EE2211 Introduction to Machine Learning

Lecture 10

Wang Xinchao
xinchao@nus.edu.sg

Course Contents

- Introduction and Preliminaries (Xinchao)
  - Introduction
  - Data Engineering
  - Introduction to Linear Algebra, Probability and Statistics
- Fundamental Machine Learning Algorithms I (Vincent)
  - Systems of linear equations
  - Least squares, Linear regression
  - Ridge regression, Polynomial regression
- Fundamental Machine Learning Algorithms II (Vincent)
  - Over-fitting, bias/variance trade-off
  - Optimization, Gradient descent
  - Decision Trees, Random Forest
- Performance and More Algorithms (Xinchao)
  - Performance Issues
  - K-means Clustering
  - Neural Networks

EE2211: Learning Outcome
A Summary of Module Content

- **I am able to understand the formulation of a machine learning task**
  - Lecture 1 (feature extraction + classification)
  - Lecture 4 to Lecture 9 (regression and classification)
  - Lecture 11 and Lecture 12 (clustering and neural network)
- **I am able to relate the fundamentals of linear algebra and probability to machine learning**
  - Lecture 2 (recap of probability and linear algebra)
  - Lecture 4 to Lecture 8 (regression and classification)
  - Lecture 12 (neural network)
- **I am able to prepare the data for supervised learning and unsupervised learning**
  - Lecture 1 (feature extraction), Page 26 to 31 [For supervised and unsupervised]
  - Lecture 2 (data wrangling) [For supervised and unsupervised]
  - Lecture 10 (Training/Validation/Test) [For supervised]
  - Programming Exercises in tutorials
- **I am able to evaluate the performance of a machine learning algorithm**
  - Lecture 5 to Lecture 9 (evaluate the difference between labels and predictions)
  - Lecture 10 (evaluation metrics)
- **I am able to implement regression and classification algorithms**
  - Lecture 5 to Lecture 9

Outline

- Dataset Partition:
  - Training/Validation/Testing

- Cross Validation

- Evaluation Metrics
  - Evaluating the quality of a trained classifier

We will talk about many metrics:
It is OK you can't memorize them all
But intuition is important!

A Real-world Scenario

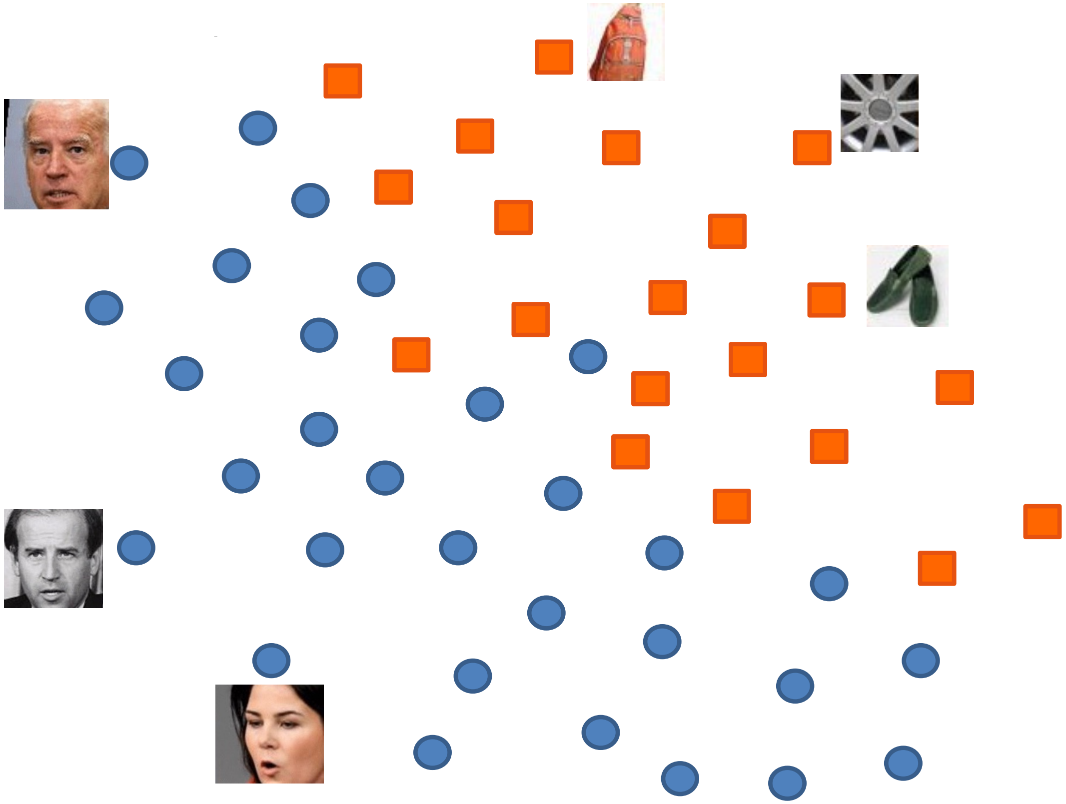
- We would like to train a Random Forest for face classification (i.e., to tell an image is a human face or not)

Faces

Non-faces

A Real-world Scenario

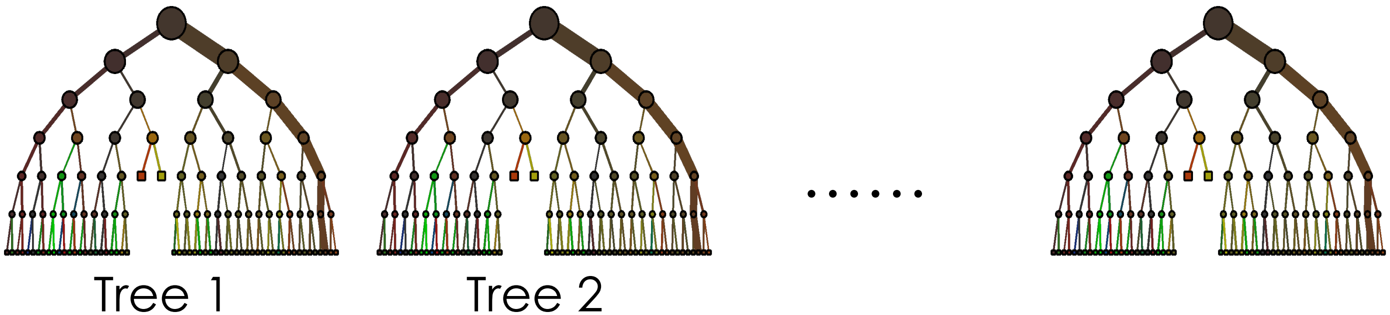
- We would like to train a Random Forest for face classification (i.e., to tell an image is a human face or not)
  - We will have **<u>one dataset</u>** to train the Random Forest



A Real-world Scenario

- We would like to train a Random Forest for face classification (i.e., to tell an image is a human face or not)
  - We will have **one dataset** to train the Random Forest
  - We will have tunable **(hyper)parameters** for the Random Forest. For example, ***the number of trees*** in the Random Forest
    - Shall we use 100 trees?
    - Shall we use 200 trees?
    - …

    We need to decide on the parameter



A Real-world Scenario

- We would like to train a Random Forest for face classification (i.e., to tell an image is a human face or not)
  - We will have **one dataset** to train the Random Forest
  - We will have tunable **(hyper)parameters** for the Random Forest. For example, ***the number of trees*** in the Random Forest
    - Shall we use 100 trees?
    - Shall we use 200 trees?
    - …

    We need to decide on the parameter
  - Once we decide the number of trees, we will the Random Forest with the **selected parameter** on **unseen** test data.

