STA 314: Statistical Methods for Machine Learning I

Lecture 2 - Linear regression

Xin Bing

Department of Statistical Sciences
University of Toronto

Review

- Supervised learning is about to estimate (learn) f under the generating mechanism

$$Y = f(X) + \epsilon$$

- For a given estimate $\hat{f}$ of f, we have learned
    - how to evaluate it
    - and its expected MSE follows the bias-variance decomposition

- Different $\hat{f}$'s are different algorithms/methods/predictors:
    - parametric methods
    - non-parametric methods

Linear regression

Let $Y \in \mathbb{R}$ be the outcome and $X \in \mathbb{R}^p$ be the (random) vector of p features.

The linear model assumes

$$\begin{aligned} Y &= f(x) + \epsilon \\ &= \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon \qquad \text{(linearity of } f\text{)} \end{aligned}$$

where:

- $\beta_0, \beta_1, \cdots, \beta_p$ are unknown constants.
  - β_0 is called the **intercept**
  - β_j, for $1 \le j \le p$, are the **coefficients** or **parameters** of the p features
- ϵ is the error term satisfying $\mathbb{E}[\epsilon] = 0$.

Linear predictor under the linear regression model

Given some estimates $\hat{\beta}_j$ of β_j for $0 \leq j \leq p$, we predict the response at any $X = \mathbf{x}$ by the linear predictor

$$\hat{y} := \hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p.$$

Question: how to choose $\hat{\beta}_0, \ldots, \hat{\beta}_p$?

Ordinary Least Squares approach (OLS)

Recall that we want to find a function g by

$$\min_g \mathbb{E}\left[(Y - g(X))^2\right].$$

Under linear model, it suffices to find $\alpha_0, \ldots, \alpha_p$ by

$$\min_{\alpha_0,\ldots,\alpha_p} \mathbb{E}\left[\left(Y - \alpha_0 - \alpha_1 X_1 - \cdots - \alpha_p X_p\right)^2\right]$$

In the **model fitting step**, we use the training data to approximate the above expectation (w.r.t. X and Y).

Specifically, given $\mathcal{D}^{train}$, we choose $\hat{\beta}_0, \cdots, \hat{\beta}_p$ by

$$(\hat{\beta}_0, \cdots, \hat{\beta}_p) = \operatorname*{argmin}_{\alpha_0,\cdots,\alpha_p} \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \alpha_0 - \alpha_1 x_{i1} - \cdots - \alpha_p x_{ip}\right)^2.$$

Ordinary Least Squares approach (OLS)

Using the matrix notation,

- $\hat{\boldsymbol{\beta}} = [\hat{\beta}_0, \cdots, \hat{\beta}_p]^\top \in \mathbb{R}^{p+1}$, $\boldsymbol{\beta} = [\beta_0, \cdots, \beta_p]^\top \in \mathbb{R}^{p+1}$, $\boldsymbol{\alpha} = [\alpha_0, \cdots, \alpha_p]^\top \in \mathbb{R}^{p+1}$,
- $\mathbf{y} = [y_1, \ldots, y_n]^\top \in \mathbb{R}^n$, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^\top \in \mathbb{R}^{n \times (p+1)}$ with $\mathbf{x}_i = [1, x_{i1}, \ldots, x_{ip}]^\top \in \mathbb{R}^{p+1}$ for $1 \leq i \leq n$.

the OLS estimator of $\boldsymbol{\beta}$ is defined as

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \underset{\boldsymbol{\alpha} \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \ \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^\top \boldsymbol{\alpha} \right)^2 \\ &= \underset{\boldsymbol{\alpha} \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \ \frac{1}{n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}\|_2^2.\end{aligned}$$

