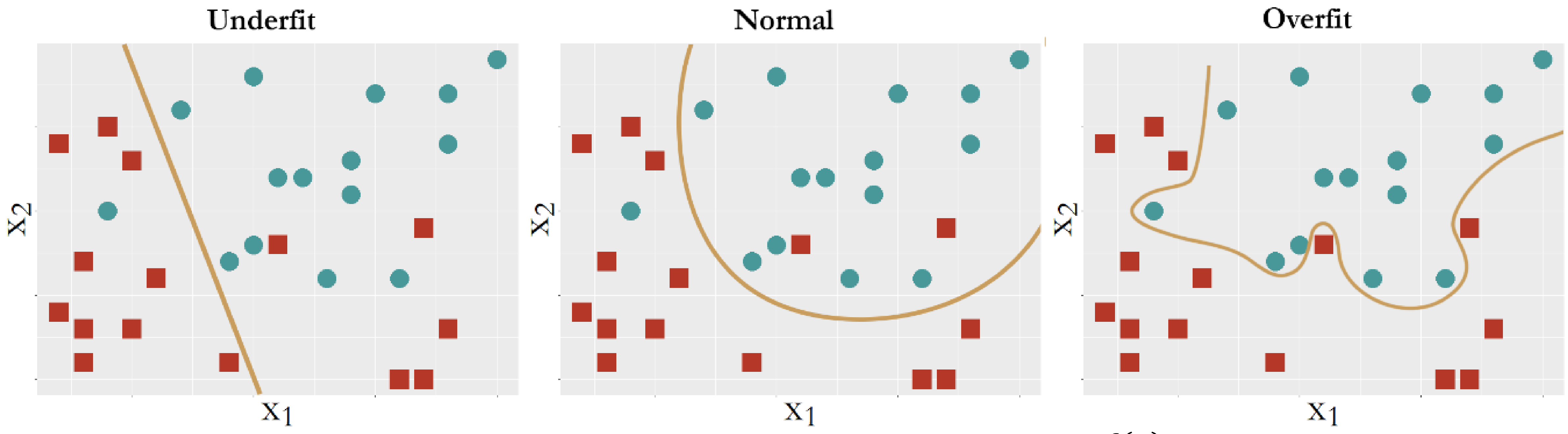
Lecture 5: Cross-Validation and Out-of-Bag (OOB) Error

Instructor: Prof. Shuai Huang

Industrial and Systems Engineering

University of Washington

Underfit, Good fit, and Overfit



$f(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

$f(\boldsymbol{x})$
$= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2$
$+ \beta_{22} x_2^2 + \beta_{12} x_1 x_2$

$f(\boldsymbol{x})$
$= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2$
$+ \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \beta_{112} x_1^2 x_2$
$+ \beta_{122} x_1 x_2^2 + \cdots$

Danger of R-squared

- When number of variables increases, in theory, the R-squared won’t decrease; in practice, it always increases. Thus, it is not a good metric to take into consideration of model complexity

$$R^2 = 1 - \frac{SSE}{SST}$$

- This is because that: ST is always fixed, while SSE could only decrease if more variables are put into the model even if these new added variables have no relationship with the outcome variable

Danger of R-squared (cont'd)

- Further, the R-squared is compounded by the variance of predictors as well. As the underlying regression model is

$$Y = \beta X + \epsilon,$$

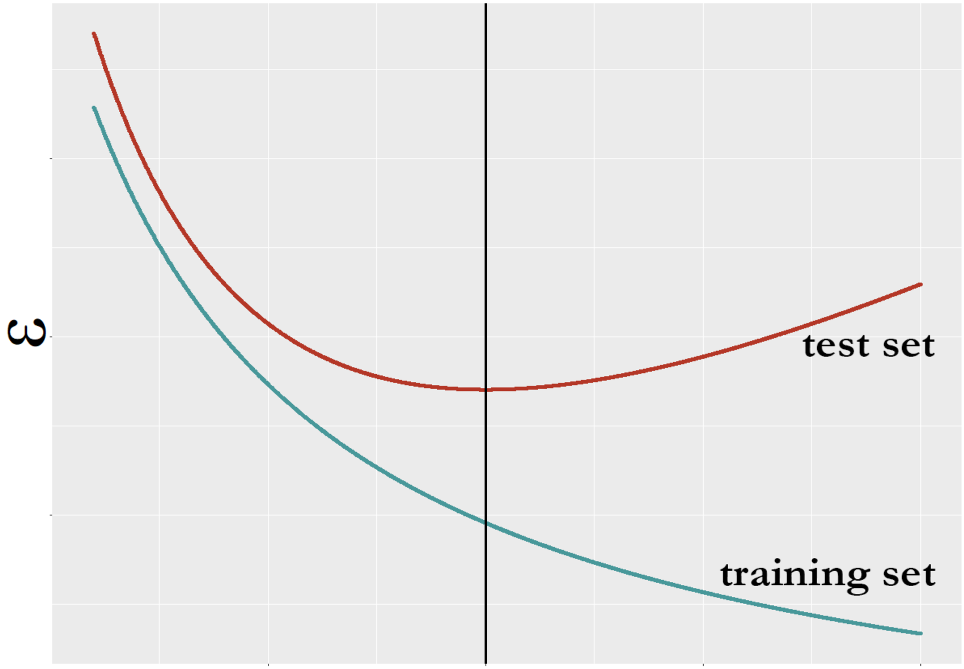
- The variance of Y, $var(Y) = \beta^2 var(X) + var(\epsilon)$. The R-squared takes the form as

