Bias and variance

STATS 303 Statistical Machine Learning

Spring 2022

Lecture 3

Gaussian density

$X \sim N(\mu, \sigma^2)$

$$p_X(x) = N(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- parameters : μ, σ^2

Suppose we are given a sample $\mathcal{X} = \{x_n\}_{n=1}^{N}$

$$l(\mu, \sigma^2 \mid \mathcal{X}) = p(\mathcal{X} \mid \mu, \sigma^2) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(x_n-\mu)^2}{2\sigma^2}\right)$$

$$\mathcal{L}(\mu, \sigma^2 \mid \mathcal{X}) = -\frac{N}{2}\log(2\pi) - N\log\sigma - \sum_{n=1}^{N} \frac{(x_n-\mu)^2}{2\sigma^2}$$

Gaussian density

Need to solve: $$\max_{\mu,\sigma^2} \underbrace{-\frac{N}{2}\log(2\pi) - N\log\sigma - \sum_{n=1}^{N}\frac{(x_n-\mu)^2}{2\sigma^2}}_{f(\mu,\sigma)}$$

Setting
$$\begin{cases} \dfrac{\partial f}{\partial \mu} = \displaystyle\sum_{n=1}^{N}\frac{x_n-\mu}{\sigma^2} = 0 \\ \dfrac{\partial f}{\partial \sigma} = -\dfrac{N}{\sigma} + \displaystyle\sum_{n=1}^{N}\frac{(x_n-\mu)^2}{\sigma^3} = 0 \end{cases}$$

We have $$\mu = \frac{\sum_{n=1}^{N} x_n}{N} = m \qquad \sigma^2 = \frac{\sum_{n=1}^{N}(x_n-m)^2}{N}$$

sample mean

Gaussian density

To summarize,

$$\hat{\mu}_{MLE} = m \quad \text{(sample mean)}$$

$$\hat{\sigma}^2_{MLE} = \frac{N-1}{N} S^2$$

(sample variance)

bias and variance

For simplicity, the discussion in the section is about a single parameter θ. The multi-dimensional case is a natural extension, but not required in this course. Please note that there are other topics in which we will discuss multi-dimensional cases.

mean squared error

true model

- $\mathcal{X}$: a *sample* from a population specified up to a single parameter θ. (example: μ, σ^2 in $N(\mu, \sigma^2)$)
- $d = d(\mathcal{X})$: an *estimator* of θ.
- To evaluate the quality of d, we measure $(d(\mathcal{X}) - \theta)^2$.
- The **Mean Squared Error (MSE)** of d for the parameter θ is

$$r(d, \theta) = \mathbb{E}[(d(\mathcal{X}) - \theta)^2]$$

Question: What is the source of randomness?

bias

- The **bias** of d is defined to be

$$b_\theta(d) := \mathbb{E}[d(\mathcal{X})] - \theta$$

- If $b_\theta(d) = 0$ for any θ, the estimator is said to be unbiased.
- Otherwise, it is said to be biased.

Recall: $\hat{\mu}_{MLE} = \dfrac{\sum_{n=1}^{N} x_n}{N}$ $\quad \hat{\sigma}^2_{MLE} = \dfrac{\sum_{n=1}^{N} (x_n - m)^2}{N}$

example: bias in MLE of Gaussian density

$$\mathbb{E}[\hat{\mu}_{MLE}] = \mathbb{E}\left[\frac{\sum_{n=1}^{N} x_n}{N}\right] = \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}[x_n] = \frac{N\mu}{N} = \mu$$

Therefore, $b_\mu(\hat{\mu}_{MLE}) = \mathbb{E}[\hat{\mu}_{MLE}] - \mu = 0$

That is, $\hat{\mu}_{MLE}$ is unbiased.

example: bias in MLE of Gaussian density

On the other hand,

$$\mathbb{E}[\hat{\sigma}^2_{MLE}] = \mathbb{E}\left[\frac{\sum_{n=1}^{N}(x_n - m)^2}{N}\right]$$

$$= \mathbb{E}\left[\frac{\sum_{n=1}^{N} x_n^2 - 2\left(\sum_{n=1}^{N} x_n\right)m + N m^2}{N}\right]$$

$$= \mathbb{E}\left[\frac{\sum_{n=1}^{N} x_n^2 - 2 N m^2 + N m^2}{N}\right]$$

$$= \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N} x_n^2 - m^2\right] = \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}[x_n^2] - \mathbb{E}[m^2]$$