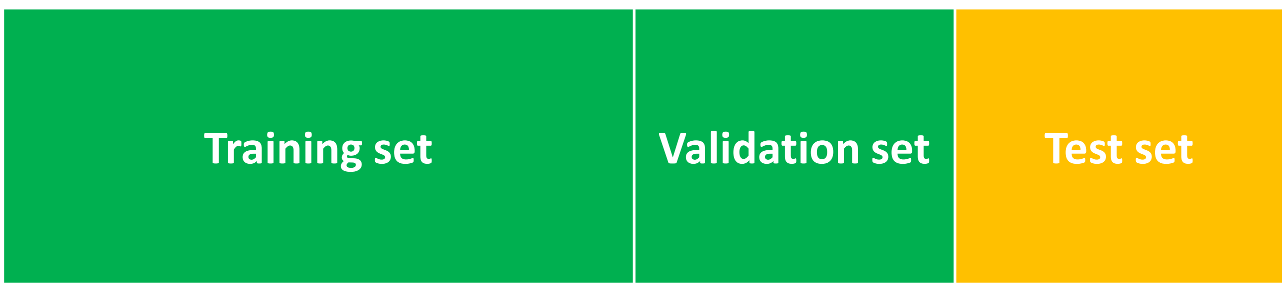
Test Data

Yes!

No!

Training, Validation, and Test

- In real-world application,
  - We don't have test data, since they are unseen
  - Imagine you develop a face detector app, you don't know whom you will test on
- In lab practice,
  - We divide the dataset into three parts

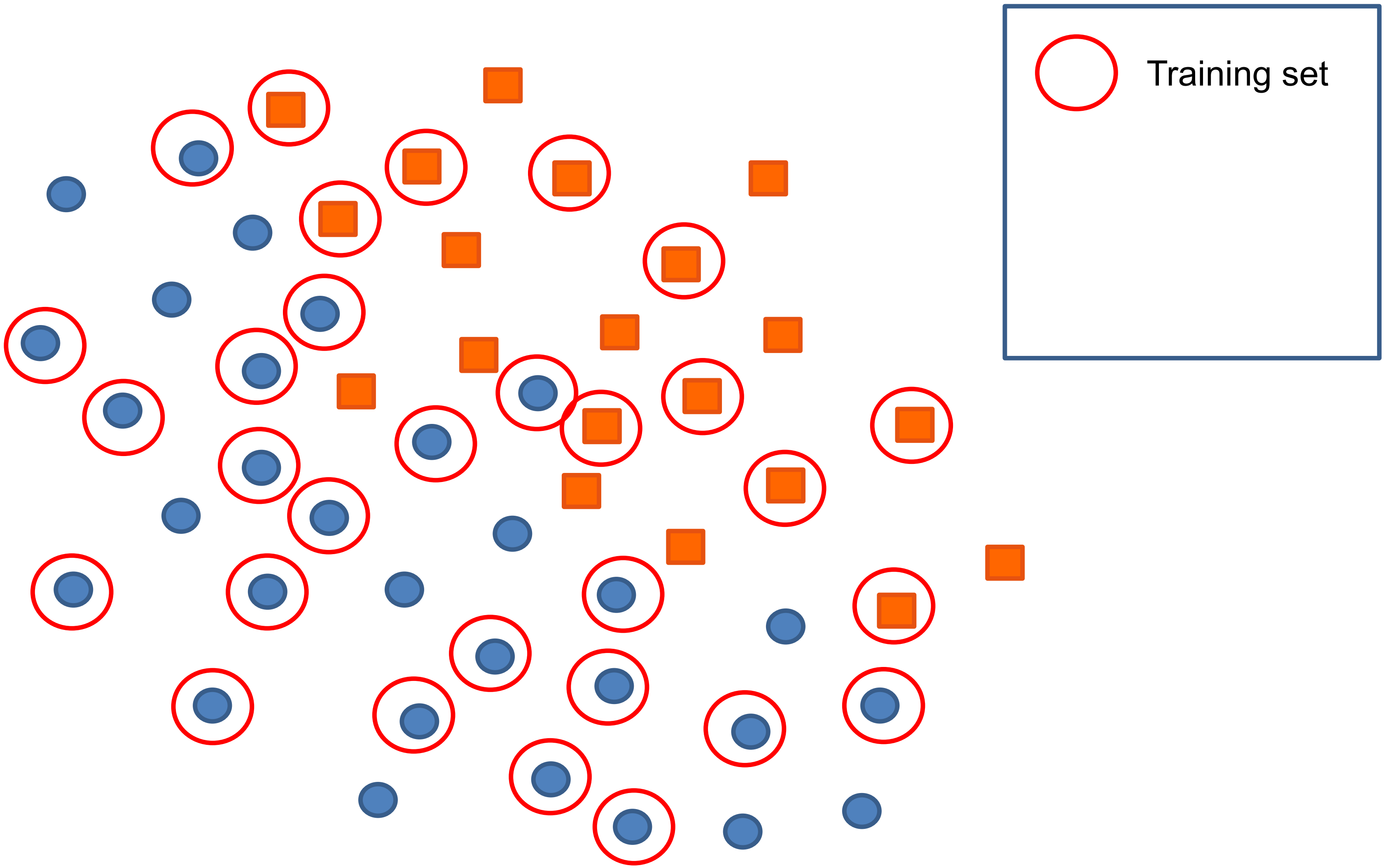


Hidden from Training !

