Introduction to Statistical Learning

Tianhang Zheng

https://tianzheng4.github.io

Course Website

https://tianzheng4.github.io/umkc-teaching/2023-fall-teaching-1/

What are on the course website:

- Lecture slides
- Lab material
- Contact Information

If you are interested in my research, feel free to contact me.

Notations (Supervised Learning)

Vector/Matrix/Tensor predictor X (also called inputs, regressors, covariates, features, independent variables)

Outcome Y (also called dependent variable, response, target)

Objectives:

1. Accurately predict the outcomes of unseen test cases
2. Understand which inputs affect the outcome, and how
3. Assess the quality of our predictions and inferences

Notations (Supervised Learning)

Task: Predict the income based on years of education, years of work, etc.

Outcome Y is Income

Predictors are years of education, years of work, etc. (Denoted by X)

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

How to assess a model

Prediction Error (regression problems)

Prediction Accuracy (classification problems)

Model Variance

Interpretability

Prediction Error

Given a training dataset D_{tr} and a testing dataset D_{te}, and a prediction function learned on D_{tr} (i.e., $g(X; D_{tr})$), the prediction error can be defined as

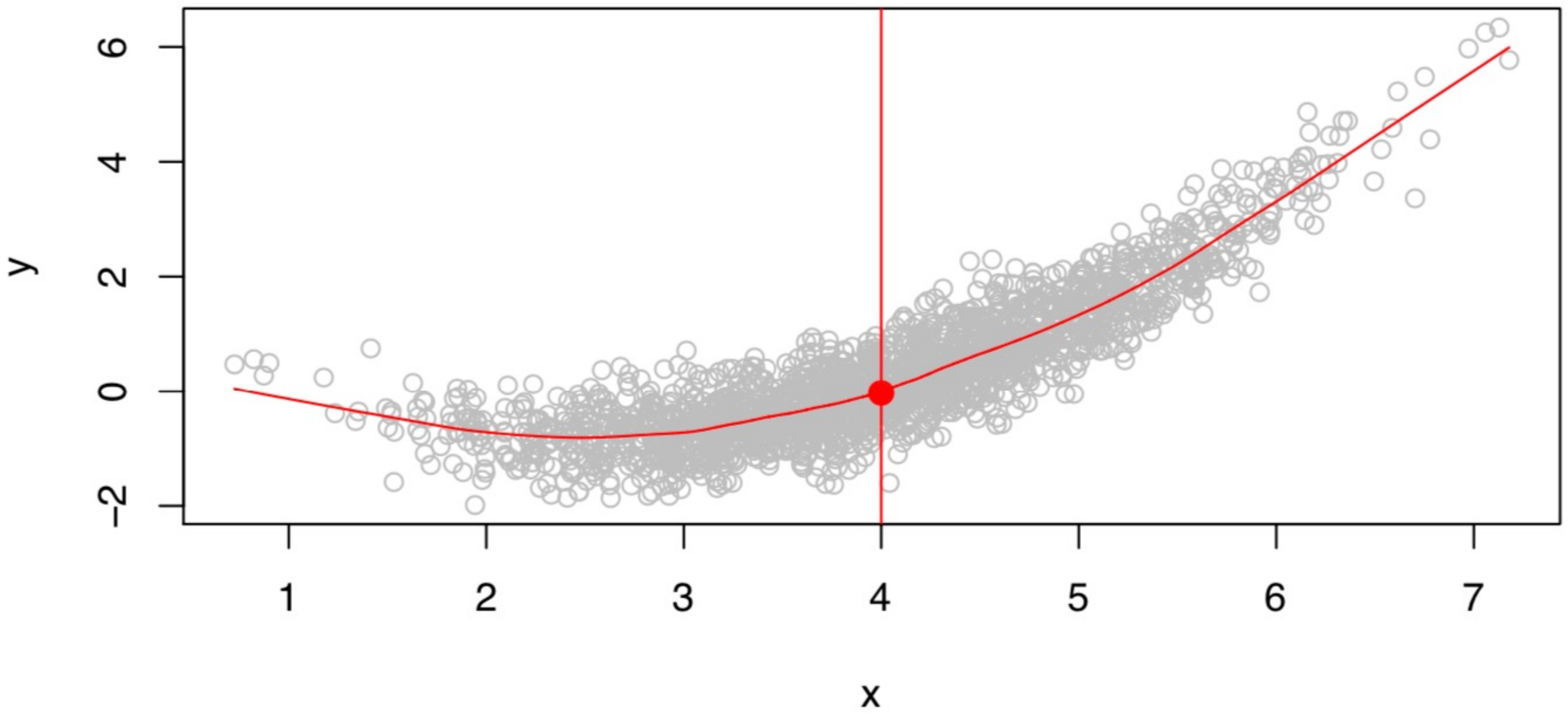
$$E_D[\left(Y - g(X; D_{tr})\right)^2]$$