Comments

- The idea of estimating f by minimizing the training MSE can be applied to (almost) all supervised learning problems. Specifically,

  $$\hat{f} = \operatorname*{argmin}_{g \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} L(y_i, g(\mathbf{x}_i)) \tag{1}$$

  where $L(\cdot, \cdot)$ is a loss function and $\mathcal{F}$ is a class of choices of g.
- The OLS approach corresponds to $L(a, b) = (a - b)^2$ and $\mathcal{F}(x) = \{\mathbf{x}^\top \boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^{p+1}\}$.
- In general, the difficulty of solving (1) varies across problems. But, the OLS approach admits a closed-form solution:

  $$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

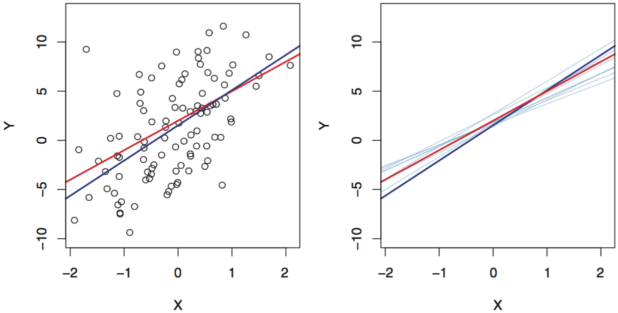
  whenever $\mathbf{X} \in \mathbb{R}^{n \times (p+1)}$ has full column rank.

What is random and what is not random?

- In Statistics, we use capitalized letters for generic random variables (e.g. X and Y).
- The parameters such as $\beta_1, \ldots, \beta_p$ or the function $f : \mathcal{X} \to \mathcal{Y}$ are treated as deterministic (non-random). Of course, being Bayesian is an exception.
- The data points (x_i, y_i) for $1 \leq i \leq n$ are actual values, observed in practice. They can be thought as the realizations of random variables (X_i, Y_i) for $1 \leq i \leq n$.
- When we talk about estimators (e.g. the OLS estimator) which, by definition, are functions of (X_i, Y_i), hence are random.
- Nevertheless, we will not distinguish between (x_i, y_i) and (X_i, Y_i) throughout the lecture, but you should have in mind that the training data (x_i, y_i) are random realizations.

The randomness in $\hat{\beta}_0$ and $\hat{\beta}_1$

We cannot hope $\hat{\beta}_0 = \beta_0$ and $\hat{\beta}_1 = \beta_1$, because they depend on the observed data which is random.



Left: The **red line** represents the true model $f(X) = 2 + 3X$. The **blue line** is the OLS fit based on the observed data.
Right: The light blue lines represent 10 OLS fits, each one computed on the basis of a different training dataset.
The fitted OLS lines are different, but their average is close to the true regression line.

Some important considerations

- Estimation of $\boldsymbol{\beta}$:
  - How close is the point estimation $\hat{\boldsymbol{\beta}}$ to $\boldsymbol{\beta}$?
  - (Inference) Can we provide confidence interval / conduct hypothesis testing of $\boldsymbol{\beta}$?

- Prediction of Y at $X = \mathbf{x}$:
  - How accurate is the point prediction $\hat{y} = \mathbf{x}^\top \hat{\boldsymbol{\beta}}$?
  - (Inference) Can we further provide confidence interval of Y?

- Variable (Model) selection:
  - Do all the predictors help to explain Y, or is there only a subset of the predictors useful?
  - Later in Lecture 3

Property of $\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}$

Take the *design matrix* $\mathbf{X} \in \mathbb{R}^{n\times(p+1)}$ to be deterministic with full column rank. Assume $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. realizations of ϵ.

- Unbiasedness: $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$
- The covariance matrix of $\hat{\boldsymbol{\beta}}$ is:

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}$$

- The above two properties imply the ℓ_2 estimation error

$$\mathbb{E}\left[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2\right] = \sigma^2 \text{trace}\left[(\mathbf{X}^\top \mathbf{X})^{-1}\right]$$

When $\mathbf{X}^\top \mathbf{X} = n\ \mathbf{I}_{p+1}$,

$$\mathbb{E}\left[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2\right] = \frac{\sigma^2(p+1)}{n}.$$

The MSE of estimating β increases as p gets larger.