Mar 6, 2023

Machine Learning Methods and Applications

Week 2. Supervised Learning: Linear Regression Models

Updates

About **grading**

- ~~Biweekly 5 homework (50%)~~ Biweekly 4 homework (4 x 10%)
- A midterm exam (20%) - Mar 27
- Final exam ~~(30%)~~ (40%) - Jun 5/12

Remember

- A learning process consists **input** and **output**.
- The major differences between Stat and ML is their purpose.
- Try to adapt the terminological differences between Stat and ML.
- A ML model predicts the target (response variable) using the features (explanatory variables).

ML models

- Supervised learning
  - Regression task
  - Classification task

- Unsupervised learning
  - Clustering task

		Task
Type of target feature	**numeric**	Regression
	categorical	Classification
	null	Clustering

Example

<u>Prediction of house sales prices</u>

Y : sales price (numeric f. / continuous v.)

X_1: surface area
X_2: number of rooms
X_3: location
…

Linear Regression Models

Simple LR: only one feature

$$Y = \beta_0 + \beta_1 X$$

Multiple LR: multiple features

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

Xs: features, βs: parameters, Y: target

Linear Regression Models

Simple LR: only one feature

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \epsilon$$

Multiple LR: multiple features

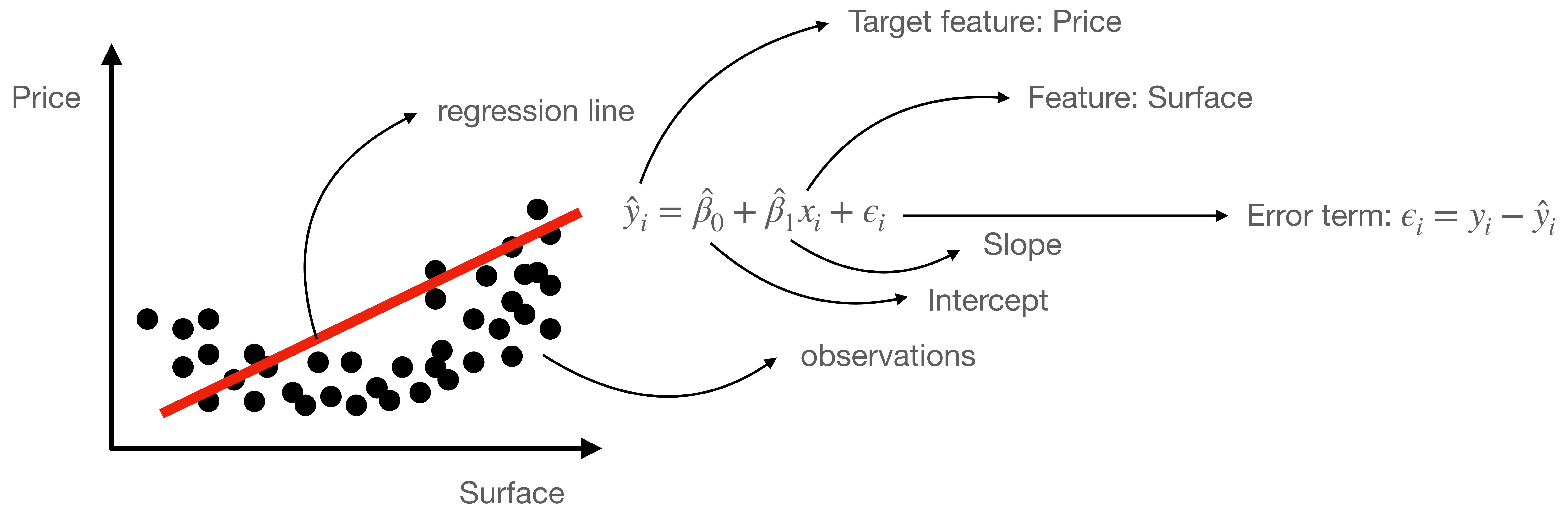
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p + \epsilon$$

xs: observed features, $\hat{\beta}$s: parameter estimations, $\hat{y}$: predicted target, ϵ: residuals

For mathematical background, you can check the section of Linear Regression in the suggested books.