Training Error: $MSE_{tr} = \frac{1}{|D_{tr}|}\sum_{\{(x,y)\in D_{tr}\}}[\left(y - g(x; D_{tr})\right)^2]$

Testing Error: $MSE_{te} = \frac{1}{|D_{te}|}\sum_{\{(x,y)\in D_{te}\}}[\left(y - g(x; D_{tr})\right)^2]$

What is an ideal model?

Given X, there may be multiple outcomes Y due to different ϵ.



What is an ideal model?

The ideal model characterizes the expectation of the outcome.

$$f(X) = E(Y|X)$$

The ideal model here is also called the regression function.

If X is a vector, $X = [X_1, X_2]$

$$f(x) = E(Y|X_1 = x_1, X_2 = x_2)$$

Decompose the Prediction Error

$$E_D\left[\left(Y - g(X)\right)^2\right] = E_D[\,((\underbrace{Y - f(X))}_{\epsilon} + (\underbrace{f(X) - g(X;D))}_{e})^2]$$

$$E_D[\,(\epsilon + e)^2] = E_D[\,\epsilon^2 + 2\epsilon e + e^2]$$

$$E_D[\,(\epsilon + e)^2] = E_D[\,\epsilon^2] + E_D[2\epsilon e] + E_D[e^2] = \epsilon^2 + 2\epsilon E_D[e] + E_D[e^2]$$

$E_D[\,(\epsilon + e)^2] = \epsilon^2 + 2\epsilon E_D[e] + E_D[e^2]$ irreducible error + reducible error

Decompose the Prediction Error

$$E_D\left[\left(Y - g(X)\right)^2\right] = bias^2 + variance + \sigma^2$$

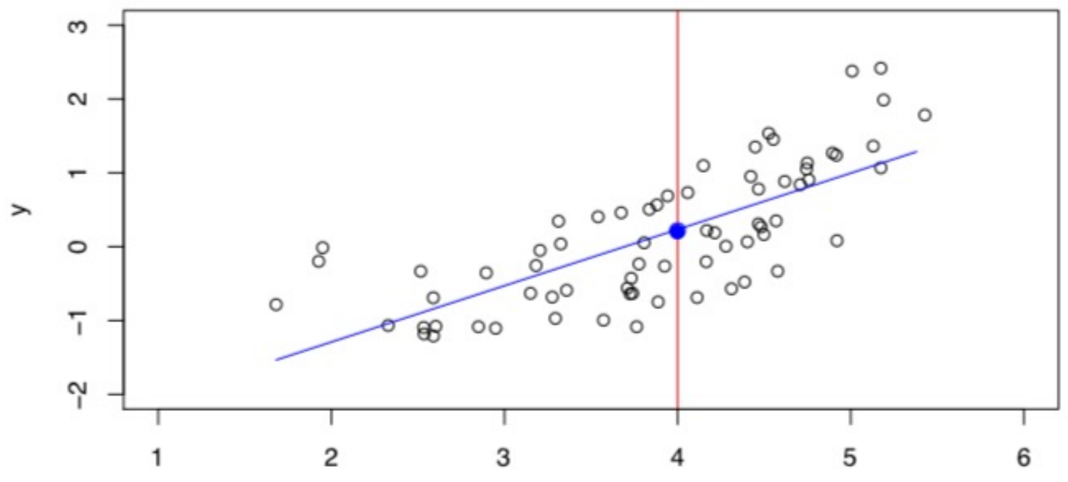
$$\left.\begin{array}{ll} \text{Bias:} & E_D[g(X;D)] - f(X) \\ \\ \text{Variance:} & \mathrm{E_D}[(\mathrm{E}_D[g(X;D)] - g(X;D))^2] \end{array}\right\} \text{Depends on model complexity}$$

