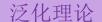


§1.1.1 一致性Consistency

@ 滕佳烨

[ref] Statistical Inference, George Casella and Roger L. Berger.





Recall:

Generalization $\mathcal{L}(\hat{f}) - \hat{\mathcal{L}}(\hat{f})$: Measuring how model performs on *unseen* data.

View 1: small generalization gap $(\mathcal{L}(\hat{f}) - \hat{\mathcal{L}}(\hat{f}))$ and small training error $(\hat{\mathcal{L}}(\hat{f}))$ leads to small population risk.

View 2: small generalization gap leads to small excess risk.

```
No-free-lunch, PAC, agnostic PAC...
```

In this section (§1), we mainly analyze the parametric model and study the **consistency** of the estimator.

- What is consistency?
- Why we need consistency?
- How to prove consistency?



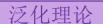
Parametric Model

(Informal) We call a model as parametric model if the learned model (or, the distribution) is indexed by a **finite-dimension** parameter.

For example, linear settings, $y|x \sim N(x^{\top}\beta^*, \sigma^2)$ is indexed by $\beta^* \in \mathbb{R}^d$ (ground truth) Use the linear classifier (indexed by $\hat{\beta}$)

Consistency

If we train the model using the training set and the algorithm returns $\hat{\beta}$, when we have infinite training samples $(n \to \infty)$, the estimator $\hat{\beta}$ should **converge** to the true parameter β^* .





Consistency

If we train the model using the training set and the algorithm returns $\hat{\beta}$, when we have infinite training samples $(n \to \infty)$, the estimator $\hat{\beta}$ should **converge** to the true parameter β^* .

Why we need consistency?

When we collect more data points, we expect the model to be *more precise*. When we have infinite data points, we expect the model to be *perfect*.





How to prove consistency?

We take **linear regression** as an example (with least-square estimator).

Consider the linear regimes $y = x^{T}\beta^{*} + \epsilon$, where $\epsilon \sim N(0, \sigma^{2})$. Fixed design regime. The estimator can be derived as

$$\hat{\beta} = (X^{\top}X)^{-1}X^{\top}Y,$$

where $X \in \mathbb{R}^{n \times d}$ is the design matrix, $Y \in \mathbb{R}^n$ is the response vector (n: training samples, d: dimension).

[Remark: we usually define *d* as the dimension of x, and *p* as the dimension of parameter. In linear regression regimes, d = p.]

Hint: for least-square estimator,

We want the estimator $\hat{\beta}$ helps to map from "x-space" to "y-space". Therefore, $\hat{\beta} = X^{-1}Y$. Since $X \in \mathbb{R}^{n \times d}$, X^{-1} denotes its pseudo inverse.

For a formal script, you can minimize the loss $L(\beta) = ||Y - X\beta||^2$. By taking derivation, we have $\frac{\partial L(\beta)}{\partial \beta} = X^{\top}(Y - X\beta) = 0$.



How to prove consistency?

We take **linear regression** as an example (with least-square estimator). Consider the linear regimes $y = x^{T}\beta^{*} + \epsilon$, where $\epsilon \sim N(0, \sigma^{2})$. Fixed design regime. The estimator can be derived as

$$\hat{\beta} = (X^{\top}X)^{-1}X^{\top}Y,$$

where $X \in \mathbb{R}^{n \times d}$ is the design matrix, $Y \in \mathbb{R}^n$ is the response vector (n: training samples, d: dimension).

The learned parameter:

$$\hat{\beta}|X \sim N(\beta^*, \sigma^2(X^\top X)^{-1})$$

Hint: $\mathbb{E}\hat{\beta}|X = (X^{\top}X)^{-1}X^{\top}\mathbb{E}Y|X = (X^{\top}X)^{-1}X^{\top}X\beta^* = \beta^*;$ $\mathbb{D}\hat{\beta}|X = (X^{\top}X)^{-1}X^{\top}[\mathbb{D}Y|X]X(X^{\top}X)^{-1} = \sigma^2(X^{\top}X)^{-1}.$

Note that $(X^{\top}X)^{-1} = \frac{1}{n} \left(\frac{1}{n} X^{\top}X\right)^{-1} \to \frac{1}{n} \Sigma_x^{-1}$ (informally), it converges to zero as $n \to \infty$. Therefore, it is consistent!



Take-away messages

(a) Basic definition on consistency.

泛化理论

- (b) With infinite samples, we hope that the estimator converge to the ground truth.
- (c) Basic framework about the linear regression. The estimator is consistent.

All the slides will be available at <u>www.tengjiaye.com/generalization</u> soon.

@ 滕佳烨

@滕佳烨

Thanks!