Statistics 101C

Shirong Xu

University of California, Los Angeles

shirong@stat.ucla.edu

October 1, 2024

Two Main Problems

We observe a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, $\mathbf{x}_i \in \mathbb{R}^p$ is p-dimensional predictors. By the type of response Y, there are two **learning problems**:

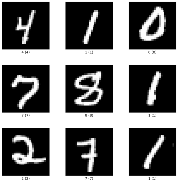
- **Regression**: The response Y is quantitative. For example, people's income, the value of a house, blood pressure of patient.
- **Classification**: The response Y is qualitative: binary (gender, like or dislike a product), categorical (brand of a product), and ordinal (ratings given by users to movies or restaurant)

Examples

Regression:

Years of Experience	Salary in 1000$
2	15
3	28
5	42
13	64
8	50
16	90
11	58
1	8
9	54

Classification (Categorical):



Description:
images (28×28 pixel grayscale images) from the MNIST dataset of handwritten digits.

Objective: Predict the number (categorical 0-9) based on the pixel values (28×28).

Example - Ordinal Classification

- **Description**: A review in Yelp community with textual data (covariates) and a rating (1-5, response).
- **Objective**: In Yelp challenge, the goal is to train a classifier predict the rating value based on the textual comment of users.

Example - Binary Classification

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1

- **Description**: Diabetes dataset contains observations with diagnostic measurements and binary response indicating whether a patient has diabetes.
- **Objective**: Predict based on diagnostic measurements whether a patient has diabetes.

Statistical Learning for regression

- Background
- Training and test mean squared errors (MSEs)
- Bias-variance trade-off

Statistical Learning for regression: Background

- **Predictors**: $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$ is p-dimensional random variable
- **Response**: Y is a quantitative random variable. Generally, Y is something we want to predict.
- **The relationship between $\boldsymbol{X}$ and Y**:

$$Y = f^*(\boldsymbol{X}) + \epsilon,$$

where $\mathbb{E}(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2$. Here $f^*(\boldsymbol{X}) = \mathbb{E}(Y|\boldsymbol{X})$.

Statistical Learning for regression: Background

- **Goal**: Find a function $f(\boldsymbol{X})$ for predicting Y (or approximate f^* well)
- **Question**: How do we assess the quality of $f(\boldsymbol{X})$ in predicting Y

Statistical Learning for regression: Background

- **Goal**: Find a function $f(\boldsymbol{X})$ for predicting Y (or approximate f^* well)
- **Question**: How do we assess the quality of $f(\boldsymbol{X})$ in predicting Y
- **Loss function**: square loss

$$L(f(\boldsymbol{X}), Y) = (Y - f(\boldsymbol{X}))^2$$

Statistical Learning for regression: Background

- **Goal**: Find a function $f(\boldsymbol{X})$ for predicting Y (or approximate f^* well)
- **Question**: How do we assess the quality of $f(\boldsymbol{X})$ in predicting Y
- **Loss function**: square loss

$$L(f(\boldsymbol{X}), Y) = (Y - f(\boldsymbol{X}))^2$$

- The averaged loss (expected error) of f:

$$R(f) = \mathbb{E}\big[L(f(\boldsymbol{X}), Y)\big] = \mathbb{E}\big[(Y - f(\boldsymbol{X}))^2\big]$$

Statistical Learning for regression: Background

- The expected squared loss (risk) can be written as

$$R(f) = \mathbb{E}\big[(Y - f(\boldsymbol{X}))^2\big] = \int\int (Y - f(\boldsymbol{X}))^2 \mathbb{P}(\boldsymbol{X}, Y) d\boldsymbol{X} dY.$$

- We can decompose $R(f)$ into

$$\begin{aligned}\mathbb{E}\big[(Y - f(\boldsymbol{X}))^2\big] = &\int\int (Y - \mathbb{E}(Y|\boldsymbol{X}))^2 \mathbb{P}(\boldsymbol{X}, Y) d\boldsymbol{X} dY \\ &+ \int\int (\mathbb{E}(Y|\boldsymbol{X}) - f(\boldsymbol{X}))^2 \mathbb{P}(\boldsymbol{X}, Y) d\boldsymbol{X} dY,\end{aligned}$$

where $\mathbb{P}(\boldsymbol{X}, Y)$ is the joint distribution of $(\boldsymbol{X}, Y)$.

