Information Theory Notes

Charles Zheng and Yuval Benjamini

December 5, 2015

These are preliminary notes.

1 Classification in high-dimension, fixed SNR regime

We observe a data point y_* which belongs to one of K classes. The distribution in the ith class is $N(\mu_i, \Omega)$. We have another dataset with r repeats per class, which we use to estimate the centroids μ_i: we obtain estimates $\hat{\mu}_i \sim N(\mu_i, r^{-1}\Omega)$. The class centroids were originally drawn i.i.d. from a multivariate normal $N(0, I)$. Furthermore Ω is unknown and have to be estimated as well: assume we have obtained estimate $\hat{\Omega}$ via some method. Without loss of generality, take the Kth class to be the true class of y_*. Write $\hat{\mu}_* = \hat{\mu}_K$.

The classification rule is given by

$$\text{Estimated class} = \text{argmin}_i \ (y_* - B\hat{\mu}_i)^T A (y_* - B\hat{\mu}_i)$$

where A and B are matrices based on $\hat{\Omega}$. The Bayes rule is given by

$$A_{Bayes} = (I + \Omega - (I + r^{-1}\Omega)^{-1})^{-1}$$

$$B_{Bayes} = (I + r^{-1}\Omega)^{-1}.$$

The "plug-in" estimates of A and B are

$$A = (I + \hat{\Omega} + (I + r^{-1}\hat{\Omega})^{-1})^{-1}$$

$$B = (I + r^{-1}\hat{\Omega})^{-1}.$$

Note that

$$(y_* - B\hat{\mu}_i)^T A(y^* - B\hat{\mu}_i) = ||A^{1/2}y_* - A^{1/2}B\hat{\mu}_i||^2.$$

Therefore the classification rule is

$$\text{Estimated class} = \text{argmin}_i\ Z_i,$$

where

$$Z_i = ||A^{1/2}y_* - A^{1/2}B\hat{\mu}_i||^2.$$

We have

$$\begin{bmatrix} A^{1/2}y \\ A^{1/2}B\hat{\mu}_* \\ A^{1/2}B\hat{\mu}_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} A^{1/2}(I+\Omega)A^{1/2} & A^{1/2}BA^{1/2} & 0 \\ & A^{1/2}B(I+\frac{\Omega}{r})BA^{1/2} & 0 \\ & & A^{1/2}B(I+\frac{\Omega}{r})BA^{1/2} \end{bmatrix}\right)$$

Therefore

$$\mathbf{E}Z_i = \begin{cases} \text{tr}[A(I+\Omega+(B(I+r^{-1}\Omega)B))] & \text{for } i \neq K \\ \text{tr}[A(I+\Omega+(B(I+r^{-1}\Omega)B)-2B)] & \text{for } i = K \end{cases},$$

$$\text{Cov}(Z_i, Z_j) = \begin{cases} 2\text{tr}[A(I+\Omega)]^2 & \text{for } i \neq j \neq K \\ 2\text{tr}[A(I+\Omega-B)]^2 & \text{for } i = K, j \neq K \\ 2\text{tr}[A(I+\Omega+B(I+r^{-1}\Omega)B)]^2 & \text{for } i = j \neq K \\ 2\text{tr}[A(I+\Omega+B(I+r^{-1}\Omega)B-2B)]^2 & \text{for } i = j = K \end{cases}.$$

1.1 Ω known

Suppose

$$\hat{\Omega} = \Omega.$$

Then,

$$\mathbf{E}Z_i = \begin{cases} p + 2\text{tr}[AB] & \text{for } i \neq K \\ p & \text{for } i = K \end{cases},$$

and

$$\text{Cov}(Z_i, Z_j) = \begin{cases} 2\text{tr}[I+AB]^2 & \text{for } i \neq j \neq K \\ 2p & \text{for } i = K, j \neq K \\ \text{tr}[I+2AB]^2 & \text{for } i = j \neq K \\ 2p & \text{for } i = j = K \end{cases}.$$

Let $\gamma_i(r)$ denote the eigenvalues of AB: we have

$$\gamma_i(r) = \frac{1}{(\lambda_i^2 + \lambda_i)/r + \lambda_i}$$

where λ_i are the eigenvalues of Ω.

The misclassification probability is

$$\text{MC} = 1 - \Pr[N(\mu(r), \sigma^2(r)) < M_{K-1}]$$

where M_{K-1} is the maximum of $K-1$ independent standard normal variates, and

$$\mu(r) = \frac{2\sum_{i=1}^p \gamma_i}{\sqrt{\sum_{i=1}^p 6\gamma_i^2 + 4\gamma_i}},$$
$$\sigma^2(r) = \frac{\sum_{i=1}^p \gamma_i^2 + 2\gamma_i}{\sum_{i=1}^p 3\gamma_i^2 + 2\gamma_i}.$$

Under the condition that

$$\text{tr}[\Omega^{-1}] \to \text{const.},\ \text{tr}[\Omega^{-2}] \to 0$$

we have

$$\mu(r) \to \sqrt{\sum_{i=1}^p \gamma_i(r)},$$
$$\sigma^2(r) \to 1.$$

2 Appendix

2.1 Gaussian min probs

Define

$$f_{ng}(\mu, \sigma^2, K) = \Pr[\sigma Z_* + \mu < \max_{i=1}^K Z_i]$$

for $Z_*, Z_1, \ldots, Z_K$ i.i.d normal.

Suppose

$$\begin{bmatrix} y_* \\ y_1 \\ \vdots \\ y_{K-1} \end{bmatrix} \sim N\left(\begin{bmatrix} a \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} b & c & \ldots & c \\ c & d & \ldots & e \\ \ldots & \ldots & \ddots & \vdots \\ c & e & \ldots & d \end{bmatrix} \right).$$

where $d > e > \frac{c^2}{b}$.

Then

$$\Pr[y_* < \min_{i=1}^{K-1} y_i] = 1 - f_{ng}\left(-\frac{a}{\sqrt{d-e}}, \frac{b+e-2c}{d-e}, K-1\right).$$