Linear Regression Models

Example of simple linear regression model on the prediction of house sales



Linear Regression Models

<u>Model assumptions</u>

1. Linear relationship
2. Constant variance among residuals
3. No autocorrelation
4. More observation than predictors
5. No multicollinearity

⚠️ LRM guarantees the correct model that show the real relationship between the features and target in case of the assumptions are satisfied. However, the model <u>**may not be**</u> reliable if any of the assumptions is violated.

Steps of training regression models

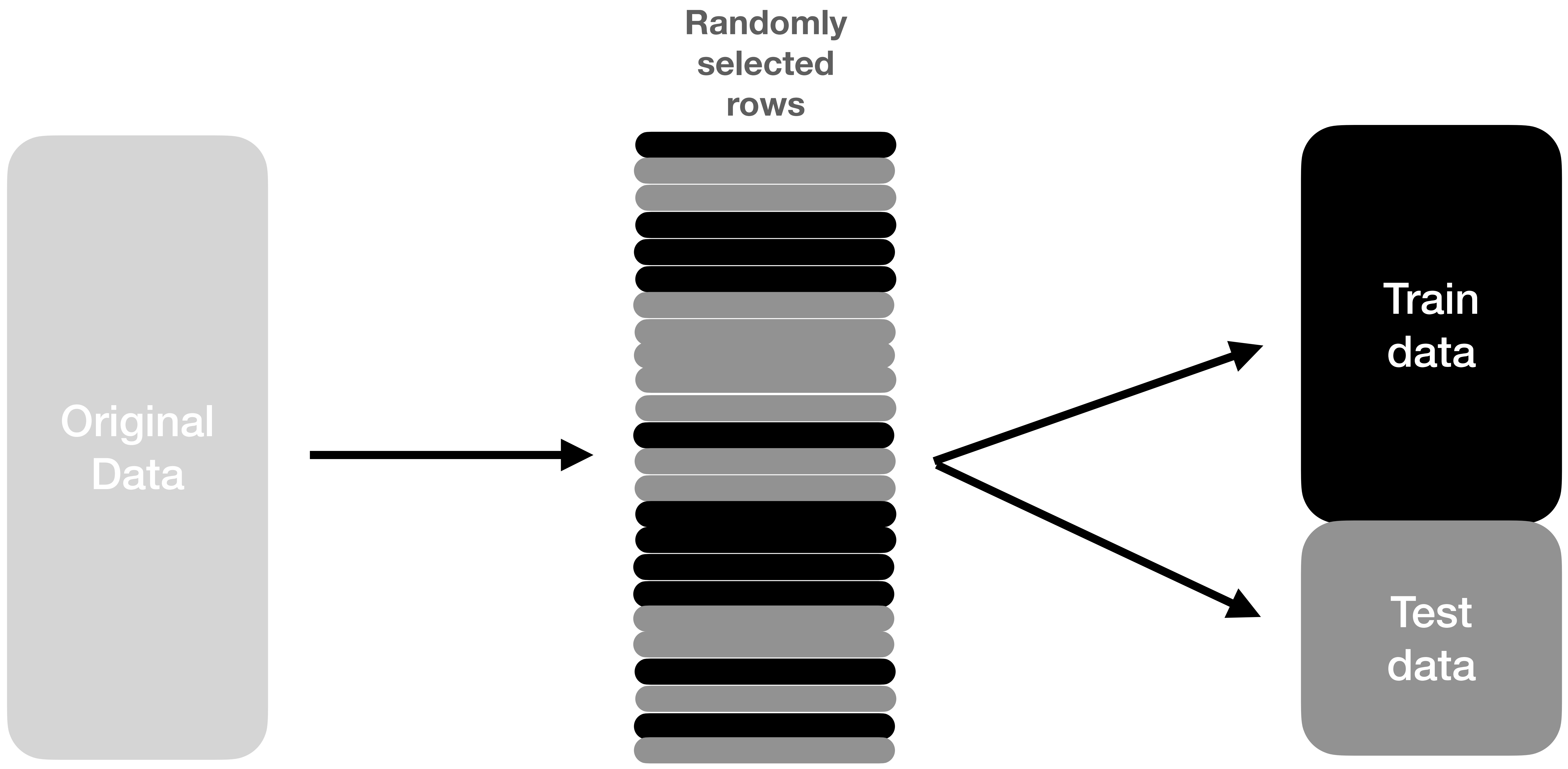
Steps

Main steps
1. Data splitting
2. Model training
3. Measuring model performance

Additional steps
X1: Checking over and underfitting
X2: Checking bias and variance of errors
X3: Checking model assumptions

