Machine Learning - Theoretical exercise 6

Téo Bouvard

March 24, 2020

Problem 1

To prove this equality in a readable way, we will denote $g \equiv g\left(x; \mathcal{D}\right)$ and $F \equiv F(x)$.

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}\left[(g-F)^2\right] &= \mathbb{E}_{\mathcal{D}}\left[(g \underbrace{-\mathbb{E}_{\mathcal{D}}\left[g\right] + \mathbb{E}_{\mathcal{D}}\left[g\right]}_{=0} -F)^2\right] \\
&= \mathbb{E}_{\mathcal{D}}\left[(g-\mathbb{E}_{\mathcal{D}}\left[g\right])^2 + 2(g-\mathbb{E}_{\mathcal{D}}\left[g\right])(\mathbb{E}_{\mathcal{D}}\left[g\right]-F) + (\mathbb{E}_{\mathcal{D}}\left[g\right]-F)^2\right] \\
&= \mathbb{E}_{\mathcal{D}}\left[(g-\mathbb{E}_{\mathcal{D}}\left[g\right])^2\right] + 2\mathbb{E}_{\mathcal{D}}\left[(g-\mathbb{E}_{\mathcal{D}}\left[g\right])(\mathbb{E}_{\mathcal{D}}\left[g\right]-F)\right] + \mathbb{E}_{\mathcal{D}}\left[(\mathbb{E}_{\mathcal{D}}\left[g\right]-F)^2\right]
\end{aligned}$$

Until now, we only injected $\mathbb{E}_{\mathcal{D}}\left[g\right]$ into the equation, expanded the identity, and used the linearity of expectation property of the expectation operator to break it down in three terms. We remark that the left term is equal to the variance. The right term can be simplified because it is the expected value of a $\mathbb{E}_{\mathcal{D}}\left[g\right]-F$, which is a constant.

$$\mathbb{E}_{\mathcal{D}}\left[(g-F)^2\right] = \mathbb{E}_{\mathcal{D}}\left[(g-\mathbb{E}_{\mathcal{D}}\left[g\right])^2\right] + 2\mathbb{E}_{\mathcal{D}}\left[(g-\mathbb{E}_{\mathcal{D}}\left[g\right])(\mathbb{E}_{\mathcal{D}}\left[g\right]-F)\right] + (\mathbb{E}_{\mathcal{D}}\left[g\right]-F)^2$$

We now see that the right term is equal to the bias squared. What do we do with the central term ? As we have remarked above, $\mathbb{E}_{\mathcal{D}}\left[g\right]-F$ is a constant so we can get it out of the expectation operator.

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}\left[(g-F)^2\right] &= \mathbb{E}_{\mathcal{D}}\left[(g-\mathbb{E}_{\mathcal{D}}\left[g\right])^2\right] + 2(\mathbb{E}_{\mathcal{D}}\left[g\right]-F)\mathbb{E}_{\mathcal{D}}\left[(g-\mathbb{E}_{\mathcal{D}}\left[g\right])\right] + (\mathbb{E}_{\mathcal{D}}\left[g\right]-F)^2 \\
&= \mathbb{E}_{\mathcal{D}}\left[(g-\mathbb{E}_{\mathcal{D}}\left[g\right])^2\right] + 2(\mathbb{E}_{\mathcal{D}}\left[g\right]-F)(\mathbb{E}_{\mathcal{D}}\left[g\right]-\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\mathcal{D}}\left[g\right]\right]) + (\mathbb{E}_{\mathcal{D}}\left[g\right]-F)^2 \\
&= \mathbb{E}_{\mathcal{D}}\left[(g-\mathbb{E}_{\mathcal{D}}\left[g\right])^2\right] + 2(\mathbb{E}_{\mathcal{D}}\left[g\right]-F)\underbrace{(\mathbb{E}_{\mathcal{D}}\left[g\right]-\mathbb{E}_{\mathcal{D}}\left[g\right])}_{=0} + (\mathbb{E}_{\mathcal{D}}\left[g\right]-F)^2 \\
&= \underbrace{\mathbb{E}_{\mathcal{D}}\left[(g-\mathbb{E}_{\mathcal{D}}\left[g\right])^2\right]}_{variance} + \underbrace{(\mathbb{E}_{\mathcal{D}}\left[g\right]-F)^2}_{bias^2}
\end{aligned}$$

Problem 2

$$
\begin{aligned}
\mu_{(\cdot)} &= \frac{1}{N}\sum_{i=1}^{N}\mu_{(i)} \\
&= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{N-1}\sum_{j=1,j\neq i}^{N} x_j \\
&= \frac{1}{N}\frac{1}{N-1}\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} x_j \\
&= \frac{1}{N}\frac{1}{N-1}\sum_{i=1}^{N}\left[\left(\sum_{j=1}^{N} x_j\right) - x_i\right] \\
&= \frac{1}{N}\frac{1}{N-1}\left[N\sum_{j=1}^{N} x_j - \sum_{i=1}^{N} x_i\right] \\
&= \frac{1}{N}\frac{1}{N-1}(N-1)\sum_{i=1}^{N} x_i \\
&= \frac{1}{N}\sum_{i=1}^{N} x_i \\
&= \hat{\mu}
\end{aligned}
$$

Problem 3

For each clustering, we compute the sum of squared error J_e.

