Artificial Intelligence and Machine Learning

Derivation of the Empirical Variance Formula

Let N independent and identically distributed random variables $\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_N$ be given. We assume they have mean $\mathbb{E}\{\mathcal{X}_n\} := \mu$ and variance $\mathbb{V}\{\mathcal{X}_n\} := \sigma^2$ $(1 \leq n \leq N)$. Our goal is to find an **unbiased estimator** for the variance parameter. *(The estimator $\mu^{ML} := \frac{1}{N}\sum_{n=1}^{N}\mathcal{X}_n$ for the mean – which we have derived in the lecture notes – is an unbiased estimator.)*

First, we show that the maximum likelihood estimator for the variance

$$(\sigma^2)^{\mathrm{ML}} := \frac{1}{N}\sum_{n=1}^{N}(\mathcal{X}_n - \mu^{\mathrm{ML}})^2$$

is biased. For this we determine the expected value of $(\sigma^2)^{\mathrm{ML}}$. We start by computing:

$$\mathbb{E}\left\{\sum_{n=1}^{N}(\mathcal{X}_n - \mu^{\mathrm{ML}})^2\right\} = \mathbb{E}\left\{\sum_{n=1}^{N}\left(\mathcal{X}_n^2 - 2\mathcal{X}_n\mu^{\mathrm{ML}} + (\mu^{\mathrm{ML}})^2\right)\right\}$$

[Pull sum inside]

$$= \mathbb{E}\left\{\sum_{n=1}^{N}\mathcal{X}_n^2 - 2\mu^{\mathrm{ML}}\sum_{n=1}^{N}\mathcal{X}_n + N(\mu^{\mathrm{ML}})^2\right\}$$

[Plug in the definition of μ^{ML}]

$$= \mathbb{E}\left\{\sum_{n=1}^{N}\mathcal{X}_n^2 - \frac{2}{N}\sum_{n=1}^{N}\mathcal{X}_n\sum_{n=1}^{N}\mathcal{X}_n + N\left(\frac{1}{N}\sum_{n=1}^{N}\mathcal{X}_n\right)^2\right\}$$

$$= \mathbb{E}\left\{\sum_{n=1}^{N}\mathcal{X}_n^2 - \frac{2}{N}\left(\sum_{n=1}^{N}\mathcal{X}_n\right)^2 + \frac{1}{N}\left(\sum_{n=1}^{N}\mathcal{X}_n\right)^2\right\}$$

$$= \mathbb{E}\left\{\sum_{n=1}^{N}\mathcal{X}_n^2 - \frac{1}{N}\left(\sum_{n=1}^{N}\mathcal{X}_n\right)^2\right\}$$

[Make use of the linearity of $\mathbb{E}$]

$$= \sum_{n=1}^{N}\mathbb{E}\left\{\mathcal{X}_n^2\right\} - \frac{1}{N}\mathbb{E}\left\{\left(\sum_{n=1}^{N}\mathcal{X}_n\right)^2\right\}$$

[Plug in definitions: For any random variable $\mathcal{Y}$ we have $\mathbb{V}\{\mathcal{Y}\} := \mathbb{E}\left\{\mathcal{Y}^2\right\} - \mathbb{E}\{\mathcal{Y}\}^2$. Moreover, by definition of the random variables $\mathcal{X}_n$ $(1 \leq n \leq N)$ we have that $\mathbb{E}\{\mathcal{X}_n\} := \mu$ and $\mathbb{V}\{\mathcal{X}_n\} := \sigma^2$]

$$= \sum_{n=1}^{N}(\mathbb{V}\{\mathcal{X}_n\} + \mu^2) - \frac{1}{N}\left(\mathbb{V}\left\{\sum_{n=1}^{N}\mathcal{X}_n\right\} + (N\mu)^2\right)$$

$$
\begin{aligned}
&= N(\sigma^2 + \mu^2) - \frac{1}{N}(N\sigma^2 + N^2\mu^2) \\
&= N\sigma^2 + N\mu^2 - \sigma^2 - N\mu^2 \\
&= (N-1)\sigma^2 \qquad (1)
\end{aligned}
$$

Using the result we obtained in (1) we are now able to show that the maximum likelihood estimator for the variance is biased:

$$
\mathbb{E}\Big\{(\sigma^2)^{\text{ML}}\Big\} = \mathbb{E}\left\{\frac{1}{N}\sum_{n=1}^{N}(\mathcal{X}_n - \mu^{\text{ML}})^2\right\}
$$

[Linearity of $\mathbb{E}$]

$$
\begin{aligned}
&= \frac{1}{N}\mathbb{E}\left\{\sum_{n=1}^{N}(\mathcal{X}_n - \mu^{\text{ML}})^2\right\} \\
&\overset{(1)}{=} \frac{N-1}{N}\sigma^2
\end{aligned}
$$

Since $\frac{N-1}{N} < 1$, we see that $(\sigma^2)^{\text{ML}}$ **systematically underestimates** the true variance of the data. We can correct for this bias by defining the **empirical variance** according to:

$$
\begin{aligned}
(\sigma^2)^{\text{Emp}} &:= \frac{N}{N-1}(\sigma^2)^{\text{ML}} \\
&= \frac{N}{N-1}\left(\frac{1}{N}\sum_{n=1}^{N}(\mathcal{X}_n - \mu^{\text{ML}})^2\right) \\
&= \boxed{\frac{1}{N-1}\sum_{n=1}^{N}(\mathcal{X}_n - \mu^{\text{ML}})^2} \qquad (2)
\end{aligned}
$$

Finally, let us verify that the empirical variance is indeed unbiased:

$$
\mathbb{E}\Big\{(\sigma^2)^{\text{Emp}}\Big\} = \mathbb{E}\left\{\frac{1}{N-1}\sum_{n=1}^{N}(\mathcal{X}_n - \mu^{\text{ML}})^2\right\}
$$

[Linearity of $\mathbb{E}$]

$$
\begin{aligned}
&= \frac{1}{N-1}\mathbb{E}\left\{\sum_{n=1}^{N}(\mathcal{X}_n - \mu^{\text{ML}})^2\right\} \\
&\overset{(1)}{=} \frac{1}{N-1}(N-1)\sigma^2 \\
&= \sigma^2
\end{aligned}
$$

□