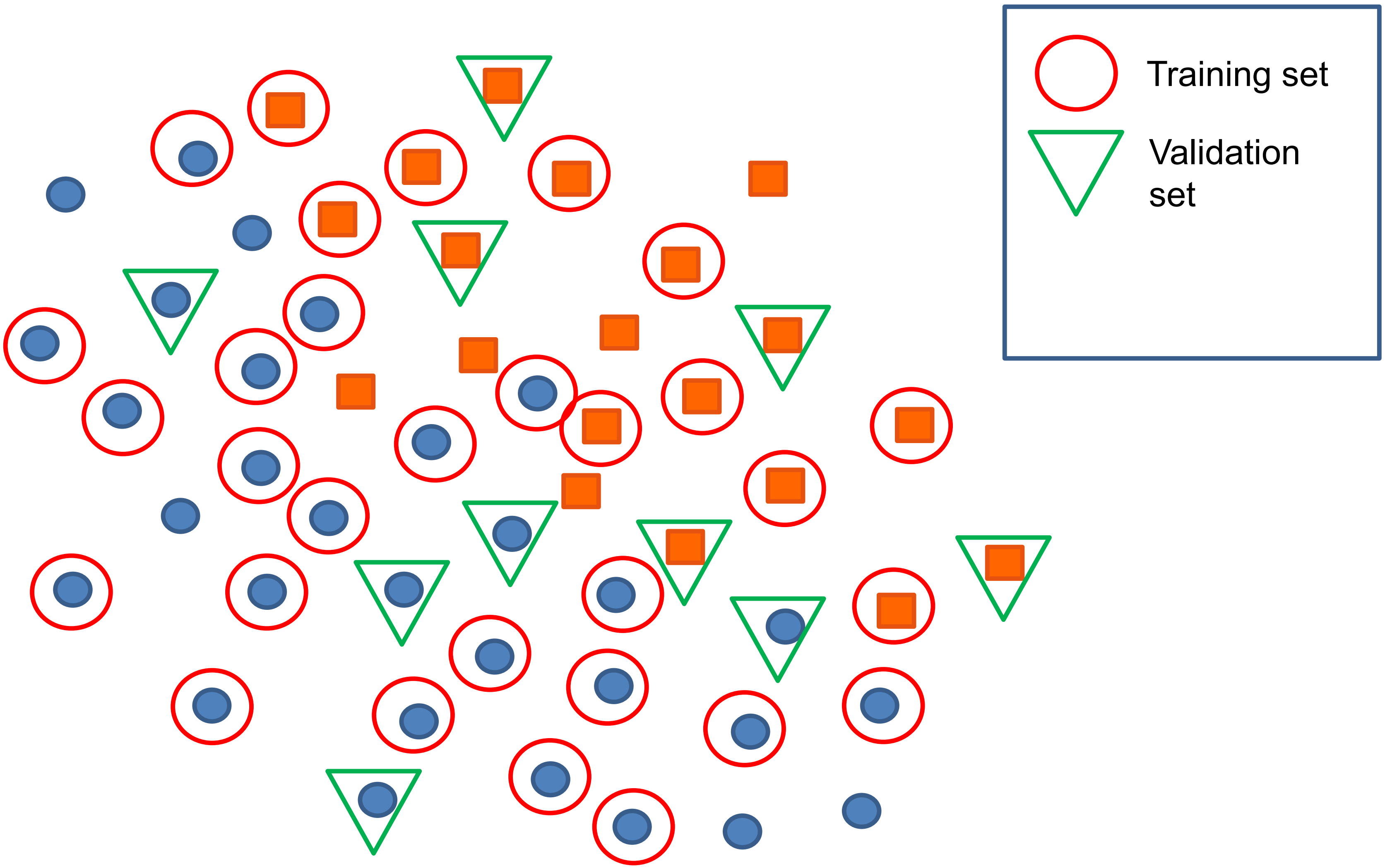
For training the ML models	For validation: choosing the parameter or model	For testing the "real" performance and generalization

  - **NEVER touch test data during training!!!**

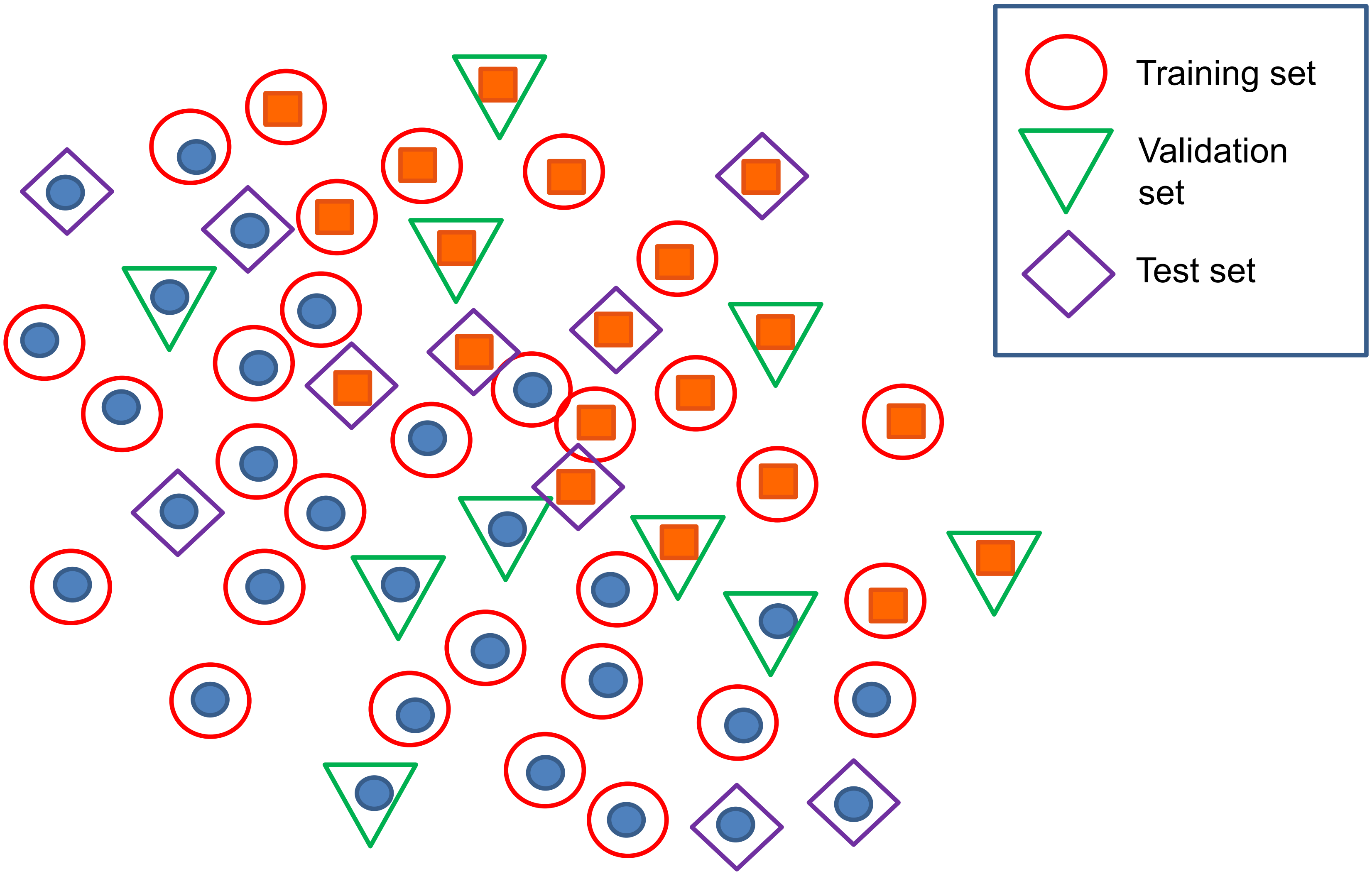
Training, Validation, and Test



Training, Validation, and Test



Training, Validation, and Test



Training, Validation, and Test

Training set	Validation set	Test set
For training the ML models	For validation: choosing the parameter or model	For testing the "real" performance and generalization

Example: Assume I want to build a Random Forest. I have a parameter to decide: shall I have

- 100 Trees?
- 200 Trees?

