Statistics Learning Theory:Linear Regression

2019.03.02

Linear Regression

1. Domain set $\mathcal{X} \in R^d$
2. Label set $\mathcal{Y}$ is the set of real number
3. Hypothesis class

$$\mathcal{H}_{reg} = L_d = \{x \to \langle w, x \rangle + b : w \in R^d, b \in R\}$$

4. loss function

$$l(h, (x, y)) = (h(x) - y)^2$$

5. Empirical Risk function

$$L_S(h) = \frac{1}{m} \sum_{i=1}^{m} (h(x_i) - y_i)^2$$

Linear Regression: Implementation of ERM rule

Least square is the algorithm that solves the ERM problem for hypothesis class of linear regression predictors with respect to the squared loss.
The ERM problem can be written as

$$arg \min_{w} L_S(h_w) = arg \min_{w} \frac{1}{m} \sum_{i=1}^{m} (\langle w, x_i \rangle - y_i)^2$$

Linear Regression: Closed form solution for ERM rule

To solve the problem we calculate the gradient of the objective function

$$\frac{2}{m}\sum_{i=1}^{m}(\langle w, x_i\rangle - y_i)x_i = 0$$

We can rewrite the problem as the problem $Aw = b$ where

$$A = (\sum_{i=1}^{m} x_i x_i^{\intercal}) \quad \text{and} \quad b = \sum_{i=1}^{m} y_i x_i$$

If A is invertible then the solution to the ERM problem is

$$w = A^{-1}b$$

Linear Regression: ERM rule

If A is not invertible, we still can find the solution to the system $Aw = b$ since b is in the range of A.
To be specific, since A is symmetric we can decompose A as

$$A = VDV^\intercal$$

where D is a diagonal matrix and V is an orthonormal matrix.
After normalizing D we can obtain

$$A^+ = VD^+V^\intercal \quad \text{and} \quad \hat{w} = A^+b$$

Let v_i denote the i'th column of V. Then, we have

$$A\hat{w} = AA^+b = VDV^\intercal VD^+V^\intercal b = VDD^+V^\intercal b = \sum_{i:D_{i,i}\neq 0} v_i v_i^\intercal b$$

Linear Regression: Consistency Theorem

Let

$$Q_n(\theta) = \text{some objective function}$$

$$\hat{\theta} = arg \max_{\theta \in \Theta} Q_n(\theta)$$

Examples:

1. NLLS : $Q_n(\theta) = -\frac{1}{n} \sum_{i=1}^{n} (y_i - m(x_i, \theta))^2$
2. ML : $Q_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log f(z_i, \theta)$
3. GMM : $Q_n(\theta) = -(\frac{1}{n} \sum_{i=1}^{n} g(z_i, \theta))' W (\frac{1}{n} \sum_{i=1}^{n} g(z_i, \theta))$

Linear Regression: Consistency Theorem

Theorem (General Consistency Theorem)

Suppose

1. Θ *is compact*
2. $\sup_{\theta \in \Theta} |Q_n(\theta) - Q_*(\theta)| \to^p 0$ *for some* $Q_* : \Theta \to R$
3. Q_* *is continuous in* $\theta \in \Theta$
4. Q_* *is uniquely maximized at* θ_0

