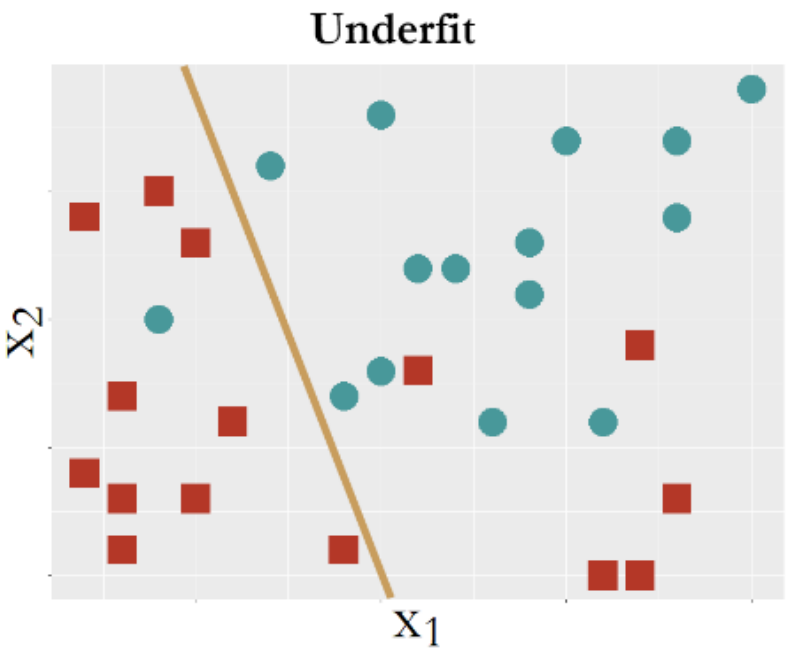
Lecture 7: Cross-Validation

Instructor: Prof. Shuai Huang

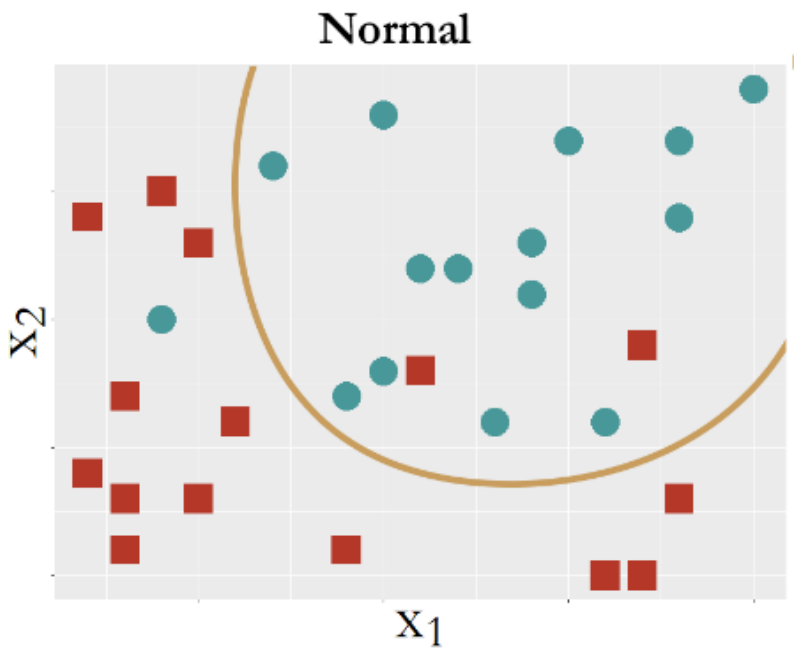
Industrial and Systems Engineering

University of Washington

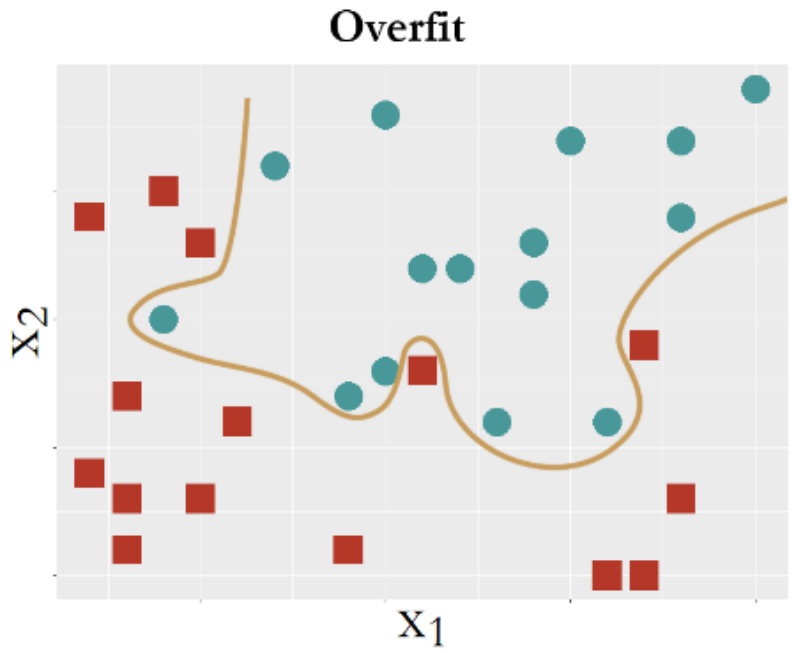
Underfit, Good fit, and Overfit



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



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