What we do next is to use the **training set** to train two classifiers,
1) C_1: Random Forest with 100 trees, and 2) C_2: Random Forest with 200 trees

They have the following accuracy:

1. **C_1: Random Forest with 100 trees**: training accuracy 92%, validation accuracy 90%
2. **C_2: Random Forest with 200 trees**: training accuracy 95%, validation accuracy 88%

Which one to choose for real application, i.e., testing?
The one with higher validation accuracy, i.e., Random Forest with 100 trees!

Python Demo: lec10.ipynb

For training the ML models

For validation: choosing the parameter or model

For testing the "real" performance and generalization

- Problem Setup
  - Dataset used: IRISdataset
    - Link: https://scikit-learn.org/stable/datasets/toy_dataset.html#iris-dataset
  - Training/Validation/Test: 100/25/25
  - Machine Learning Task and Model: Polynomial regression
  - Parameters to select: Order 1 to 10

Cross Validation

- In practice, we do the **k-fold cross validation**

Test

4-fold cross validation

					Classifiers Trained
Fold 1	Train	Train	Train	Validation	C_1^1 C_2^1
Fold 2	Train	Train	Validation	Train	C_1^2 C_2^2
Fold 3	Train	Validation	Train	Train	C_1^3 C_2^3
Fold 4	Validation	Train	Train	Train	C_1^4 C_2^4

Example: which one to select for test?

	Fold 1 Accuracy on Validation Set 1	Fold 2 Accuracy on Validation Set 2	Fold 3 Accuracy on Validation Set 3	Fold 4 Accuracy on Validation Set 4	Average Accuracy on All Validation Sets
Classifier with Param1 (e.g. 100 trees)	88% C_1^1	89% C_1^2	93% C_1^3	92% C_1^4	90.5% ✓
Classifier with Param2 (e.g. 200 trees)	90% C_2^1	88% C_2^2	91% C_2^3	91% C_2^4	90%

Cross Validation

Other common partitioning:

- 10-Fold CV
- 5-Fold CV
- 3-Fold CV

We may decide on the size of the test set, for example, 15%, 20%, 30% of the whole dataset, the rest for training/validation.

The test set contains the examples that the learning algorithm has never seen before, so if our model performs well on predicting the labels of the examples from the test set, we say that our model **generalizes well**.

Training, Validation, and Test

- Validation is however not always used:
  - Validation is used when you need to **pick parameters or models**
  - If you have no models or parameters to compare, you may consider partition the data into only training and test

Outline

- Dataset Partition:
  – Training/Validation/Testing
- Cross Validation
- Evaluation Metrics
  – Evaluating the quality of a trained classifier

We will talk about many metrics:
It is OK you can’t memorize them all
But intuition is important!

Evaluation Metrics

Regression

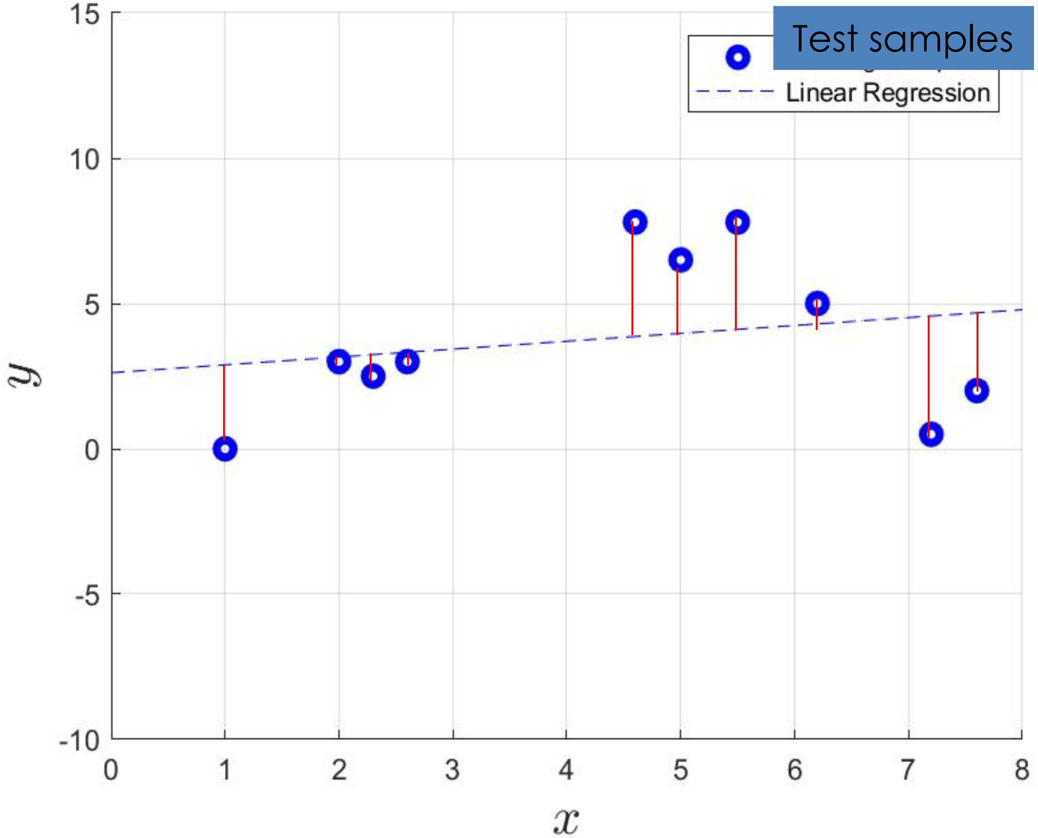
Mean Square Error

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

Mean Absolute Error

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

where y_i denotes the target output and $\hat{y}_i$ denotes the predicted output for sample i.