$$\text{R-squared}=\frac{\beta^2 var(X)}{\beta^2 var(X)+var(\epsilon)}.$$

- Thus, it seems that R-squared is not only impacted by how well X can predict Y, but also by the variance of X as well.

The truth about training error

- Just as the R-squared, it will continue to decrease if the model is mathematically more complex (therefore, more able to shape itself to make its prediction correct on data points that are due to noise)



Fix R-squared: AIC/BIC/?IC…

- The definition of AIC (Akaike Information Criterion)

$$AIC = 2k - 2\ln(\hat{L})$$

- The definition of BIC (Bayesian Information Criterion)

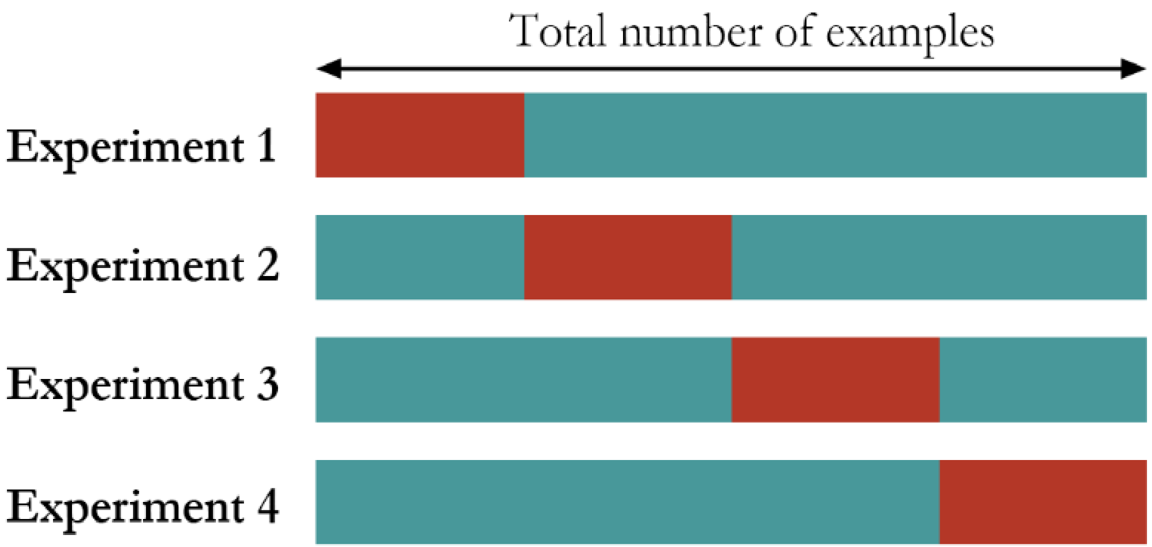
$$BIC = \ln(N)\,k - 2\ln(\hat{L})$$

Training and testing data

- A simple strategy: if a model is good, then it should perform well on an unseen testing data (that represents the future data – which is of course unseen in the model training stage)