Step 1. Data Splitting

Train / Test split is used rather than just validating the model on train set, it gives also an estimate of how well the model performs on new data (test set)



Step 1. Data Splitting

A major goal of the machine learning process is to find an algorithm $f(X)$ that most accurately predicts future values $(\hat{Y})$ based on a set of features (X). In other words, we want an algorithm that not only fits well to our past data, but more importantly, one that predicts a future outcome accurately. This is called the **generalizability** of our algorithm. How we “spend” our data will help us understand how well our algorithm generalizes to unseen data.

Step 2. Model training

In machine learning, the process to estimate the model coefficients is called model training. We estimate the coefficient and get the model formula in this step.

Step 3. Measuring model performance

After training model, we must check the model performance on unseen (test) data. We can use the following metrics (MSE: Mean squared error, RMSE: Root mean squared error, MAE: Mean absolute error) to measure the performance of a regression model.

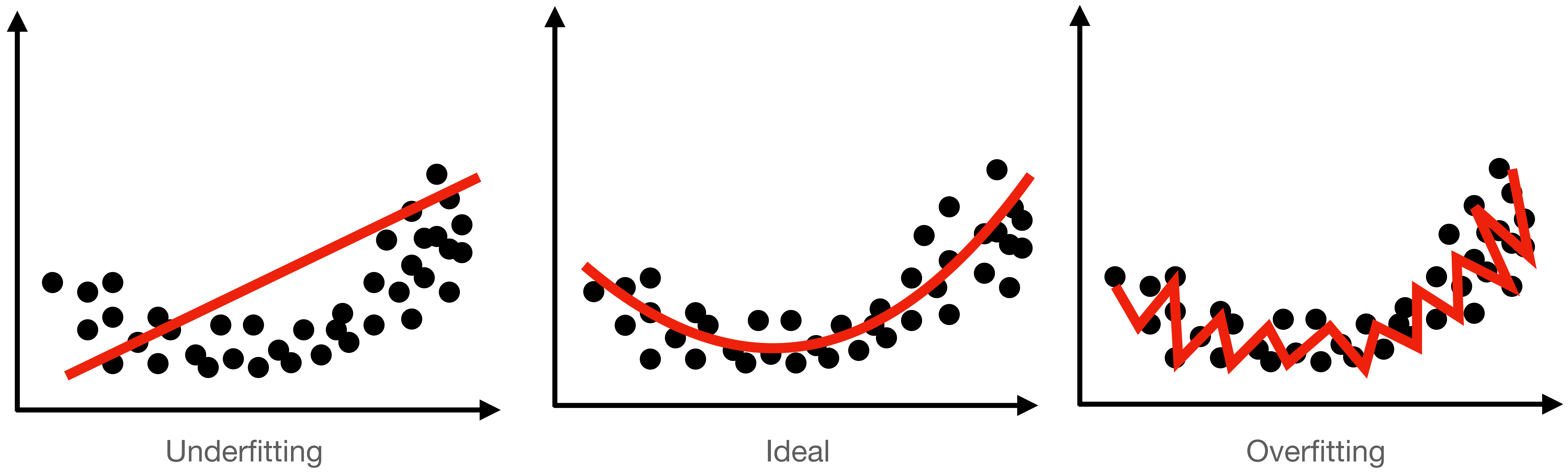
$$MSE(f) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$RMSE(f) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$MAE(f) = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

These metrics based on the difference between the observed and predicted values of target feature (aka error $\epsilon_i = y_i - \hat{y}_i$)