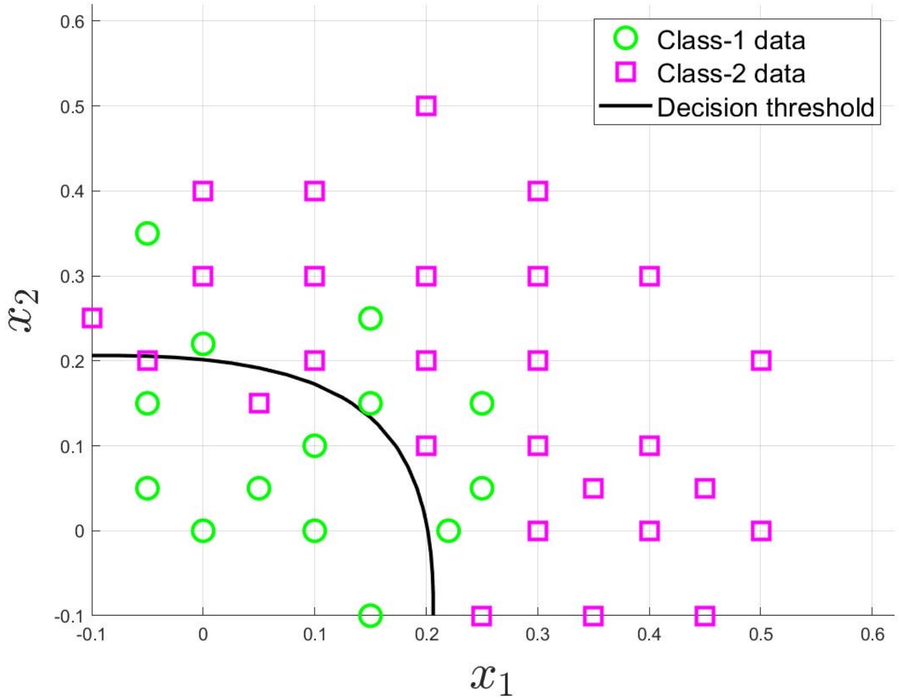
Evaluation Metrics

Classification

Class-1: Positive Class
Class-2: Negative Class

Confusion Matrix

	Class-1 (predicted)	Class-2 (predicted)
Class-1 (actual)	7 (TP)	7 (FN)
Class-2 (actual)	2 (FP)	25 (TN)



TP: True Positive
FN: False Negative (i.e., **Type II Error**)
FP: False Positive (i.e., **Type I Error**)
TN: True Negative

Evaluation Metrics

	Class-1 (predicted)	Class-2 (predicted)
Class-1 (actual)	7 (TP)	7 (FN)
Class-2 (actual)	2 (FP)	25 (TN)

Classification

- How many samples in the dataset have the real label of Class-2?

  2+25=27
- How many samples are there in total?

  7+7+2+25 = 41
- How many sample are correctly classified? How many are incorrectly classified?

  correctly classified -> 7+25 = 32;

  incorrectly classified -> 7 + 2 = 9

Evaluation Metrics

Classification

Confusion Matrix for Binary Classification

	$\hat{P}$ (predicted)	$\hat{N}$ (predicted)	
P (actual)	TP	FN	Recall TP/(TP+FN)
N (actual)	FP	TN	
	Precision TP/(TP+FP)		Accuracy (TP+TN)/(TP+TN+FP+FN)

Evaluation Metrics

Classification

Cost Matrix for Binary Classification

	$\hat{\mathrm{P}}$ (predicted)	$\hat{\mathrm{N}}$ (predicted)
P (actual)	$C_{p,p}$ * TP	$C_{p,n}$ * FN
N (actual)	$C_{n,p}$ * FP	$C_{n,n}$ * TN

Total cost:
$C_{p,p}$ * TP +
$C_{p,n}$ * FN +
$C_{n,p}$ * FP +
$C_{n,n}$ * TN

Main Idea: To assign different ***penalties*** for different entries. Higher penalties for more severe results.

Usually, $C_{p,p}$ and $C_{n,n}$ are set to 0; $C_{n,p}$ and $C_{p,n}$ may and may not equal

Evaluation Metrics

- Example of cost matrix
  - Assume we would like to develop a self-driving car system
  - We have an ML system that detects the pedestrians using camera, by conducing a binary classification
    - When it detects a person (positive class), the car should stop
    - When no person is detected (negative class), the car keeps going

Credit: automotiveworld.com

True Positive (cost $C_{p,p}$)
There is person, ML detects person and car stops

True Negative (cost $C_{n,n}$)
There is no person, car keeps going

False Positive (cost $C_{n,p}$)
There is no person, ML detects person and car stops

False Negative (cost $C_{p,n}$)
There is person, ML fails to detect person and car keeps going

$$C_{n,p} < C_{p,n} \; (>, <, \text{or} =)$$

Evaluation Metrics

- Handling unbalanced data
  - Assume we have 1000 samples, of which 10 are *positive* and 990 are *negative*
  - Accuracy = 990/1000=0.99!
  - Yet, half of the Class-1 are Classified to Class-2!

	Class-1 (predicted)	Class-2 (predicted)
Class-1 (actual)	5 (TP)	5 (FN)
Class-2 (actual)	5 (FP)	985 (TN)

The goal is to highlight the problems of the results!

In this case, we shall

1) Use cost matrix, assign different costs for each entry
2) Use Precision and Recall! Precision = 0.5 and Recall = 0.5

Evaluation Metrics

Classification

(True Positive Rate) TPR = TP/(TP+FN) Recall
(False Negative Rate) FNR = FN/(TP+FN)

(True Negative Rate) TNR = TN/(FP+TN)
(False Positive Rate) FPR = FP/(FP+TN)

TPR + FNR = 1 (100% of positive-class data)
TNR + FPR = 1 (100% of negative-class data)

	$\hat{P}$ (predicted)	$\hat{N}$ (predicted)
P (actual)	TP	FN
N (actual)	FP	TN

Evaluation Metrics

Classification

Prediction function y = f(x)

sample	N1	N2	P1	N3	P2	P3
input x	-4	-3	-2.5	-2	-1.5	-0.5
Prediction y	-1.1	-0.5	-0.1	0.2	0.6	0.9
Label	-1	-1	1	-1	1	1

If threshold set to be y=0,
N3, P2, P3 will be taken as +1
P1, N2, N1 will be taken as -1