1. We first compute the centroid of each cluster.

$$
\begin{aligned}
m_1 &= \frac{1}{2}(x_1 + x_2) = \frac{1}{2}\left(\begin{bmatrix}0\\0\end{bmatrix} + \begin{bmatrix}1\\1\end{bmatrix}\right) = \begin{bmatrix}0.5\\0.5\end{bmatrix} \\
m_2 &= \frac{1}{2}(x_3 + x_4) = \frac{1}{2}\left(\begin{bmatrix}1\\0\end{bmatrix} + \begin{bmatrix}2\\0.5\end{bmatrix}\right) = \begin{bmatrix}1.5\\0.25\end{bmatrix}
\end{aligned}
$$

And then the squared error for each point.

$$
\begin{aligned}
\|x_1 - m_1\|^2 &= (0-0.5)^2 + (0-0.5)^2 = 0.5 \\
\|x_2 - m_1\|^2 &= (1-0.5)^2 + (1-0.5)^2 = 0.5 \\
\|x_3 - m_2\|^2 &= (1-1.5)^2 + (0-0.25)^2 = 0.3125 \\
\|x_4 - m_2\|^2 &= (2-1.5)^2 + (0.5-0.25)^2 = 0.3125
\end{aligned}
$$

And finally, the we compute J_e as the sum of all errors.

$$
J_e^{(1)} = 0.5 + 0.5 + 0.3125 + 0.3125 = 1.625
$$

2. We repeat the same process for the second clustering. Centroids first,

$$
\begin{aligned}
m_1 &= \frac{1}{2}(x_1 + x_4) = \frac{1}{2}\left(\begin{bmatrix}0\\0\end{bmatrix} + \begin{bmatrix}2\\0.5\end{bmatrix}\right) = \begin{bmatrix}1\\0.25\end{bmatrix}\\
m_2 &= \frac{1}{2}(x_2 + x_3) = \frac{1}{2}\left(\begin{bmatrix}1\\1\end{bmatrix} + \begin{bmatrix}1\\0\end{bmatrix}\right) = \begin{bmatrix}1\\0.5\end{bmatrix}
\end{aligned}
$$

Then individual errors,

$$
\begin{aligned}
\|x_1 - m_1\|^2 &= (0-1)^2 + (0-0.25)^2 = 1.0625\\
\|x_4 - m_1\|^2 &= (2-1)^2 + (0.5-0.25)^2 = 1.0625\\
\|x_2 - m_2\|^2 &= (1-1)^2 + (1-0.5)^2 = 0.25\\
\|x_3 - m_2\|^2 &= (1-1)^2 + (0-0.5)^2 = 0.25
\end{aligned}
$$

And sum of errors

$$
J_e^{(2)} = 1.0625 + 1.0625 + 0.25 + 0.25 = 2.625
$$

3. We repeat the same process for the third clustering. Centroids first,

$$
\begin{aligned}
m_1 &= \frac{1}{3}(x_1 + x_2 + x_3) = \frac{1}{3}\left(\begin{bmatrix}0\\0\end{bmatrix} + \begin{bmatrix}1\\1\end{bmatrix} + \begin{bmatrix}1\\0\end{bmatrix}\right) = \begin{bmatrix}\frac{2}{3}\\\frac{1}{3}\end{bmatrix}\\
m_2 &= \frac{1}{1}(x_4) = \begin{bmatrix}2\\0.5\end{bmatrix}
\end{aligned}
$$

Then individual errors,

$$
\begin{aligned}
\|x_1 - m_1\|^2 &= (0-\frac{2}{3})^2 + (0-\frac{1}{3})^2 = \frac{5}{9}\\
\|x_2 - m_1\|^2 &= (1-\frac{2}{3})^2 + (1-\frac{1}{3})^2 = \frac{5}{9}\\
\|x_3 - m_1\|^2 &= (1-\frac{2}{3})^2 + (0-\frac{1}{3})^2 = \frac{2}{9}\\
\|x_4 - m_2\|^2 &= (2-2)^2 + (0.5-0.5)^2 = 0
\end{aligned}
$$

And sum of errors

$$
J_e^{(3)} = \frac{5}{9} + \frac{5}{9} + \frac{5}{9} + 0 = \frac{5}{3}
$$

Since $\frac{5}{3} < 1.625 < 2.625$, clustering 3. is the clustering minimizing J_e.