Statistical Learning for regression: Background

- The expected squared loss (risk) can be written as

$$R(f) = \mathbb{E}\left[(Y - f(\boldsymbol{X}))^2\right] = \int\int (Y - f(\boldsymbol{X}))^2 \mathbb{P}(\boldsymbol{X}, Y) d\boldsymbol{X} dY.$$

- We can decompose $R(f)$ into

$$\begin{aligned}\mathbb{E}\left[(Y - f(\boldsymbol{X}))^2\right] = &\int\int (Y - \mathbb{E}(Y|\boldsymbol{X}))^2 \mathbb{P}(\boldsymbol{X}, Y) d\boldsymbol{X} dY \\ &+ \int\int (\mathbb{E}(Y|\boldsymbol{X}) - f(\boldsymbol{X}))^2 \mathbb{P}(\boldsymbol{X}, Y) d\boldsymbol{X} dY,\end{aligned}$$

where $\mathbb{P}(\boldsymbol{X}, Y)$ is the joint distribution of $(\boldsymbol{X}, Y)$.
- $R(f)$ attains its minimum at $f(\boldsymbol{X}) = \mathbb{E}(Y|\boldsymbol{X})$.
- If you have $\mathbb{E}(Y|\boldsymbol{X})$, you're done. Since you already have the "best" function.

Statistical Learning for regression: Background

- In practice, we do not know the exact from of $\mathbb{E}(Y|\boldsymbol{X})$.
- **Question**: What do we usually do?

Statistical Learning for regression: Background

- In practice, we do not know the exact from of $\mathbb{E}(Y|\boldsymbol{X})$.
- **Question**: What do we usually do?
  - Impose a structure on f, for example
$$f(\boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p.$$
  - Suppose a function class
$$\mathcal{F} = \{f(\boldsymbol{x}) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i : \beta_i \in \mathbb{R}, i = 0, \ldots, p\}$$
.
  - Minimize the averaged squared loss on training dataset
$$\widehat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (f(\boldsymbol{x}_i) - y_i)^2$$

Statistical Learning for regression: Bias-Variance tradeoff

- Based on the training dataset $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, we obtain an estimator $\widehat{f}$

$$\widehat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (f(\boldsymbol{x}_i) - y_i)^2$$

Statistical Learning for regression: Bias-Variance tradeoff

- Based on the training dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, we obtain an estimator $\widehat{f}$

$$\widehat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (f(\mathbf{x}_i) - y_i)^2$$

- Suppose for a new data point $\mathbf{x}_0$ (testing step), we aim to predict its y, the quality of $\widehat{f}$ at $\boldsymbol{X} = \mathbf{x}_0$:

$$\begin{aligned}
&\mathbb{E}\big[(\widehat{f}(\boldsymbol{X}) - Y)^2 | \boldsymbol{X} = \mathbf{x}_0\big] \\
=&\big[\widehat{f}(\mathbf{x}_0) - \mathbb{E}(Y|\boldsymbol{X} = \mathbf{x}_0)\big]^2 + \mathbb{E}\big[Y - \mathbb{E}(Y|\boldsymbol{X} = \mathbf{x}_0)|\boldsymbol{X} = \mathbf{x}_0\big]^2 \\
=&\underbrace{\big[\widehat{f}(\mathbf{x}_0) - \mathbb{E}(Y|\boldsymbol{X} = \mathbf{x}_0)\big]^2}_{Reducible} + \underbrace{\sigma^2}_{non-reducible}
\end{aligned}$$

Statistical Learning for regression: Bias-Variance tradeoff

- Reducible part can be decomposed into two components

$$
\begin{aligned}
&\mathbb{E}\left[\widehat{f}(x_0) - \mathbb{E}(Y|\boldsymbol{X} = x_0)\right]^2 \\
=&\underbrace{\mathbb{E}\left[\widehat{f}(x_0) - \mathbb{E}(\widehat{f}(\boldsymbol{X}))\right]^2}_{Variance} + \underbrace{\left[\mathbb{E}(\widehat{f}(x_0)) - \mathbb{E}(Y|\boldsymbol{X} = x_0)\right]^2}_{Bias^2},
\end{aligned}
$$

where the expectation is taken with respect to what?

Statistical Learning for regression: Bias-Variance tradeoff

- Reducible part can be decomposed into two components

$$\mathbb{E}\left[\widehat{f}(x_0) - \mathbb{E}(Y|\boldsymbol{X} = x_0)\right]^2$$
$$= \underbrace{\mathbb{E}\left[\widehat{f}(x_0) - \mathbb{E}(\widehat{f}(\boldsymbol{X}))\right]^2}_{Variance} + \underbrace{\left[\mathbb{E}(\widehat{f}(x_0)) - \mathbb{E}(Y|\boldsymbol{X} = x_0)\right]^2}_{Bias^2},$$

where the expectation is taken with respect to what?