Step X1. Overfitting and Underfitting

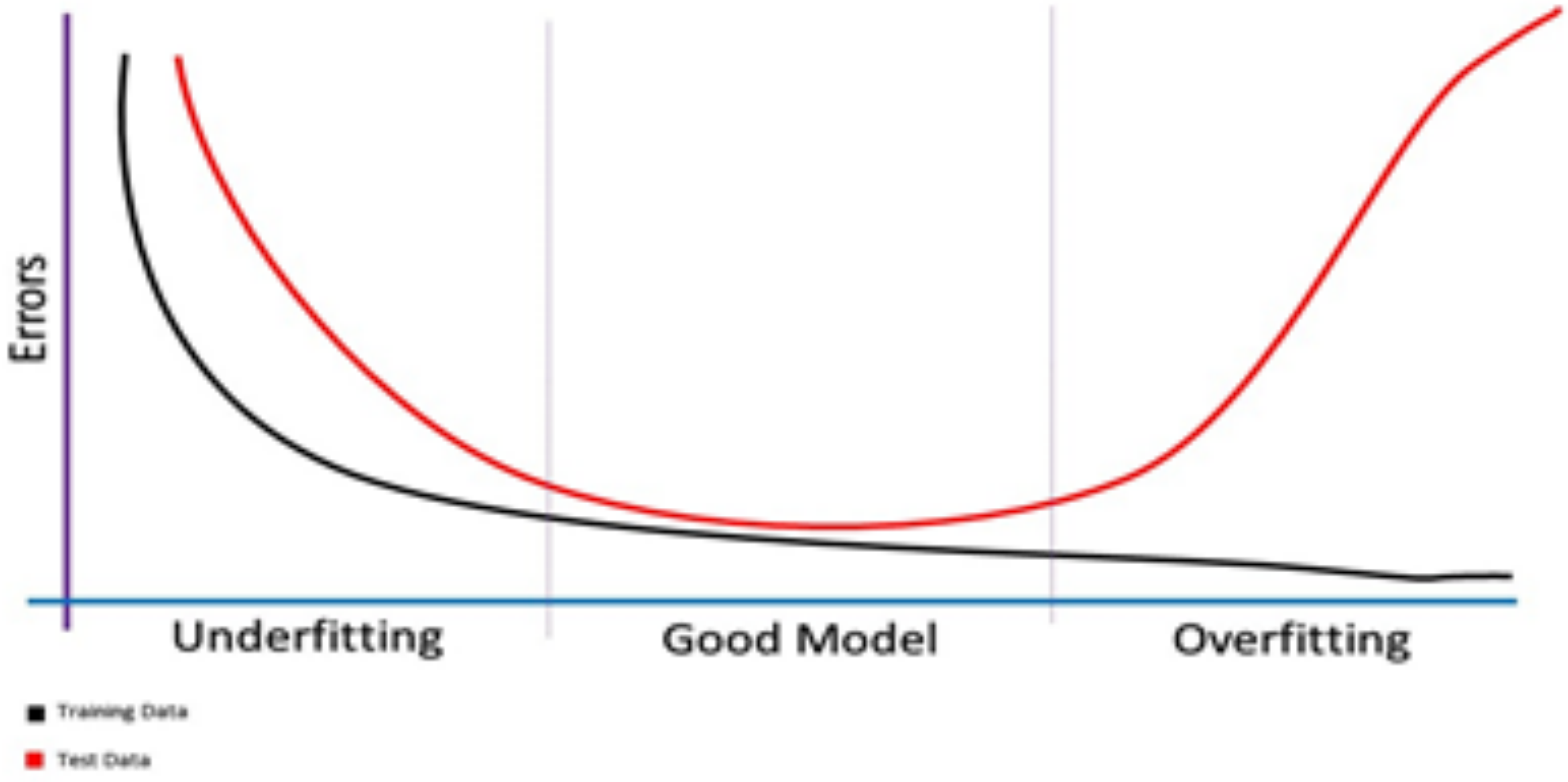


Overfitting is that a model learns from train set too well. This negatively impacts the performance of the model on test set.

Underfitting is the insufficient learning of a model from the train set. Thus it generalizes to test set.

