CLASSIFICATION EVALUATION

Practical aspects in machine learning

Dr Zohar Barnett-Itzhaki

TODAY'S LECTURE

TODAY'S LECTURE

Evaluation approaches

TODAY'S LECTURE

Evaluation approaches

Numerical evaluation of model

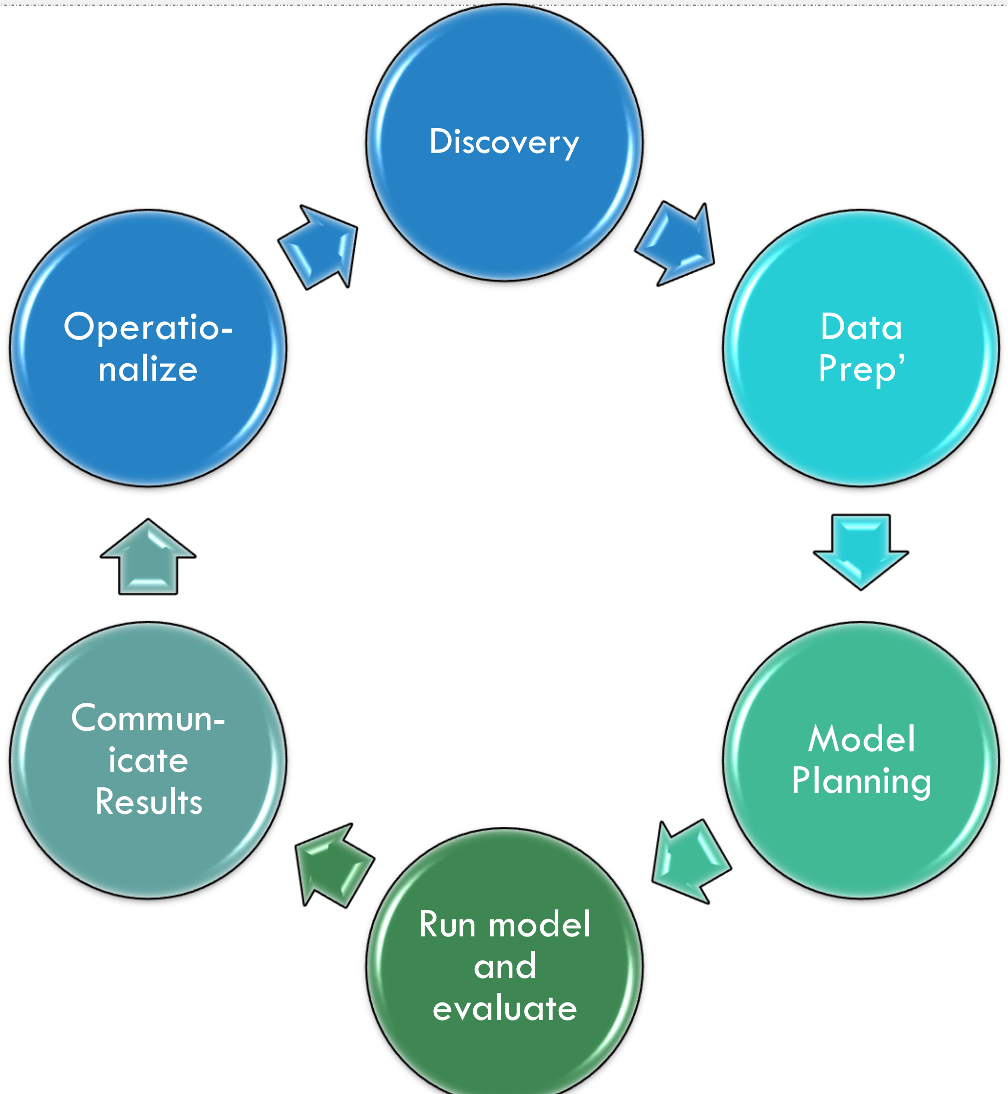
TODAY'S LECTURE

- Evaluation approaches
- Numerical evaluation of model