Irreducible error: $\epsilon\ or\ \sigma$

Regression Models (to Estimate g(X))

Linear Models:

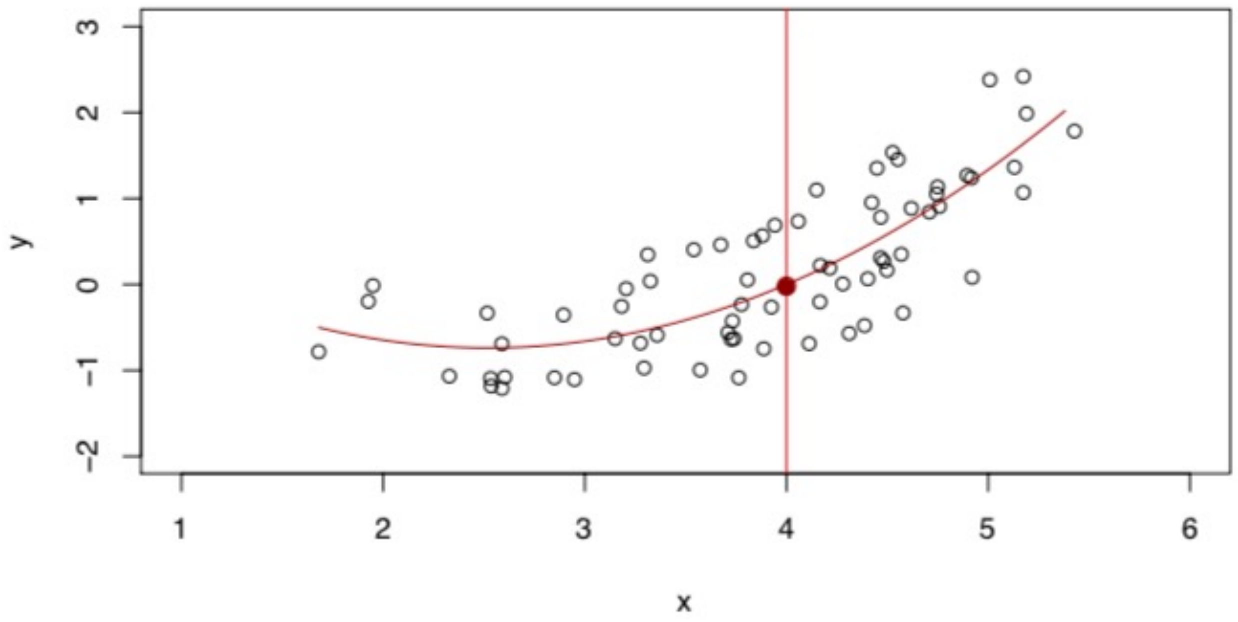
$$g(X) = \beta_0 + \beta_1 X_1 + \ \ldots + \beta_p X_p \qquad (X = [X_1, \ldots, X_p])$$



Regression Models (to Estimate g(X))

Quadratic Models:

$$g(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 \qquad (X = [X_1])$$



Prediction Accuracy

Given a training dataset D_{tr} and a testing dataset D_{te}, and a prediction function learned on D_{tr} (i.e., $g(X; D_{tr})$), the prediction accuracy can be defined as

$$Acc = E_D[I(Y = g(X; D_{tr})]$$

Training Accuracy: $Acc_{tr} = \frac{1}{|D_{tr}|} \sum_{\{(x,y) \in D_{tr}\}} [I(y = g(x; D_{tr})]$

Testing Accuracy: $Acc_{te} = \frac{1}{|D_{te}|} \sum_{\{(x,y) \in D_{te}\}} [I(y = g(x; D_{tr})]$

Classification Models

The output of classification models are labels

- Logistic Regression Models
- Support Vector Machine
- K-Nearest Neighbors

Model Bias and Variance

Model Bias $E_D[g(X;D)] - f(X)$

Model Variance: The variance of parameters (but what is the standard?)

