Statistic

for machine learning

Tran Trong Khiem

AI lab tranning

2024/05/29

1. The bias-variance tradeoff

Bias of an estimator

The bias of an estimator is defined as

$$\text{bias}(\hat{\theta}(\cdot)) = \mathbb{E}[\hat{\theta}(D)] - \theta^* \tag{1}$$

Where θ^* is the true parameter value. If the bias is zero, the estimator is called **unbiased**.
For example, the MLE for a Gaussian mean is unbiased :

$$\text{bias}(\hat{\mu}) = \mathbb{E}[\hat{\mu}] - \mu \qquad = \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N} x_n\right] - \mu = 0 \tag{2}$$

where $\bar{x}$ is the sample mean.
The MLE for a Gaussian variance is given by $\sigma^2_{\text{mle}} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})^2$, is not an unbiased estimator of σ^2.

$$\mathbb{E}\left[\sigma^2_{\text{mle}}\right] = \frac{N-1}{N}\sigma^2$$

Variance of an estimator

We define the variance of an estimator as follows:

$$V(\hat{\theta}) = \mathbb{E}[\hat{\theta}^2] - \left(\mathbb{E}[\hat{\theta}]\right)^2 \tag{3}$$

where the expectation is taken with respect to $p(D|\theta^*)$.
=>This measures how much our estimate will change as the data changes.
We would like the **variance of our estimator to be as small as possible**
Cramer-Rao lower bound, provides a **lower bound on the variance of any unbiased estimator.**
Let $X_1, \ldots, X_N \sim p(X|\theta^*)$ and $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_N)$ be an unbiased estimator of θ^*.Then, under various smoothness assumptions on $p(X|\theta^*)$, we have

$$V(\hat{\theta}) \geq \frac{1}{N\mathcal{F}(\theta^*)} \tag{4}$$

where $\mathcal{F}(\theta^*)$ is the Fisher information.
MLE achieves the Cramer Rao lower bound, and hence has the smallest asymptotic variance of any unbiased estimator. Thus MLE is said to be asymptotically optimal.

The bias-variance tradeoff

Assuming our goal is to minimize the mean squared error (MSE), $\hat{\theta} = \hat{\theta}(D)$ denote the estimate, $\bar{\theta} = E[\hat{\theta}(D)]$ denote the expected value of estimate (vary D).

$$\mathbb{E}\left[(\hat{\theta} - \theta^*)^2\right] = \mathbb{E}\left[(\hat{\theta} - \bar{\theta}) + (\bar{\theta} - \theta^*)\right]^2 \tag{5}$$

$$= \mathbb{E}\left[(\hat{\theta} - \bar{\theta})^2 + 2(\bar{\theta} - \theta^*)(\hat{\theta} - \bar{\theta}) + (\bar{\theta} - \theta^*)^2\right] \tag{6}$$

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

$$= \mathbb{E}\left[(\hat{\theta} - \bar{\theta})^2\right] + (\bar{\theta} - \theta^*)^2 \tag{7}$$

$$= \mathrm{Var}(\hat{\theta}) + \mathrm{bias}^2(\hat{\theta}) \tag{8}$$

This called **bias-variance tradeoff**

$$\mathrm{MSE} = \mathrm{variance} + \mathrm{bias}^2 \tag{9}$$

MAP estimator for a Gaussian mean

Suppose we want to estimate the mean of a Gaussian from $\mathbf{x} = (x_1, \ldots, x_N)$. Assume the data is sampled from $x_n \sim \mathcal{N}(\theta^* = 1, \sigma^2)$.We have :

$$\mathbb{V}[\bar{x}|\theta^*] = \frac{\sigma^2}{N}$$

The MAP estimate under a Gaussian prior of the form $\mathcal{N}(\theta_0, \sigma^2/\kappa_0)$ is given by

$$\tilde{x} = \frac{N}{N+\kappa_0}\bar{x} + \frac{\kappa_0}{N+\kappa_0}\theta_0 = w\bar{x} + (1-w)\theta_0 \tag{10}$$

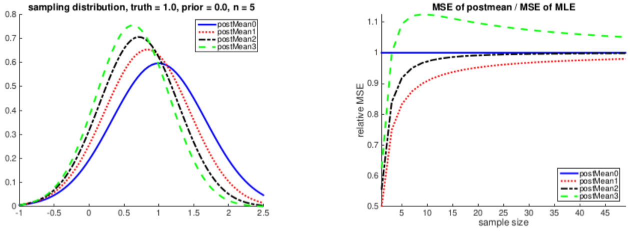
where $0 \leq w \leq 1$ controls how much we trust the MLE compared to our prior.The bias and variance are given by

$$\begin{aligned} \mathbb{E}[\tilde{x}] - \theta^* &= w\theta^* + (1-w)\theta_0 - \theta^* && (11) \\ &= (1-w)(\theta_0 - \theta^*) && (12) \end{aligned}$$

$$V[\tilde{x}] = w^2 \frac{\sigma^2}{N} \tag{13}$$

MAP estimator for a Gaussian mean

- The MAP estimate is biased (assuming $w < 1$), it has lower variance.
- Left: Sampling distribution of the MAP estimate (equivalent to the posterior mean) under a $\mathcal{N}(\theta_0 = 0, \sigma^2/\kappa_0)$ prior with different prior strengths κ_0.
- Right: plot $\frac{\text{MSE}(\tilde{x})}{\text{MSE}(\bar{x})}$ vs N. We see that the MAP estimate has lower MSE than the MLE for $\kappa_0 \in \{1, 2\}$.

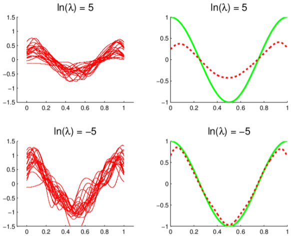


(a) (b)

MAP estimator for linear regression

MAP estimation for linear regression under a Gaussian prior, $p(w) = \mathcal{N}(w|0, \lambda^{-1}I)$.
The zero-mean prior encourages the weights to be small, which reduces overfitting
λ, controls the strength of this prior.

- $\lambda = 0$ MAP become MLE
- $\lambda > 0$ results in a biased estimate
- We see that as we increase the strength of the regularizer, the variance decreases, but the bias increases



MAP estimator for linear regression