- ROC

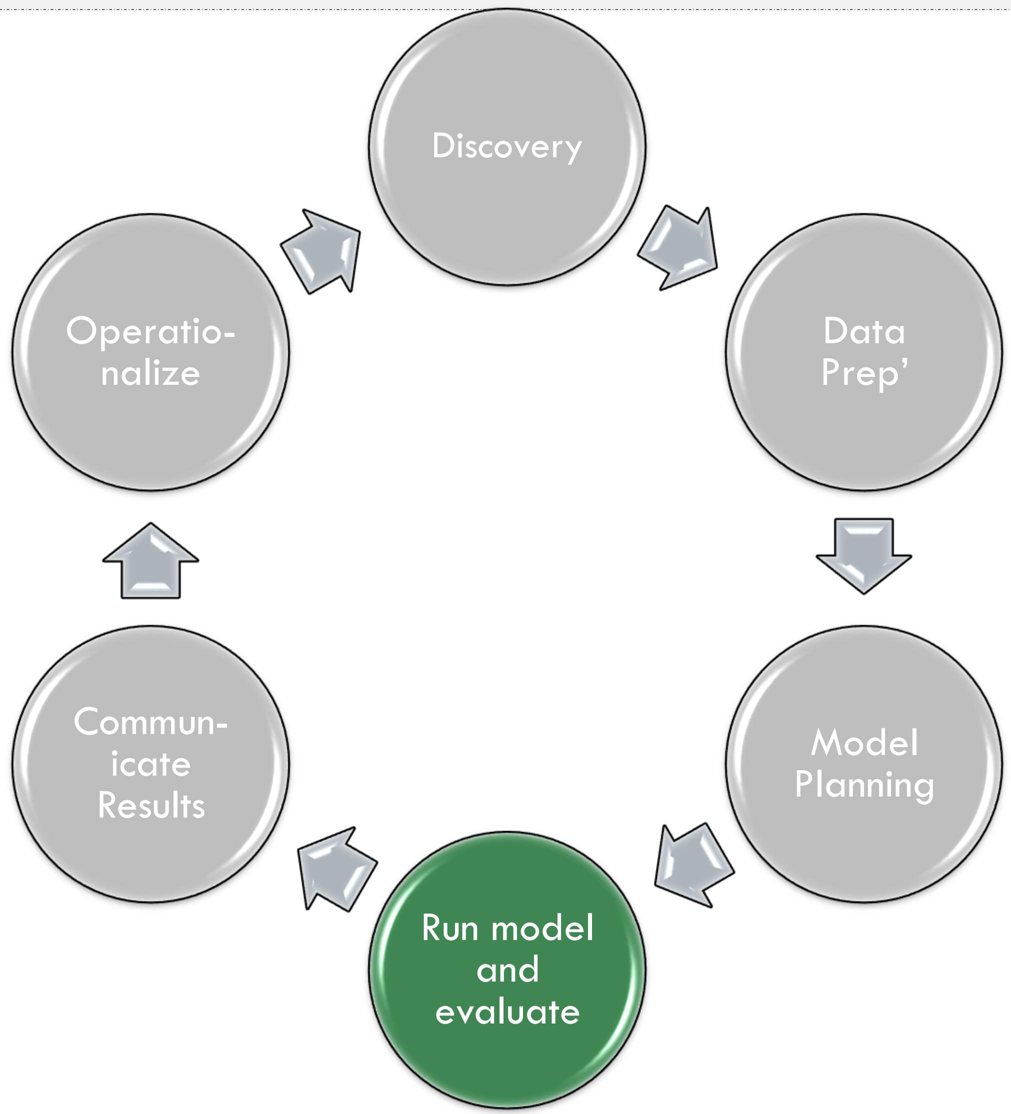
TODAY'S LECTURE

Evaluation approaches

DATA ANALYTICS LIFE CYCLE



DATA ANALYTICS LIFE CYCLE



TESTING OUR HYPOTHESES

Models (hypotheses) are built based on training samples

Then the models are testes on new samples

But how can we evaluate its performances?