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

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

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

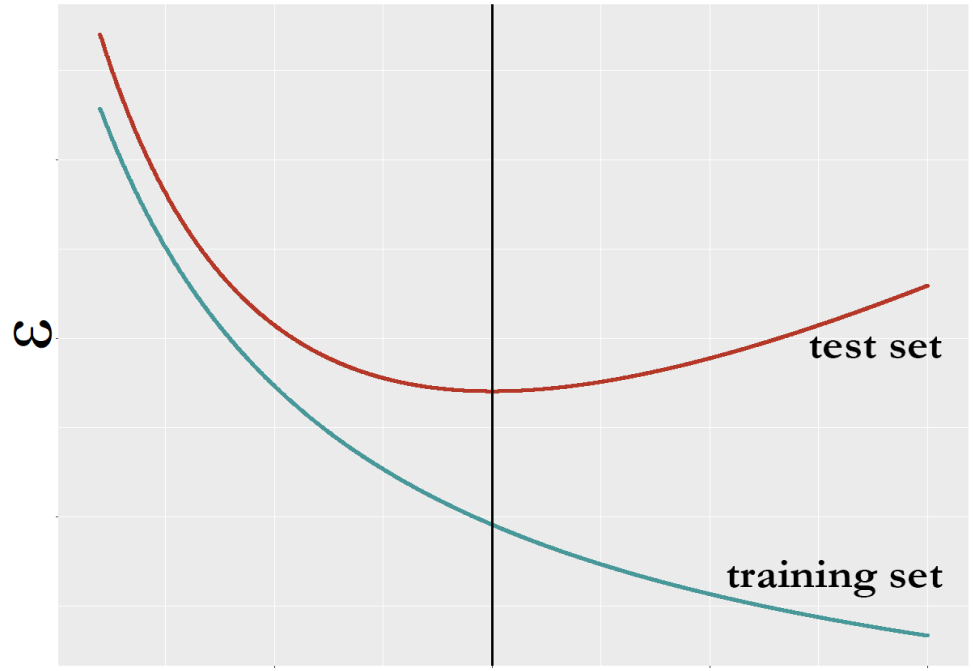
- The variance of Y, $var(Y) = \beta^2 var(X) + var(\epsilon)$. The R-squared takes the form as

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

- 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)