Better metric: $\mathrm{E_D}[(\mathrm{E}_D[g(X;D)] - g(X;D))^2]$

Bias and Variance Trade-off

If a model is more complicated (e.g., with more parameters):

The bias is expected to be smaller

The variance is expected to be larger

Interpretability

The importance of each predictor?

Why a model makes a particular decision?

Example: Linear Model

$$Income = 5 \times Year\ of\ work + 4 \times Year\ of\ Edu + 0.1 \times Height$$

Accuracy and Interpretability Trade-off

If a model is more complicated (e.g., with more parameters):

The accuracy may be higher (but may suffer from overfitting)

The interpretability may be worse (It is easy to interpret linear models)

Other Metrics (Binary Classification)

A + B: B is the prediction, and A means the correctness of the prediction

True Positive: TP

False Positive: FP

TP + FP = 1

True Negative: TN

False Negative: FN

TN + FN = 1

Other Metrics (Binary Classification)

$$\text{Precision} = \frac{TP}{TP + FP}$$

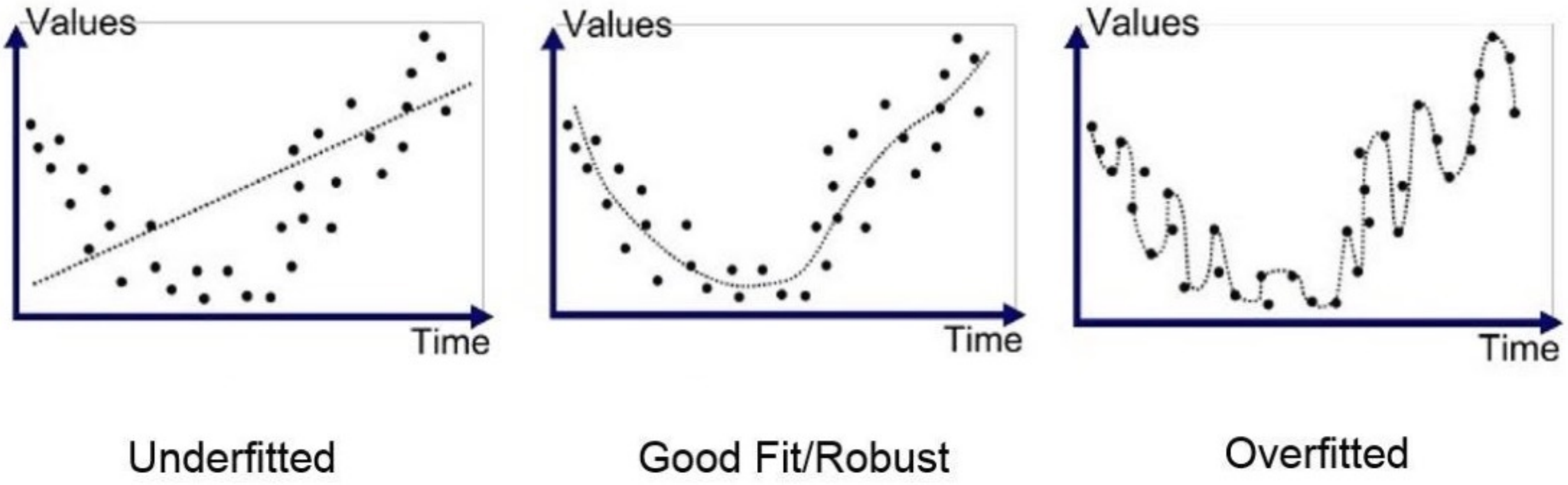
How many positive predictions are correct?

$$\text{Recall} = \frac{TP}{TP + FN}$$

How many positive cases are correctly predicted?