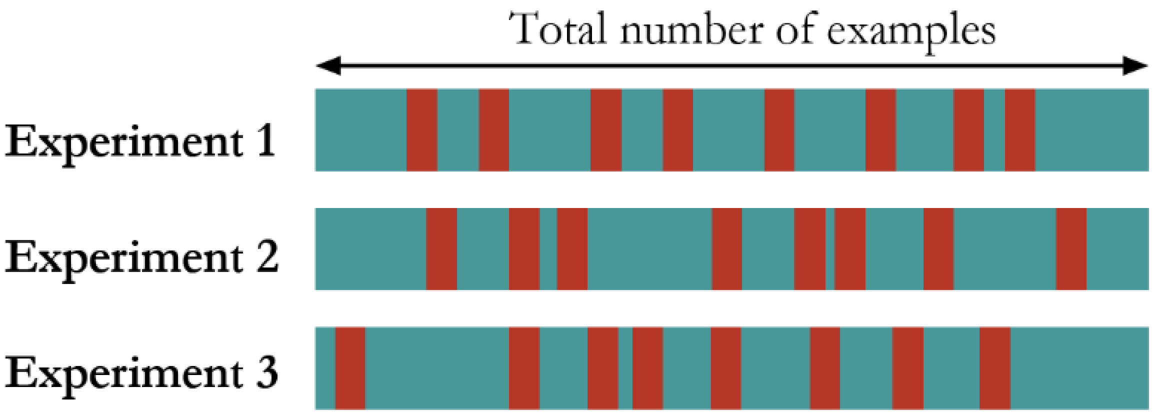
K-Fold cross-validation

- For example, K=4



Random sampling method

- How to conduct the training/testing data scheme, when we only have access to a dataset (usually we take this dataset as “training data” – a concept taken for granted)?



Other dimensions of "error"

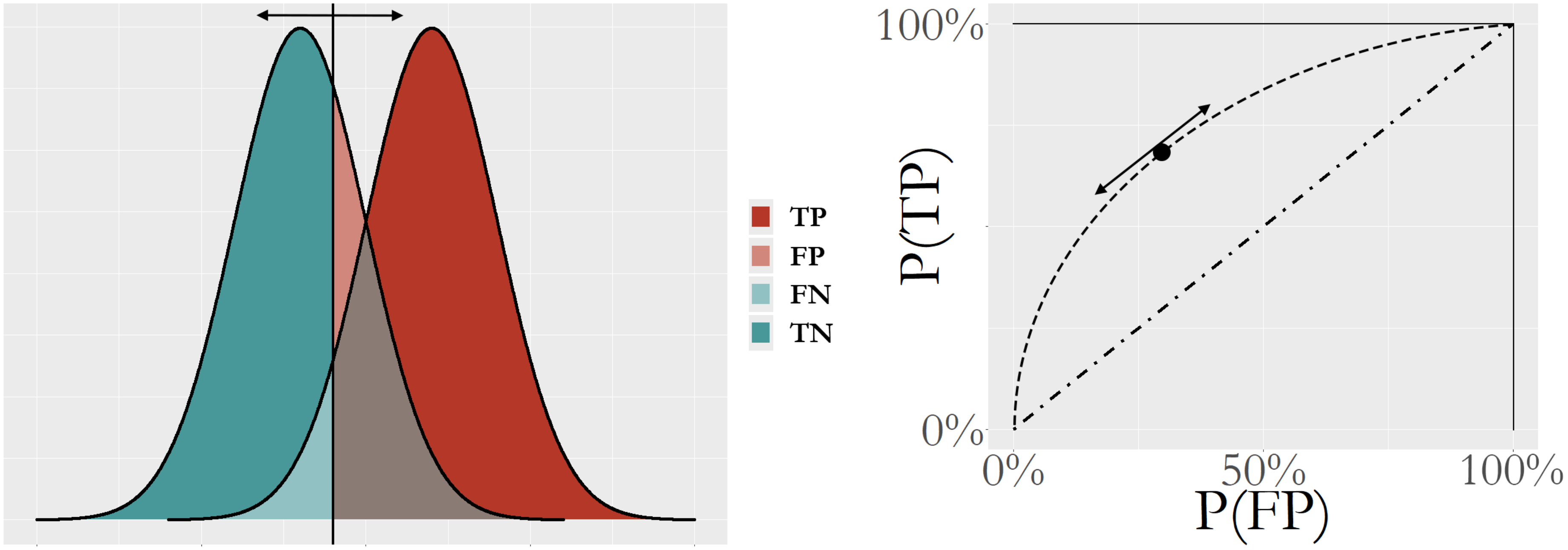
- The TP, FP, FN, TN

Table 5.1: The confusion matrix

The confusion matrix		**Reality**	
		Positive	Negative
Model Prediction	**Positive**	**True positive (TP)**	**False positive (FP)**
	Negative	**False negative (FN)**	**True negative (TN)**

The ROC curve (Receiver Operating Characteristics)