example: bias in MLE of Gaussian density

$$\mathbb{E}[x_n^2] = (\mathbb{E}[x_n])^2 + \mathrm{Var}(x_n) = \mu^2 + \sigma^2$$

$$\mathbb{E}[m^2] = \mathbb{E}\left[\left(\frac{\sum_{n=1}^{N} x_n}{N}\right)^2\right] = \mathbb{E}\left[\frac{\left(\sum_{n=1}^{N} x_n\right)\left(\sum_{m=1}^{N} x_m\right)}{N^2}\right]$$

$$= \mathbb{E}\left[\frac{\sum_{n=1}^{N} x_n^2 + \sum_{n\neq m} x_n x_m}{N^2}\right] = \frac{\sum_{n=1}^{N} \mathbb{E}[x_n^2] + \sum_{n\neq m} \mathbb{E}[x_n]\mathbb{E}[x_m]}{N^2}$$

$$= \frac{N(\mu^2+\sigma^2) + N(N-1)\mu^2}{N^2} = \mu^2 + \frac{1}{N}\sigma^2$$

$$\mathbb{E}[\hat{\sigma}^2_{MLE}] = (\mu^2+\sigma^2) - \left(\mu^2 + \frac{1}{N}\sigma^2\right) = \frac{N-1}{N}\sigma^2$$

example: bias in MLE of Gaussian density

$$b_\sigma(\hat{\sigma}^2_{MLE}) = \mathbb{E}[\hat{\sigma}^2_{MLE}] - \sigma^2 = \frac{N-1}{N}\sigma^2 - \sigma^2 = -\frac{1}{N}\sigma^2$$

Therefore, $\hat{\sigma}^2_{MLE}$ is biased.

If we consider $\frac{N}{N-1}\hat{\sigma}^2_{MLE} = \frac{\sum_{n=1}^{N}(x_n - m)^2}{N-1} =: S^2$, (sample variance)

$$\mathbb{E}\left[\frac{N}{N-1}\hat{\sigma}^2_{MLE}\right] - \sigma^2 = \frac{N}{N-1}\cdot\frac{N-1}{N}\sigma^2 - \sigma^2 = 0$$

bias-variance formula

MSE $\quad r(d,\theta) = \mathbb{E}\left[(d-\theta)^2\right]$

$$= \mathbb{E}\left[\left((d-\mathbb{E}d)+(\mathbb{E}d-\theta)\right)^2\right]$$

$$= \mathbb{E}\left[(d-\mathbb{E}d)^2 + (\mathbb{E}d-\theta)^2 + 2(d-\mathbb{E}d)(\mathbb{E}d-\theta)\right]$$

$$= \mathbb{E}\left[(d-\mathbb{E}d)^2\right] + \mathbb{E}\left[(\mathbb{E}d-\theta)^2\right] + 2\,\mathbb{E}\left[(d-\mathbb{E}d)(\mathbb{E}d-\theta)\right]$$

bias-variance formula

(copied from previous page)

$$= \mathbb{E}\left[(d-\mathbb{E}d)^2\right] + \mathbb{E}\left[(\overbrace{\mathbb{E}d-\theta}^{\text{no randomness}})^2\right] + 2\,\mathbb{E}\left[(d-\mathbb{E}d)(\mathbb{E}d-\theta)\right]$$

$$= \mathbb{E}\left[(d-\mathbb{E}d)^2\right] + (\mathbb{E}d-\theta)^2 + 2(\mathbb{E}d-\mathbb{E}d)(\mathbb{E}d-\theta)$$

$$= \mathbb{E}\left[(d-\mathbb{E}d)^2\right] + (\mathbb{E}d-\theta)^2$$

= Variance of d + squared bias of d

score function

- Given a (log-) likelihood function $\mathcal{L}(\theta|\mathcal{X}) = \sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\theta)$, we define the **(Fisher) score function** to be

$$\mathcal{S}(\theta|\mathcal{X}) := \frac{\partial \mathcal{L}(\theta|\mathcal{X})}{\partial \theta}$$

- For MLE, $\mathcal{S}(\hat{\theta}_{\mathrm{MLE}}|\mathcal{X}) = 0$.
- Fact: $\mathbb{E}[\mathcal{S}(\theta|\mathcal{X})] = 0$ (why?).

Questions?

Reference

- *Bayesian inference:*
  - *[Al] Ch.4.4, 16.1, 16.2*
  - *[Bi] Ch.2.2.1 (for Dirichlet distribution)*
  - *[HaTF] Ch.8.3*
- *Parametric classification and regression:*
  - *[Al] Ch.4.5*