  - **Variance**: represents the variability of the predicted value. The randomness comes from the training dataset.
  - **Squared Bias**: The second term is the squared bias. If $\mathcal{F}$ is chosen well, so that the mean across all training data sets is the true function, then bias is 0.

Training MSE v.s. Testing MSE

- Let $D_r = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ and $D_e = \{(\boldsymbol{x}'_i, y'_i)\}_{i=1}^m$ be training and testing datasets, respectively. Train an estimator from D_r

$$\widehat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (f(\boldsymbol{x}_i) - y_i)^2$$

- Evaluate $\widehat{f}$ by the mean squared error (MSE):

$$\text{Training MSE} : \frac{1}{n} \sum_{i=1}^{n} (\widehat{f}(\boldsymbol{x}_i) - y_i)^2$$

$$\text{Testing MSE} : \frac{1}{m} \sum_{i=1}^{m} (\widehat{f}(\boldsymbol{x}'_i) - y'_i)^2$$

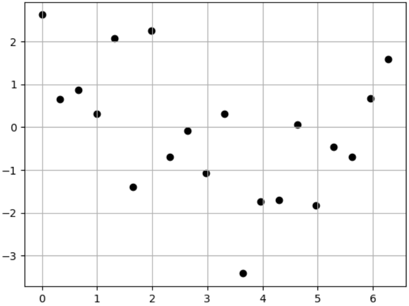
- **Question**: Which one can be used for assessing the quality of $\widehat{f}$?

An example.

- We generate $\{(x_i, y_i)\}_{i=1}^n$ in the following way

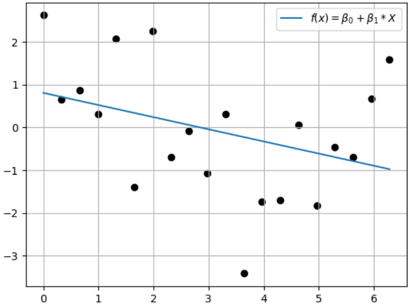
$$y_i = sin(x_i) + cos(x_i) + \epsilon_i$$

  - $x_i \sim \mathrm{Unif}(0, 2\pi)$
  - $\epsilon \sim N(0, 1)$
  - Set $n = 20$



An example: Model 1

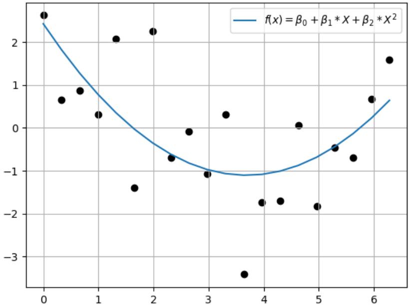
- We fit a linear model $f(x) = \beta_0 + \beta_1 x$



- Training MSE is 1.9918
- Testing MSE is 1.6304

An example: Model 2

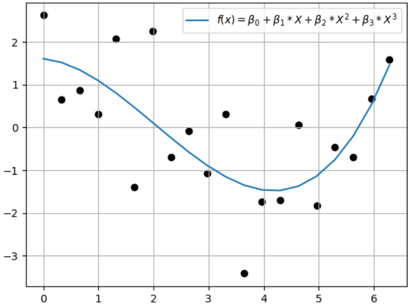
- We fit a linear model $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$



- Training MSE is 1.2848
- Testing MSE is 1.2837

An example: Model 3

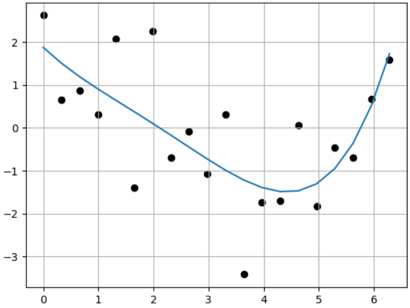
- We fit a linear model $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$



- Training MSE is 1.1101
- Testing MSE is 1.1374

An example: Model 4

- We fit a linear model $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4$



- Training MSE is 1.0883
- Testing MSE is 1.1924

An example: Conclusion

Metrics	Model 1	Model 2	Model 3	Model 4
Training MSE	1.9918	1.2848	1.1101	1.0883
Testing MSE	1.6304	1.2837	1.1374	1.1924

- **Conclusions**:
  (1) Training MSE is **non-increasing** with respect to the flexibility of model, i.e., as training model $\mathcal{F}$ becomes more flexible, training MSE always becomes smaller.
  (2) Testing MSE decreases first and then increases with respect to the flexibility of model.
- **The behavior of Testing MSE**: Bias-variance trade-off
  (1) Models with greater flexibility have a smaller bias.
  (2) More flexible methods have a greater variance

Taylor Expansion

In the previous example, $Y = sin(X) + cos(X) + \epsilon$.