Q. Overfitting or underfitting? Why?

Q. Overfitting or underfitting? Why?



Step X1.1. Overfitting

How to solve overfitting problem

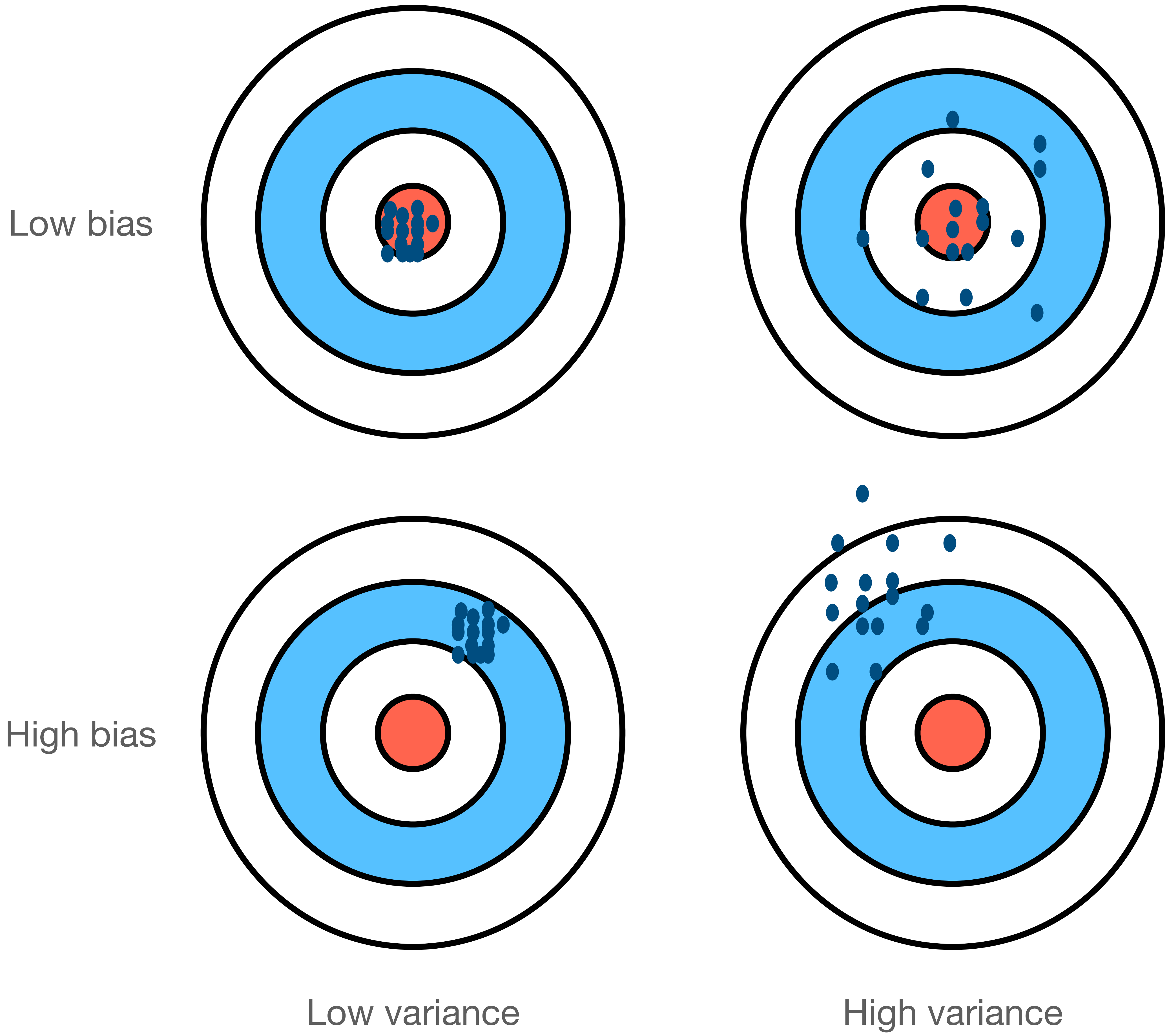
- using cross-validation
- training with more data
- removing some features
- regularization
- training an ensemble model

Step X1.2. Underfitting

<u>How to solve underfitting problem</u>

- training model with more features
- training with more data
- use more complex model
- dealing with noise problem in data

Step X2. Bias-Variance Trade-Off



Darts is a competitive sport in which two or more players bare-handedly throw small sharp-pointed missiles known as darts at a round target known as a dartboard.

In regression models, you can assume any $y - \hat{y}$ as dart throw, and the ideal case is low variance-bias, which is close to the dart-target.

Step X3. Model assumptions

1. Linear relationship
2. Constant variance among residuals
3. No autocorrelation
4. More observation than predictors
5. No multicollinearity

Application

See the R codes on the course GitHub repository!

The video recording of today's lecture will be available on YouTube, and slides on **GitHub**.
Feel free to contact me via e-mail: mustafacavus@eskisehir.edu.tr