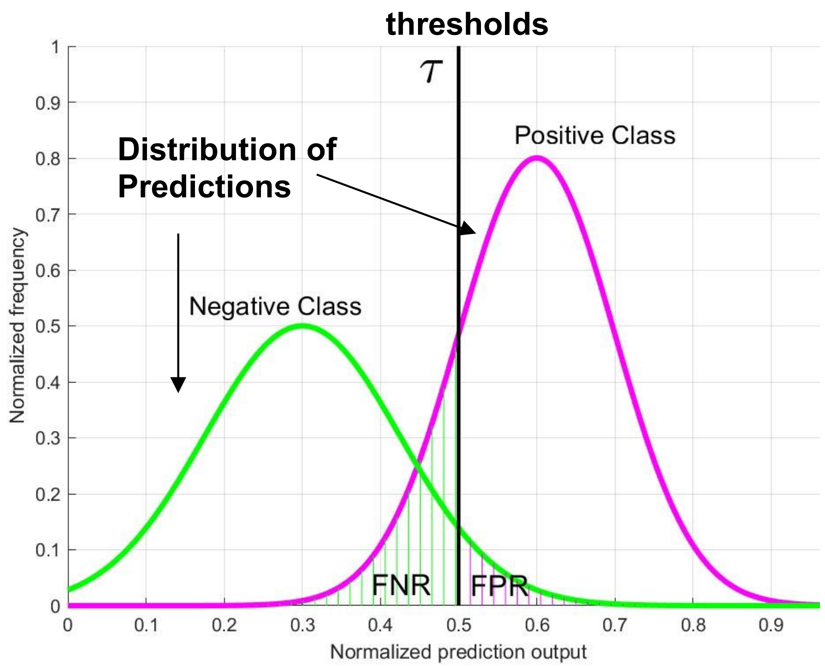
	$\hat{\mathrm{P}}$ (predicted)	$\hat{\mathrm{N}}$ (predicted)
P (actual)	$TP = 2$	$FN = 1$
N (actual)	$FP = 1$	$TN = 2$

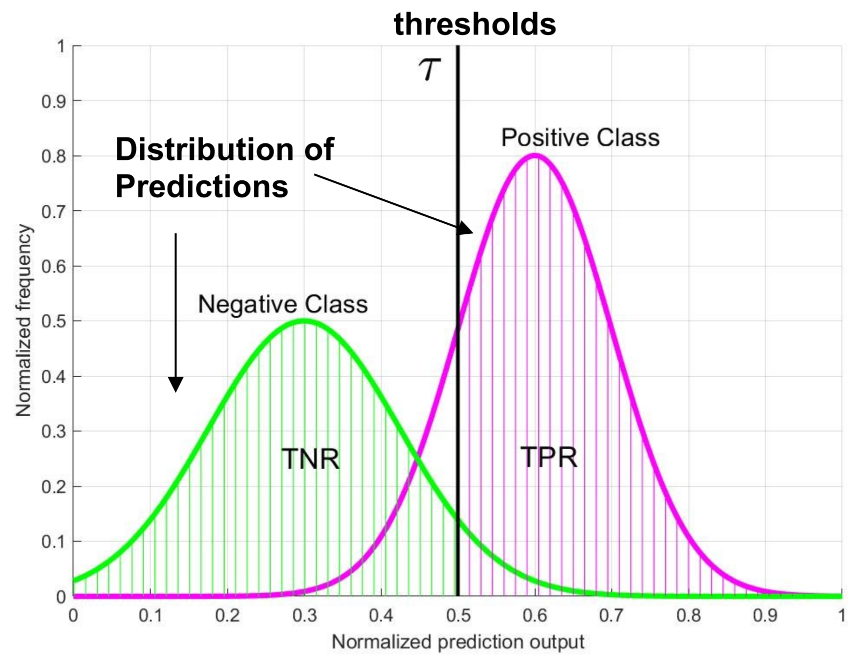
Evaluation Metrics

(True Positive Rate) TPR = TP/(TP+FN)
(False Negative Rate) FNR = FN/(TP+FN)
(True Negative Rate) TNR = TN/(FP+TN)
(False Positive Rate) FPR = FP/(FP+TN)

Classification

TPR + FNR = 1 (100% of positive-class data)
TNR + FPR = 1 (100% of negative-class data)





Normalized prediction output: change the output to the range of [0,1]
Reference: Lec 2, Page 25

Evaluation Metrics

Classification

Prediction function y = f(x)

We can change the threshold!

sample	N1	N2	P1	N3	P2	P3
input x	-4	-3	-2.5	-2	-1.5	-0.5
Prediction y	-1.1	-0.5	-0.1	0.2	0.6	0.9
Label	-1	-1	1	-1	1	1

If threshold set to be y=0.4,
P2, P3 will be taken as +1
N3, P1, N2, N1 will be taken as -1

	$\hat{P}$ (predicted)	$\hat{N}$ (predicted)
P (actual)	$TP = 2$	$FN = 1$
N (actual)	$FP = 0$	$TN = 3$

Evaluation Metrics

Classification:

TP, FP, FN, TN will change wrt thresholds!

If threshold set to be y=0,
N3, P2, P3 will be taken as +1
P1, N2, N1 will be taken as -1

	$\widehat{\mathrm{P}}$ (predicted)	$\widehat{\mathrm{N}}$ (predicted)
P (actual)	$TP = 2$	$FN = 1$
N (actual)	$FP = 1$	$TN = 2$

If threshold set to be y=0.4,
P2, P3 will be taken as +1
N3, P1, N2, N1 will be taken as -1

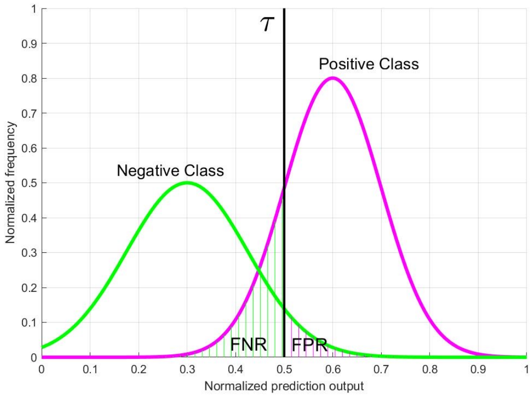
	$\widehat{\mathrm{P}}$ (predicted)	$\widehat{\mathrm{N}}$ (predicted)
P (actual)	$TP = 2$	$FN = 1$
N (actual)	$FP = 0$	$TN = 3$

Imagine we vary the thresholds at y!