- Consider a logistic regression model

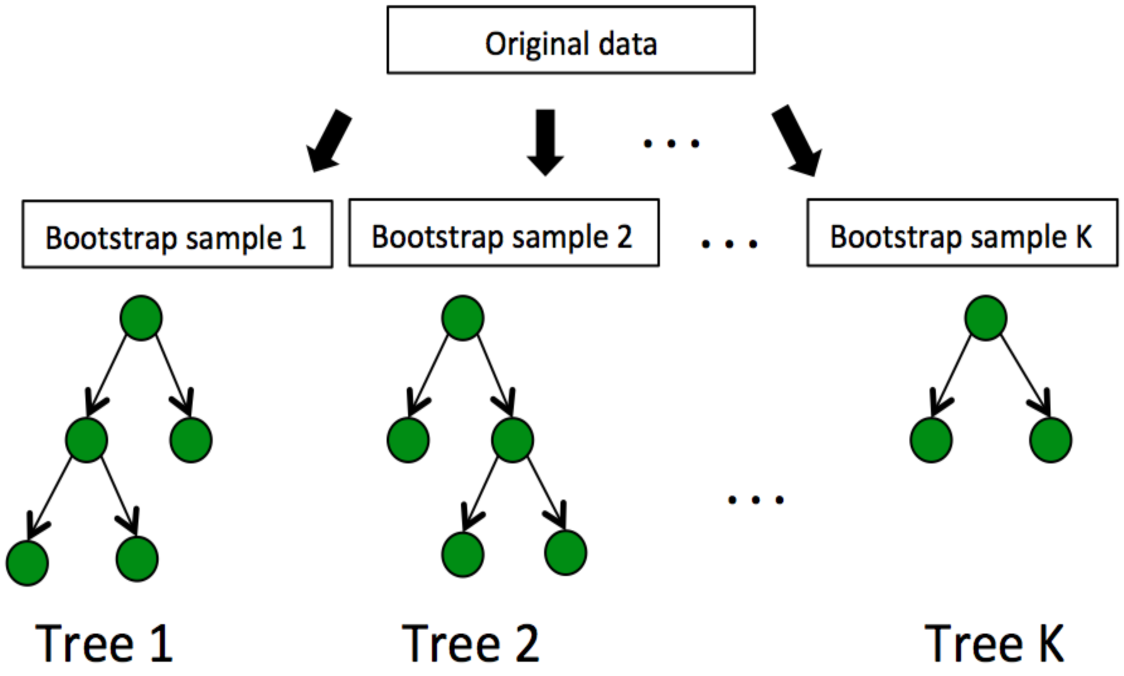


R lab

- Download the markdown code from course website
- Conduct the experiments
- Interpret the results
- Repeat the analysis on other datasets

The Out-of-Bag (OOB) error

- The out-of-bag (OOB) error in a random forest model provides a computationally convenient approach to evaluate the model without using a testing dataset, neither a cross-validation procedure



The idea behind the OOB error

- The probability of a data point from the training data is missing from a bootstrapped dataset is
$$\left(1-\frac{1}{N}\right)^N.$$
- When N is sufficiently large, we can have
$$lim_{N\to\infty}\left(1-\frac{1}{N}\right)^N = e^{-1} \approx 0.37.$$
- Therefore, roughly 37% of the data points from S are not contained in any bootstrapped dataset B_i.
- And thus, not used for training tree i. These excluded data points are referred as the out-of-bag samples for the bootstrapped dataset B_i and tree i.

Further develop the line of argument

- As there are 37% of probability that a data point is not used for training a tree, we can infer that, a data point is not used for training about 37% of the trees.
- Therefore, for each data point, in theory, there are 37% of trees trained without this data point. These trees can be used to predict on this data point, which can be considered as testing an unseen data.
- The out-of-bag error estimation can then be calculated by aggregating the out-of-bag testing error of all the data points.
- The out-of-bag error can be calculated after random forests are built, and are significantly less computationally than cross-validation.

A Simple Example

- Suppose that we have a training dataset of 5 instances (IDs as 1,2,3,4,5).

Bootstrap	Tree
1,1,4,4,5	1
2,3,3,4,4	2
1,2,2,5,5	3

Table 5.3: The out-of-bag (OOB) errors

Tree	Training data	1 (C1)	2 (C2)	3 (C2)	4 (C1)	5 (C2)
1	1,1,4,4,5		C1	C2		
2	2,3,3,4,4	C1				C2
3	1,2,2,5,5			C2	C1	

- We can see that, as the data instance (ID = 1) is not used in training Tree 2, we can use Tree 2 to predict on this data instance, and we see that it correctly predicts the class as C1.
- Similarly, Tree 1 is used to predict on data instance (ID=2), and the prediction is wrong. Finally, we can see that the overall out-of-bag (OOB) error is 1/6.

R lab

- Download the markdown code from course website
- Conduct the experiments
- Interpret the results
- Repeat the analysis on other datasets