Parametric Methods (cont.)

CAI 5107: Machine Learning

Instructor: Anowarul Kabir

Email: akabir@usf.edu

Fall 2025

Evaluating an Estimator: Bias & Variance

- Let, *X* be a sample from a population parameterized by θ , and *d=d(X)* be an estimator of θ
- So, MSE is: $r(d, \theta) = E[(d(\mathcal{X}) - \theta)^2]$
- The bias of an estimator is defined as: $b_\theta(d) = E[d(\mathcal{X})] - \theta$
- Unbiased estimator of θ , if $b_\theta(d) = 0$ for all θ values

Is sample average an unbiased estimator of a density mean?

- Assume, x_t is drawn from known population desnsity mean μ , m is the sample average, then

$$E[m] = E\left[\frac{\sum_t x^t}{N}\right] = \frac{1}{N}\sum_t E[x^t] = \frac{N\mu}{N} = \mu$$

Is sample variance an unbiased estimator of population variance?

- The MLE of σ^2 is s^2.

$$
\begin{aligned}
s^2 &= \frac{\sum_t (x^t - m)^2}{N} = \frac{\sum_t (x^t)^2 - Nm^2}{N} \\
E[s^2] &= \frac{\sum_t E[(x^t)^2] - N \cdot E[m^2]}{N}
\end{aligned}
$$

- Given that $\mathrm{Var}(X) = E[X^2] - E[X]^2$, we get $E[X^2] = \mathrm{Var}(X) + E[X]^2$

$$E[(x^t)^2] = \sigma^2 + \mu^2 \text{ and } E[m^2] = \sigma^2/N + \mu^2$$

$$E[s^2] = \frac{N(\sigma^2 + \mu^2) - N(\sigma^2/N + \mu^2)}{N} = \left(\frac{N-1}{N}\right)\sigma^2 \neq \sigma^2$$

Model vs. Variance and Bias

$$
\begin{aligned}
r(d,\theta) &= E\left[(d-\theta)^2\right] \\
&= E\left[(d-E[d]+E[d]-\theta)^2\right] \\
&= E\left[(d-E[d])^2+(E[d]-\theta)^2+2(E[d]-\theta)(d-E[d])\right] \\
&= E\left[(d-E[d])^2\right]+E\left[(E[d]-\theta)^2\right]+2E\left[(E[d]-\theta)(d-E[d])\right] \\
&= E\left[(d-E[d])^2\right]+(E[d]-\theta)^2+2(E[d]-\theta)E[d-E[d]] \\
&= \underbrace{E\left[(d-E[d])^2\right]}_{variance}+\underbrace{(E[d]-\theta)^2}_{bias^2}
\end{aligned}
$$

$$r(d,\theta)=\mathrm{Var}(d)+(b_\theta(d))^2$$

Variance and Bias

- **Variance**: measures how much, on average, d_i vary around the expected value (from one dataset to another)
- **Bias**: measures how much the expected value varies from the correct value of θ

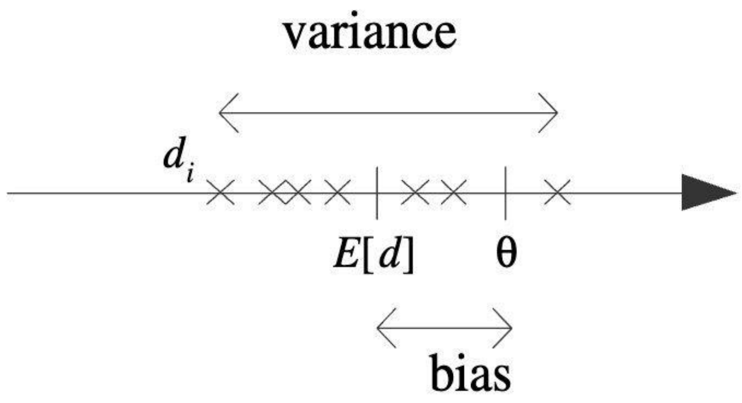

Figure 4.1 θ is the parameter to be estimated. d_i are several estimates (denoted by '×') over different samples $\mathcal{X}_i$. Bias is the difference between the expected value of d and θ. Variance is how much d_i are scattered around the expected value. We would like both to be small.

Maximum a posteriori (MAP)

- Given a sample *X*
  - We want to choose a parameter θ for *X*

- Goal:

$$\theta_{map} = arg\max_{\theta} p(\theta|X)$$

- As opposed to MLE:

$$\theta_{mle} = arg\max_{\theta} p(X|\theta)$$

MAP vs. MLE

- MAP is equivalated to MLE:
  - If no prior information of θ is known.
  - $p(\theta)$ is uniform
- MAP helps to reduce overfitting
  - Regularization
- MLE and MAP both are point estimation
  - They loose information unless posterior is unimodal and makes a narrow pick around the points

Baye's Estimator

- The expected value of the posterior density

$$\theta_{Bayes} = E[\theta|\mathcal{X}] = \int \theta p(\theta|\mathcal{X})d\theta$$

- The best estimate of a random variable is its mean.

- For normal density, the mode is the expected value. Then

$$\theta_{Bayes} = \theta_{MAP}$$

Discussion

- When θ is nomally distributed with known parameters
  - and samples are normally distributed with unknown θ and known variance
  - Baye's estimator is a weighted average of known prior and sample mean
- As sample size *N* increases, Baye's estimator gets closer to the sample average
- When *N* is small, prior guess has higher effect

Reading materials

- Chapter 4: Parametric Methods
  - 4.3 - 4.4