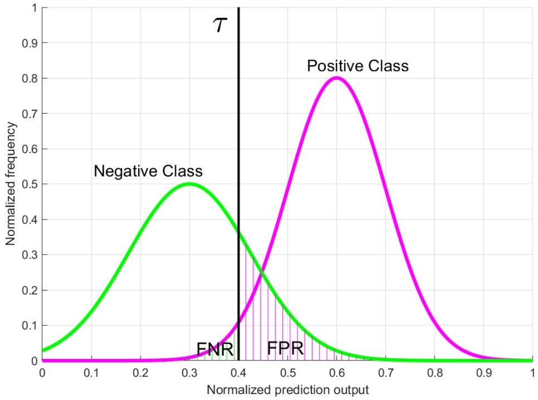
Evaluation Metrics

Classification

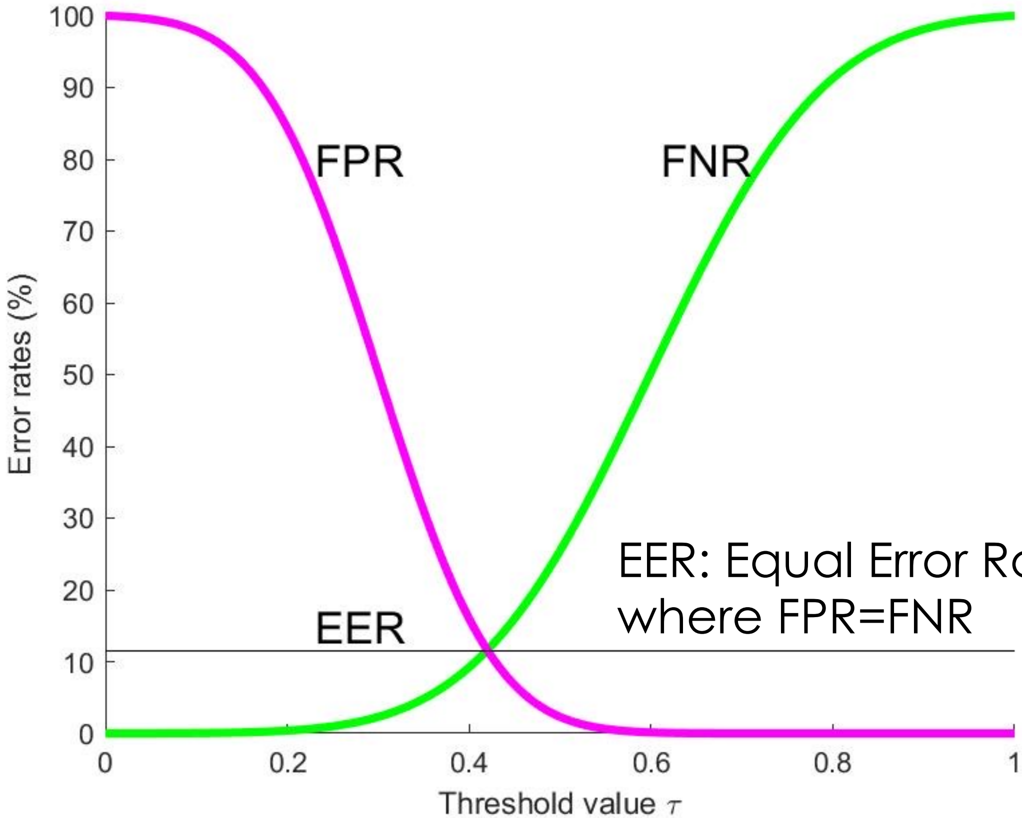
Threshold 1



Threshold 2



Sliding the threshold

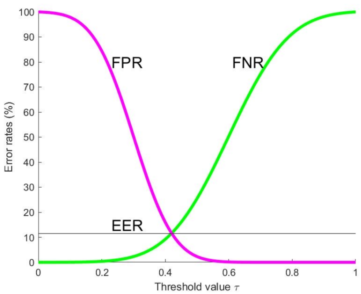


EER: Equal Error Rate where FPR=FNR

EER: Higher better or Lower better?