$$F_1 = \frac{2}{\text{recall}^{-1} + \text{precision}^{-1}}$$

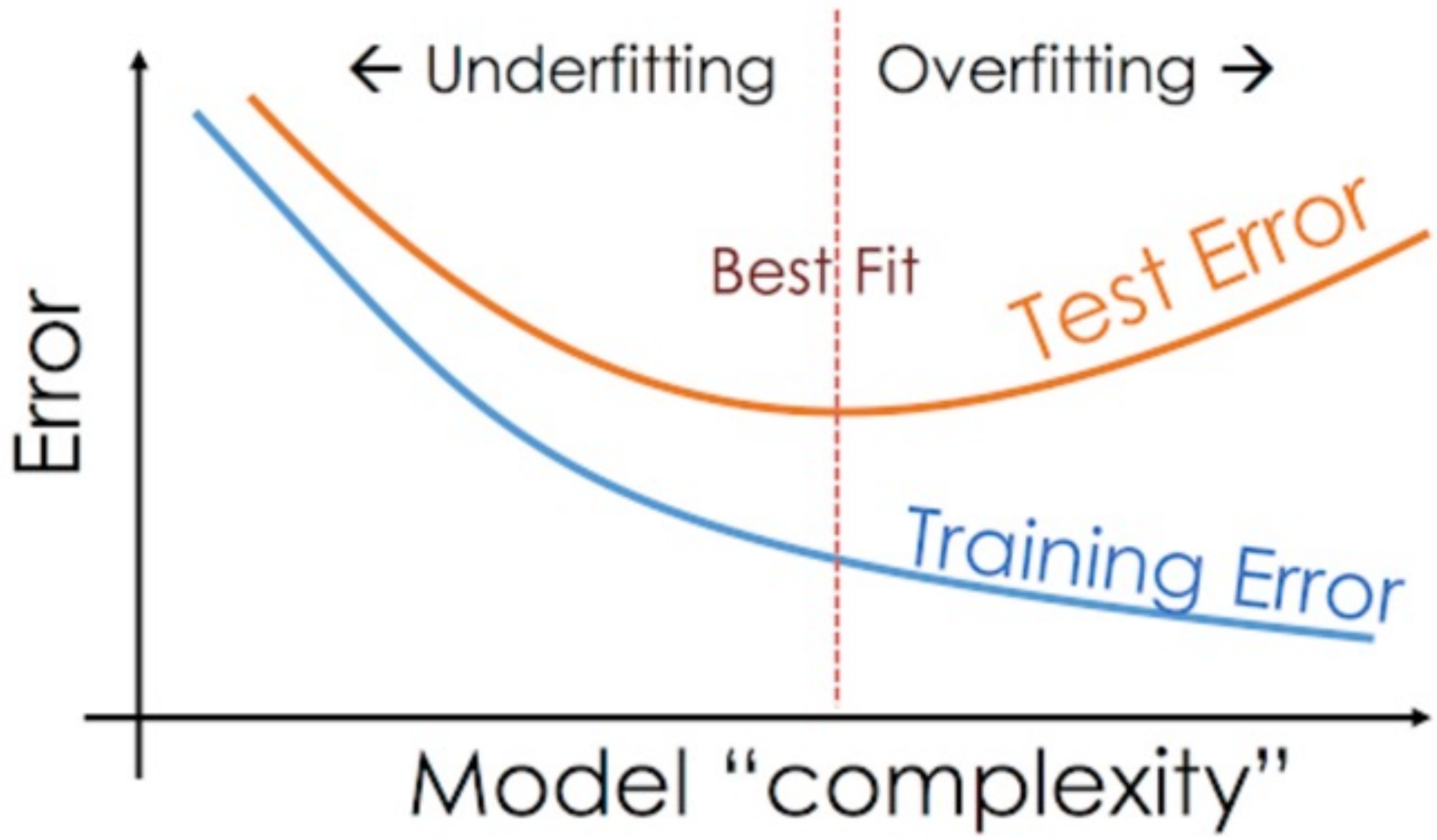
Goodfit, Underfit and Overfit



High bias

High variance

Underfit and Overfit



Q & A