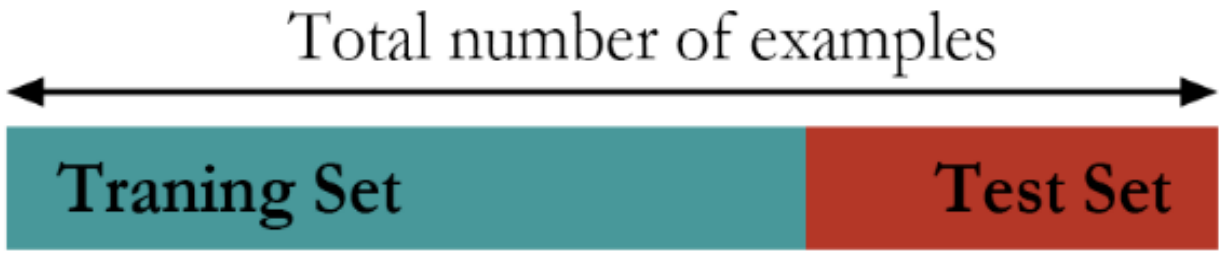
$$AIC = 2k - 2\ln(\hat{L})$$

- The definition of BIC (Bayesian Information Criterion)

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

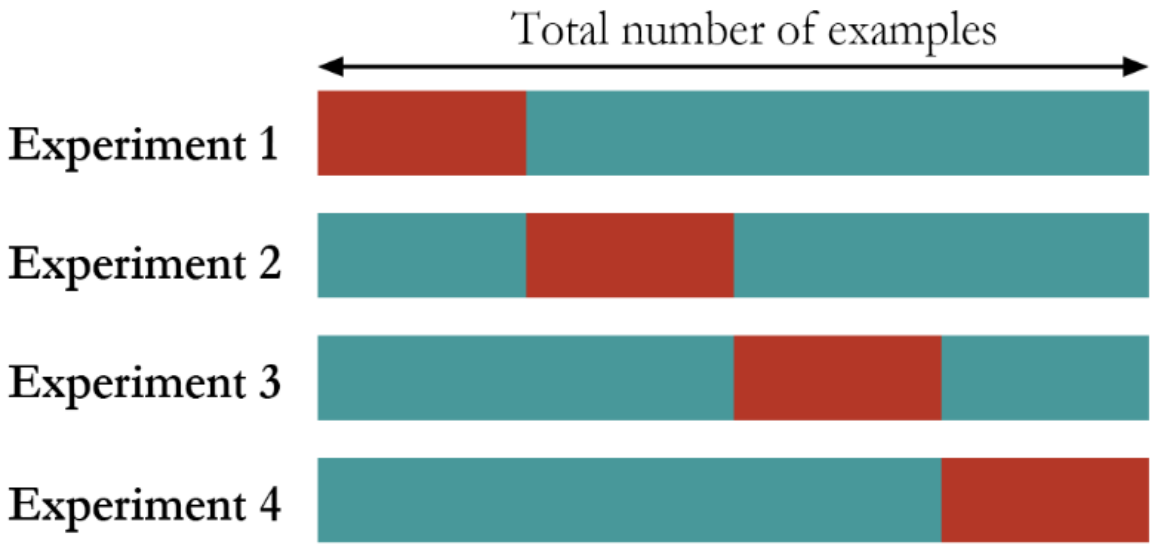
Training and testing data

- 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)



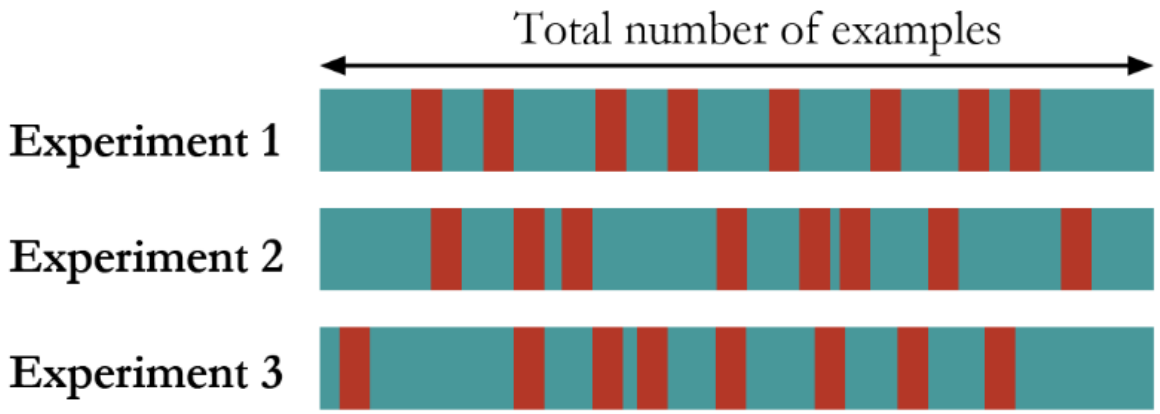
K-Fold cross-validation

- 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”