Lower is better

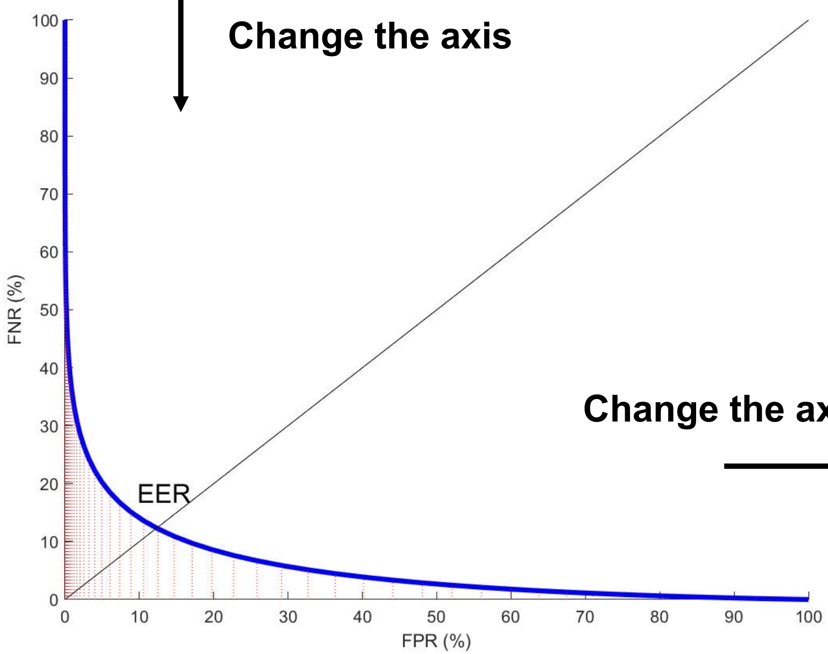
Evaluation Metrics



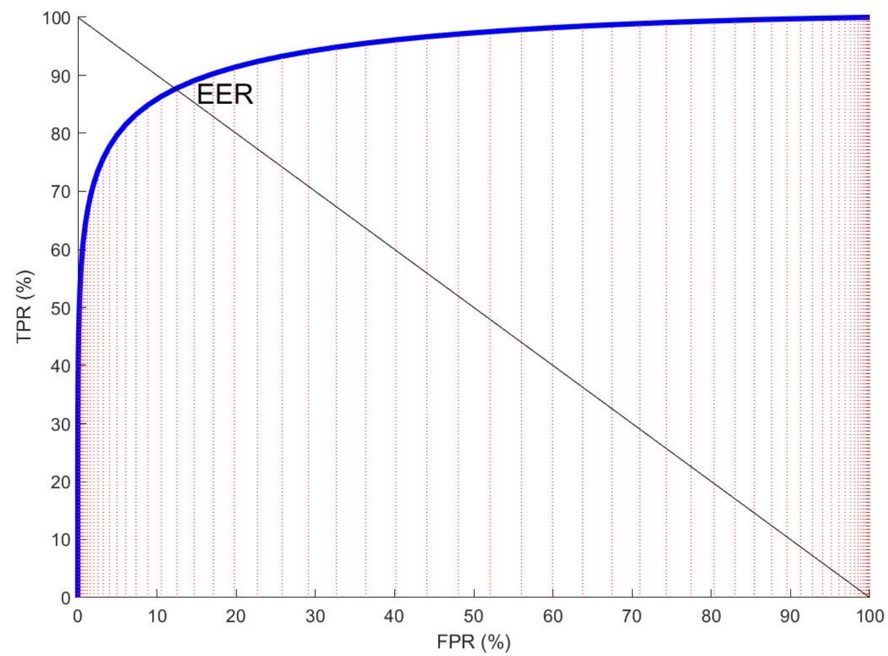
Change the axis

ROC: Widely used

TPR + FNR = 1



Change the axis



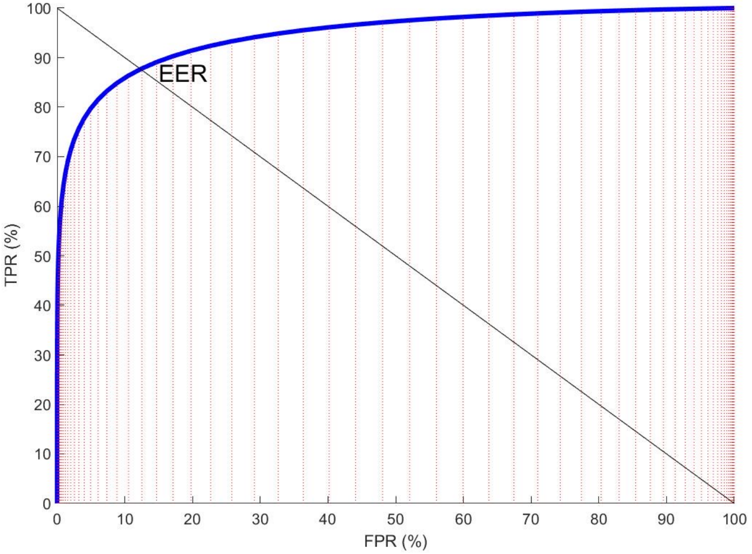
Detection Error Trade-off (DET) curve

Receiver Operating Characteristic (ROC) Curve

Evaluation Metrics

Area Under the Curve (ROC curve)

AUC provides an aggregate measure of performance across all possible classification thresholds.

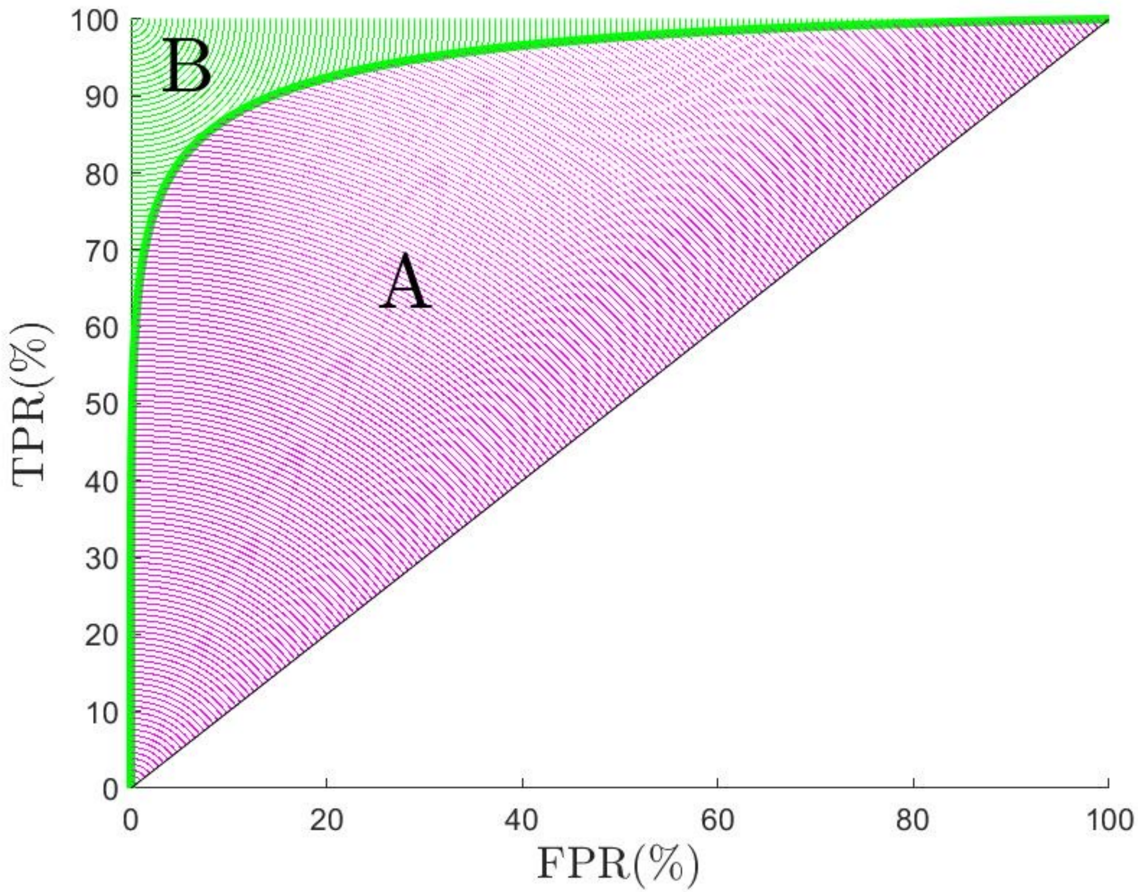


AUC ranges in value from 0 (100% wrong prediction) to 1 (100% correct prediction). It is classification-threshold-invariant.

Evaluation Metrics

Classification

Gini coefficient
G=A/(A+B)



Evaluation Metrics

Classification

Confusion Matrix for Multicategory Classification

	$P_{\hat{1}}$ (predicted)	$P_{\hat{2}}$ (predicted)		$P_{\hat{C}}$ (predicted)
P_1 (actual)	$P_{1,\hat{1}}$	$P_{1,\hat{2}}$	$\cdots$	$P_{1,\hat{C}}$
P_2 (actual)	$P_{2,\hat{1}}$	$P_{2,\hat{2}}$	$\cdots$	$P_{2,\hat{C}}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
P_C (actual)	$P_{C,\hat{1}}$	$P_{C,\hat{2}}$		$P_{C,\hat{C}}$

Other Issues

- Computational speed and memory consumptions are also important factors
  – Especially for mobile or edge devices

- Other factors
  – Parallelable, Modularity, Maintainability

- Not focus of this module


DANKE
THANK YOU
Danke
THANK YOU
MERCI
DANKE!
DANKE
THANK YOU!
GRAZIE
THANK YOU
Thank You
THANKS
GRAZIE
MERCI
GRACIAS
MERCI
DANKE SCHÖN
DANKE
Merci
Grazie
MERCI
THANKS
GRACIAS
Danke schön
THANK YOU!
Thanks!
THANK YOU
DANKE