Then

$$\hat{\theta} \to^p \theta_0$$

Linear Regression: Consistency Theorem Proof

Pick $\varepsilon > 0$, since $\hat{\theta}$ maximizes $Q_n(\theta)$

$$Q_n(\hat{\theta}) > Q_n(\theta_0) - \frac{\varepsilon}{3}$$

By condition 2, for any $\theta \in \Theta$

$$|Q_n(\theta) - Q_*(\theta)| < \frac{\varepsilon}{3}$$

with probability approaching one

Linear Regression: Consistency Theorem Proof

Thus with probability approaching one

$$Q_n(\hat{\theta}) - Q_*(\hat{\theta}) < \frac{\varepsilon}{3}$$

$$Q_*(\theta_0) - Q_n(\theta_0) < \frac{\varepsilon}{3}$$

Combining these inequalities

$$Q_*(\hat{\theta}) + \frac{\varepsilon}{3} > Q_n(\hat{\theta})$$

$$> Q_n(\theta_0) - \frac{\varepsilon}{3}$$

$$> Q_*(\theta_0) - \frac{2\varepsilon}{3}$$

Therefore,

$$Q_*(\hat{\theta}) > Q_*(\theta_0) - \varepsilon$$

Linear Regression: Consistency Theorem Proof

Since we want to prove that $\hat{\theta} \to^p \theta_0$, we want to show that

$$Pr\{\hat{\theta} \in \mathcal{N}\} \to 1$$

for any open set $\mathcal{N} \subset \Theta$ containing θ_0
Now pick any open set $\mathcal{N}$ containing θ_0, since $\mathcal{N}$ is open, $\mathcal{N}^c$ is closed. By condition 1, Θ is compact, $\Theta \cap \mathcal{N}^c$ is also compact. Since Q_* is continuous, Weierstrass theorem guarantees there exists $\theta_* \in \Theta \cap \mathcal{N}^c$ such that

$$\sup_{\Theta \cap \mathcal{N}^c} Q_*(\theta) = Q_*(\theta_*)$$

Linear Regression: Consistency Theorem Proof

Since Q_* is uniquely maximized at θ_0, we have

$$Q_*(\theta_0) > Q_*(\theta_*)$$

and set

$$\varepsilon^{'} = Q_*(\theta_0) - Q_*(\theta_*) > 0$$

Using the previous inequality and set $\varepsilon = \varepsilon^{'}$

$$Q_*(\hat{\theta}) > Q_*(\theta_0) - \varepsilon^{'}$$

$$= Q_*(\theta_*)$$

$$= \sup_{\Theta \cap \mathcal{N}^c} Q_*(\theta)$$

This means that

$$\hat{\theta} \in \mathcal{N}$$

with probability approaching one

Linear Regression: Consistency Theorem Remark

1. For each application, most efforts are devoted to check Conditions 2 and 4
2. Condition 4 is called identification condition. If $Q_*(\theta)$ is maximized at multiple points, we cannot tell where $\hat{\theta}$ converges in general
3. Condition 2 sats that objective function $Q_n(\theta)$ uniformly converges in probability to the limit objective function $Q_*(\theta)$
4. For condtion 2, we typically need some kind of uniform law of large numbers

Linear Regression: Uniform law of large number

Theorem

1. Θ *is compact*
2. $g(z,\theta)$ *is almost surely continuous at each* $\theta \in \Theta$
3. *There is* $d(z)$ *such that* $|g(z,\theta) \leq d(z)$ *for all* $\theta \in \Theta$ *and almost every* z *and* $E[d(z)] < \infty$

Then

$$\sup_{\theta \in \Theta} |\bar{g}(\theta) - E[g(z,\theta)]| \to^{p} 0$$

Linear Regression: Consistency of NLLSE

Theorem

Suppose

1. $\{(y_i, x_i')\}_{i=1}^n$ *is iid and* $E[y|x] = m(x,\theta)$ *almost surely only if* $\theta = \theta_0$
2. Θ *is compact*
3. $m(x,\theta)$ *is almost surely continuous at each* $\theta \in \Theta$
4. $E[y^2] < \infty$ *and* $E[\sup_{\theta\in\Theta} |m(x,\theta)|^2] < \infty$

Then

$$\hat{\theta} \rightarrow^p \theta_0$$

Linear Regression: Consistency of NLLSE

It is sufficient to check condition $1-4$ in the general theorem.
Condition 1 is guaranteed by our second condition.
Condition 2: $Q_*(\theta) = -E[\{y - m(x,\theta)\}^2]$, we want to show that

$$\sup_{\theta\in\Theta} |\frac{1}{n}\sum_{i=1}^{n}\{y_i - m(x_i,\theta)\}^2 - E[\{y - m(x,\theta)\}^2]| \to^p 0$$

The above is holded by applying the ULLN. Since ULLN also guarantees continuity of $Q_*(\theta)$. Thus condition 3 is also satisfied.

Linear Regression: Consistency of NLLSE

It remains to check condition 4, identification of θ_0. i.e.

$$Q_*(\theta) = -E[\{y - m(x, \theta)\}^2]$$

is uniquely maximized at θ_0. Since $m(x) = E[y|x]$ solves

$$\min_g E[\{y - g(x)\}^2]$$