Machine Learning - Theoretical exercise 6

Téo Bouvard

March 19, 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 simplifed 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,

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

Then individual errors,

$$\|x_1 - m_1\|^2 = (0-1)^2 + (0-0.25)^2 = 1.0625$$
$$\|x_4 - m_1\|^2 = (2-1)^2 + (0.5-0.25)^2 = 1.0625$$
$$\|x_2 - m_2\|^2 = (1-1)^2 + (1-0.5)^2 = 0.25$$
$$\|x_3 - m_2\|^2 = (1-1)^2 + (0-0.5)^2 = 0.25$$

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,

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

Then individual errors,

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

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 its closest cluster, which leads to the following repartition.

$$C_1 = \{x_3, x_4, x_6, x_7\}$$
$$C_2 = \{x_1, x_2, x_5\}$$

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

- Sample chosen : x_1