Problem 4

a) First of all, we assign each sample to the cluster having the closest mean, which leads to the following initial partitioning.

$$
\begin{aligned}
C_1 &= \{x_3, x_4, x_6, x_7\}\\
C_2 &= \{x_1, x_2, x_5\}
\end{aligned}
$$

We then follow the algorithm by taking the random samples in the given order.

1. Sample chosen : x_1

$$\begin{aligned}\rho_1 &= \frac{4}{5}\|x_1 - m_1\|^2 = 4\\ \rho_2 &= \frac{3}{2}\|x_1 - m_2\|^2 = 6\\ \rho_1 &< \rho_2 \implies \text{transfer}\end{aligned}$$

so we transfer x_1 from C_2 to C_1

$$\begin{aligned}m_1^* &= m_1 + \frac{x_1 - m_1}{5} = \begin{bmatrix}\frac{9}{5} & \frac{13}{5}\end{bmatrix}^T\\ m_2^* &= m_2 - \frac{x_1 - m_2}{2} = \begin{bmatrix}2 & 1\end{bmatrix}^T\\ C_1 &= \{x_1, x_3, x_4, x_6, x_7\}\\ C_2 &= \{x_2, x_5\}\end{aligned}$$

2. Sample chosen : x_2

$$\begin{aligned}\rho_1 &= \frac{5}{6}\|x_2 - m_1\|^2 = \frac{10}{3}\\ \rho_2 &= \frac{2}{1}\|x_2 - m_2\|^2 = 2\\ \rho_1 &> \rho_2 \implies \text{no transfer}\end{aligned}$$

3. Sample chosen : x_3

$$\begin{aligned}\rho_1 &= \frac{5}{6}\|x_3 - m_1\|^2 = \frac{1}{6}\\ \rho_2 &= \frac{2}{1}\|x_3 - m_2\|^2 = 8\\ \rho_2 &> \rho_1 \implies \text{no transfer}\end{aligned}$$

4. Sample chosen : x_7

$$\begin{aligned}\rho_1 &= \frac{5}{6}\|x_7 - m_1\|^2 = \frac{17}{6}\\ \rho_2 &= \frac{2}{1}\|x_7 - m_2\|^2 = 20\\ \rho_2 &> \rho_1 \implies \text{no transfer}\end{aligned}$$

5. Sample chosen : x_6

$$\begin{aligned}\rho_1 &= \frac{5}{6}\|x_6 - m_1\|^2 = \frac{4}{3}\\ \rho_2 &= \frac{2}{1}\|x_6 - m_2\|^2 = 10\\ \rho_2 &> \rho_1 \implies \text{no transfer}\end{aligned}$$

6. Sample chosen : x_4

$$\begin{aligned}\rho_1 &= \frac{5}{6}\|x_4 - m_1\|^2 = \frac{85}{24}\\ \rho_2 &= \frac{2}{1}\|x_4 - m_2\|^2 = 25\\ \rho_2 &> \rho_1 \implies \text{no transfer}\end{aligned}$$

7. Sample chosen : x_5

$$\begin{aligned}\rho_1 &= \frac{5}{6}\|x_5 - m_1\|^2 = \frac{17}{12}\\ \rho_2 &= \frac{2}{1}\|x_5 - m_2\|^2 = 1\\ \rho_2 &< \rho_1 \implies \text{transfer}\end{aligned}$$

so we transfer x_5 from C_2 to C_1

$$\begin{aligned}&m_1^* = m_1 + \frac{x_5 - m_1}{6} = \begin{bmatrix}\frac{23}{12} & \frac{29}{12}\end{bmatrix}^T\\ &m_2^* = m_2 - \frac{x_5 - m_2}{1} = \begin{bmatrix}\frac{3}{2} & \frac{1}{2}\end{bmatrix}^T\\ &C_1 = \{x_1, x_2, x_3, x_4, x_6, x_7\}\\ &C_2 = \{x_2\}\end{aligned}$$

The algorithm terminates here because we cannot compute ρ_2 as it would lead to a division by zero. We get a final clustering of

$$\begin{aligned}&C_1 = \{x_1, x_2, x_3, x_4, x_6, x_7\}\\ &C_2 = \{x_2\}\end{aligned}$$

which is obviously an incorrect result. The reason this is incorrect is because we initially considered m_1 and m_2 to be the cluster means, when in reality they should have been recomputed according the initial partitioning.

b) The criterion function used is the sum of the squared errors over all clusters, i.e.

$$J = \sum_{i=1}^{M} \sum_{x \in X_i} \|x - m_i\|^2$$

with

$$m_i = \frac{1}{N} \sum_{x \in X_i} x$$

For each random sample, we only apply a transfer if the increase of the criterion function from the new cluster is lower than the decrease from the old cluster. This leads to all transfers resulting in a net decrease of the global criterion function. Since the criterion function is a positive function which is strictly decreasing, and the number of different clustering is finite ($N!$), the algorithm is guaranteed to converge.