泛化理论

第一章 传统统计模型

§1.1.2 岭回归

@ 滕佳烨

Recall:

Consistency: when we have infinite training samples ($n \to \infty$), the estimator $\hat{\beta}$ should converge to the true parameter β^*.

- Basic linear regression: $y = x^\top \beta^* + \epsilon$
- LSE (least-square estimator): minimize MSE (min square error) $||Y - X\beta||^2$

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

Today's topic:

1. Bias-variance tradeoff
2. Ridge regression

Bias-variance tradeoff

We consider the term $\mathbb{E}\left\|\hat{\beta}-\beta^*\right\|^2$ which measures the distance between $\hat{\beta}$ and β^*.
(The expectation is taken over the randomness of Y. We still consider fixed design.)
It can be split as the bias and variance component

$$\mathbb{E}\left\|\hat{\beta}-\beta^*\right\|^2=\left\|\mathbb{E}\hat{\beta}-\beta^*\right\|^2+\mathbb{E}\left\|\hat{\beta}-\mathbb{E}\hat{\beta}\right\|^2$$

For LSE estimator $\hat{\beta}=(X^\top X)^{-1}X^\top Y$.

- It is unbiased: $\mathbb{E}\hat{\beta}|X=\beta^*$
- Its variance is $\mathbb{E}\left\|\hat{\beta}-\mathbb{E}\hat{\beta}\right\|^2|X=Trace[\sigma^2(X^\top X)^{-1}]$
- LSE estimator has the smallest variance among all the unbiased estimator (MVUE, minimal variance unbiased estimator).

Question: when the eigenvalues of $(X^\top X)$ is small, the estimator has large variance!
Can we find a biased estimator with small variance, while controlling the bias?

Intuition for Ridge Regression

LSE estimator: $\hat{\beta} = (X^\top X)^{-1} X^\top Y.$

- The variance term $\sigma^2 (X^\top X)^{-1}$ may be too large due to small eigenvalues of $(X^\top X)$.
- The variance term is stemmed from $(X^\top X)^{-1}$ in $\hat{\beta}$
- Boost the eigenvalue of $(X^\top X)$!

Ridge regression estimator: $\hat{\beta}_r(\rho) = (X^\top X + \rho I)^{-1} X^\top Y.$

- The minimal eigenvalue is at least ρ
- The estimator is biased: $\mathbb{E}\hat{\beta}_r | X \neq \beta^*$
- The variance term can be smaller than $\hat{\beta}$

It can be proved that under some mild assumptions, there are some ρ^* such that

$$\mathbb{E}\left\|\hat{\beta}_r(\rho^*) - \beta^*\right\|^2 < \mathbb{E}\left\|\hat{\beta} - \beta^*\right\|^2$$

Hint: prove that the derivation $\frac{d\left\|\hat{\beta}_r(\rho) - \beta^*\right\|^2}{d\rho}|_{\rho=0} < 0$. Note that $\hat{\beta}_r(0) = \hat{\beta}$.

Another intuition for Ridge Regression

1. The corresponding loss for Ridge regression:

$$L(\beta) = \frac{1}{n}\|Y - X\beta\|^2 + \rho\|\beta\|^2$$

Penalty on β! Prior: beta is not too large.

2. Computational stability: $\hat{\beta} = (X^\top X)^{-1} X^\top Y.$

It is not stable to compute $(X^\top X)^{-1}$ with small eigenvalues…

Take-away messages

(a) Bias-variance tradeoff.
(b) Ridge regression: bias but less variance.
(c) Some intuition behind ridge regression: reduce the variance; prior; stable.

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

@ 滕佳烨

Thanks!