W3WI DS304.1 Applied Machine Learning Fundamentals

Derivation of the Empirical Variance Formula

Let the n independent random variables $\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_n$ be given. We assume they have mean $\mathbb{E}\{\mathcal{X}_i\} := \mu$ and variance $\mathbb{V}\{\mathcal{X}_i\} := \sigma^2$ ($1 \leq i \leq n$). We aim to find an **unbiased estimator** for the variance. (The estimator $\mu^{\mathrm{ML}} := \frac{1}{n}\sum_{i=1}^{n} \mathcal{X}_i$ for the mean is already an unbiased estimator.)

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

$$\left(\sigma^2\right)^{\mathrm{ML}} := \frac{1}{n}\sum_{i=1}^{n}\left(\mathcal{X}_i - \mu^{\mathrm{ML}}\right)^2.$$

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

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

[Pull sum inside]

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

[Plug in the definition of μ^{ML}]

$$\begin{aligned}
&= \mathbb{E}\left\{\sum_{i=1}^{n}\mathcal{X}_i^2 - \frac{2}{n}\sum_{i=1}^{n}\mathcal{X}_i\sum_{i=1}^{n}\mathcal{X}_i + n\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{X}_i\right)^2\right\} \\
&= \mathbb{E}\left\{\sum_{i=1}^{n}\mathcal{X}_i^2 - \frac{2}{n}\left(\sum_{i=1}^{n}\mathcal{X}_i\right)^2 + \frac{1}{n}\left(\sum_{i=1}^{n}\mathcal{X}_i\right)^2\right\} \\
&= \mathbb{E}\left\{\sum_{i=1}^{n}\mathcal{X}_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n}\mathcal{X}_i\right)^2\right\}
\end{aligned}$$

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

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

[Definition of the variance: $\mathbb{V}\{\mathcal{X}_i\} := \mathbb{E}\{\mathcal{X}_i^2\} - \left(\mathbb{E}\{\mathcal{X}_i\}^2\right)$; $\mathbb{E}\{\mathcal{X}_i\} := \mu$; $\mathbb{V}\{\mathcal{X}_i\} := \sigma^2$]

$$\begin{aligned}
&= \sum_{i=1}^{n}\left(\mathbb{V}\{\mathcal{X}_i\} + \mu^2\right) - \frac{1}{n}\left(\mathbb{V}\left\{\sum_{i=1}^{n}\mathcal{X}_i\right\} + (n\mu)^2\right) \\
&= n\left(\sigma^2 + \mu^2\right) - \frac{1}{n}\left(n\sigma^2 + n^2\mu^2\right) \\
&= n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2 \\
&= (n-1)\sigma^2 \qquad (1)
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{E}\Big\{\big(\sigma^2\big)^{\mathrm{ML}}\Big\} &= \mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\big(\mathcal{X}_i - \mu^{\mathrm{ML}}\big)^2\right\} \\
&= \frac{1}{n}\mathbb{E}\left\{\sum_{i=1}^{n}\big(\mathcal{X}_i - \mu^{\mathrm{ML}}\big)^2\right\} \\
&\overset{(1)}{=} \frac{n-1}{n}\sigma^2
\end{aligned}$$

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

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