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots$$

$$\cos(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots$$

$$f(x) = \left(x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots\right) + \left(1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots\right)$$

Therefore, if we consider a model with polynomial terms with higher degrees, we are closer to the truth. But it does not mean we achieve higher performance on the testing. Why?

Another example: Bias-variance tradeoff

1. We generate 1000 datasets: the j-th dataset is $D_j = \{(x_i^{(j)}, y_i^{(j)})\}_{i=1}^{30}$

$$y_i^{(j)} = sin(x_i^{(j)}) + cos(x_i^{(j)}) + \epsilon_i,$$

   where $x_i \in \text{Unif}(-2\pi, 2\pi)$.

2. Consider polynomial model with degree $d = 1, 2, \ldots, 7$,

$$\mathcal{F}_d = \{f(x) = \beta_0 + \sum_{i=1}^{d} \beta_i x_i : \beta_i \in \mathbb{R}, i = 0, \ldots, d\}$$

3. Estimate $\widehat{f}$ in 1,000 replications

$$\widehat{f}^{(j)} = \arg\min_{f \in \mathcal{F}_d} \sum_{i=1}^{30} (f(x_i^{(j)}) - y_i^{(j)})^2$$

4. Generate 50,000 testing samples $\{(x_i', y_i')\}_{i=1}^{50{,}000}$:

$$y_i' = sin(x_i') + cos(x_i')$$

Another example: Bias-variance tradeoff

5 Estimate the Bias:

$$\text{Estimate of Bias} : \frac{1}{50000}\sum_{i=1}^{50000}(\bar{f}(x_i') - y_i')^2,$$

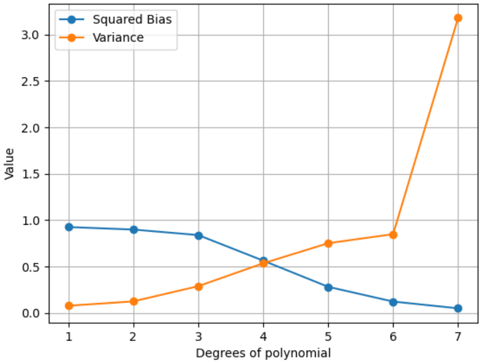
where $\bar{f}(x_i') = \frac{1}{1000}\sum_{j=1}^{1000}\widehat{f}^{(j)}(x_i')$

6 Estimate the variance:

$$\text{Estimate of Variance} : \frac{1}{50000}\sum_{i=1}^{50000}\left(\frac{1}{1000}\sum_{j=1}^{1000}(\widehat{f}^{(j)}(x_i') - \bar{f}(x_i'))^2\right),$$

where $\bar{f}(x_i') = \frac{1}{1000}\sum_{j=1}^{1000}\widehat{f}^{(j)}(x_i')$

Another example: Bias-variance tradeoff



- Degrees of polynomial increases ⇒ Model becomes more flexible ⇒ Squared Bias decreases
- Degrees of polynomial increases ⇒ Model becomes more flexible ⇒ Variance increases

Take home messages

- A more flexible function class is not always preferred. In practice, a "medium" model usually has higher performance in predicting unobserved samples (testing data)

- In real-life situation f is unobserved, it is impossible to compute the bias and variance of an estimated function. Nevertheless, we should always keep the bias-variance tradeoff in mind.

- The bias-variance tradeoff point depends on the sample size.

Assignment 1: Part 1

- 1 Reproduce the Bias-Variance plot in Page 23 (codes).
- 2 Give the explaination (for the example from Pages 14-20). The last model

$$\mathcal{F}_4 = \{f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4\}$$

approximates the ground truth model $f(x) = \sin(x) + \cos(x)$ better. But the testing performance is worse than the third one, that is

$$\mathcal{F}_3 = \{f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3\}.$$