# Advanced Data Analysis
# Homework Week - 12

Aswin Vijay

## Question 3

We need to prove that,

$$\begin{aligned} \operatorname{rank}\left(S^b\right) &\leq c-1 \\ S_b &= \sum_{y=1}^{c} n_y \mu_y \mu_y^T \end{aligned} \tag{1}$$

It can be also written as,

$$S_b = \sum_{y=1}^{c} n_y\left(\mu_y - \mu\right)\left(\mu_y - \mu\right)^T \tag{2}$$

where $\mu_y$ denotes the mean of training samples in class $y$. $\mu$ is $c$ is the number of classes.

$$\begin{aligned} \mu_y &= \frac{1}{n_y} \sum_{i:y_i=y} x_i \\ \mu &= \frac{1}{n} \sum_{i=1}^{n} x_i \end{aligned} \tag{3}$$

If we perform a rank analyis of Eq. 1, we see that $S_b$ is of the form $\sum \mu_y \mu_y^T$. So the rank of $S_b$ is rank of $\mu_y \mu_y^T$, where $\mu_y$ is a column vector. The sum in Eq.2 can be represented by the following matrix product,

$$\begin{aligned} S_b &= \mathrm{MM}^T \text{ where} \\ \mathrm{M} &= \sqrt{n_y}[\mu_1 - \mu, \mu_2 - \mu, ..., \mu_c - \mu] \end{aligned} \tag{4}$$

Now we use the following property,
- For a given real matrix $A$, $\operatorname{rank}(A) = \operatorname{rank}(AA^T) = \operatorname{rank}(A^T A)$
- Rank of $S_b$ is therefore rank of $M$.

Since there are $c$ classes, the column space of $M$ is contained within the $c$-dimensional space spanned by the $c$ class means. However, the class means are not all linearly independent since the overall mean $\mu$ is already in the column space of $M$ as $\mu_i - \mu$. Therefore, the maximum number of linearly independent vectors in the column space of $M$ is $c-1$.

Thus, the rank of $S^b$ is at most $c-1$. Thus proven.