior of this loss by first looking at the limiting behavior of the lower bound on the integral:

$$\lim_{d\to\infty} \frac{\sum_{i=1}^{d} \frac{1}{i}}{\sqrt{(1+1/n)\sum_{i=1}^{d} 1/i + d/n}}$$

Since the harmonic series is bounded by ln, we can look at the behavior of:

$$\lim_{d\to\infty} \frac{\ln(d)}{\sqrt{(1+1/n)\ln(d) + d/n}}$$

From this expression, it is easy to see that the $\frac{d}{n}$ term in the denominator dominates in the limit. Thus:

$$\lim_{d\to\infty} \frac{\sum_{i=1}^{d} \frac{1}{i}}{\sqrt{(1+1/n)\sum_{i=1}^{d} 1/i + d/n}} \to 0$$

From this

$$\lim_{d\to\infty} \int_{\frac{\sum_{i=1}^{d} \frac{1}{i}}{\sqrt{(1+1/n)\sum_{i=1}^{d} 1/i + d/n}}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} \to \int_{0}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} = \frac{1}{2}$$