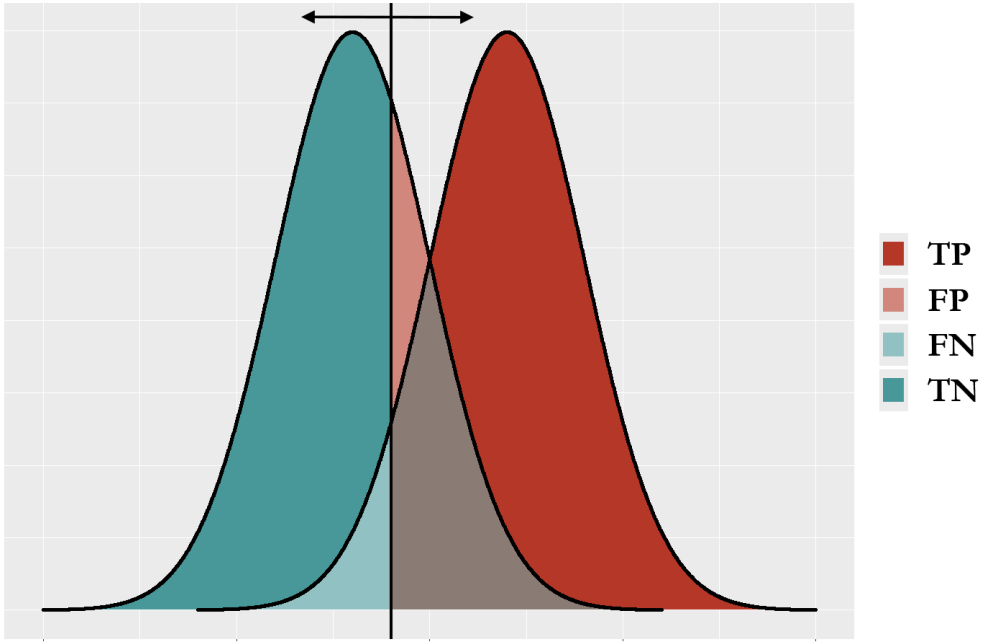
- The TP, FP, FN, TN

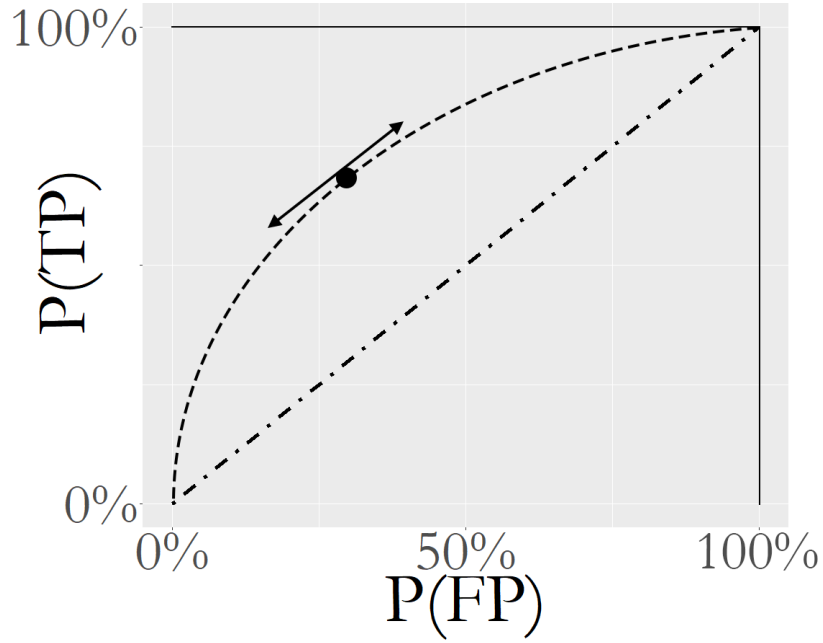
Table 5.1: The confusion matrix

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