How can we be sure it can generalize?

HYPOTHESIS EVALUATION APPROACHES

- Classical train / test approach
- Cross validation

HYPOTHESIS EVALUATION APPROACHES

- Classical train / test approach
- Cross validation

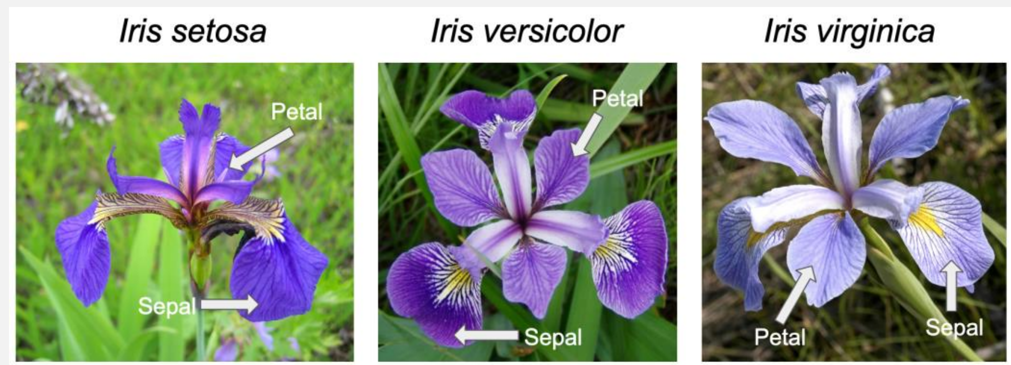
CLASSICAL APPROACH

- The common approach for evaluating a hypothesis is based on splitting the initial data to two groups:
  - Training set - usually 70% of the data
  - Test set – usually 30% of the data
- The split must be **randomized** so that both groups will be representative, why?

IMPORTANCE OF RANDOMIZATION

- Recall the Iris data:

50 rows: setosa
50 rows: versicolor
50 rows: virginica

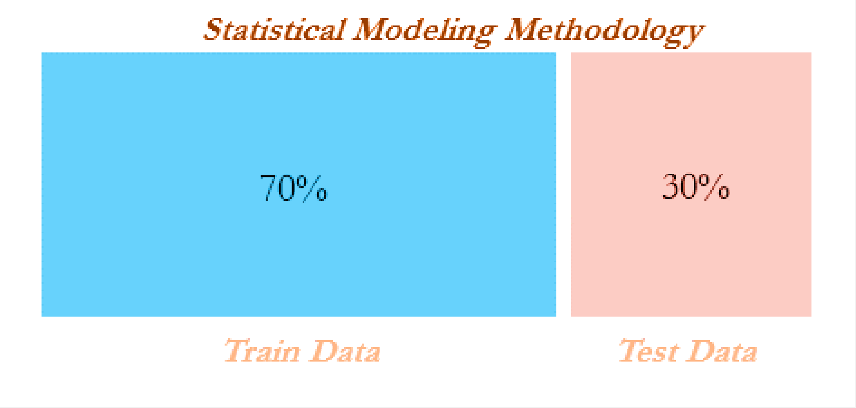


ML EVALUATION PROCESS USING THE CLASSICAL APPROACH

1. Randomly split the dataset into a training set (70%) and a test set (30%)

2. Learn a model based on the training set

3. Compute the error rate using the test set



*note: the 80%-20% approach is also acceptable

HYPOTHESIS EVALUATION APPROACHES

- Classical train / test approach
- Cross validation

WHAT MAY BE PROBLEMATIC IN THE TRAIN-TEST APPROACH?

- We might miss important data (we don't use 30% of the data to build the model)

- In small datasets this is crucial! – splitting small datasets will result in non significant datasets

CROSS VALIDATION

- Multiple random splits of the given samples to training and test sets
- This approach is widely used, also when we have big enough datasets

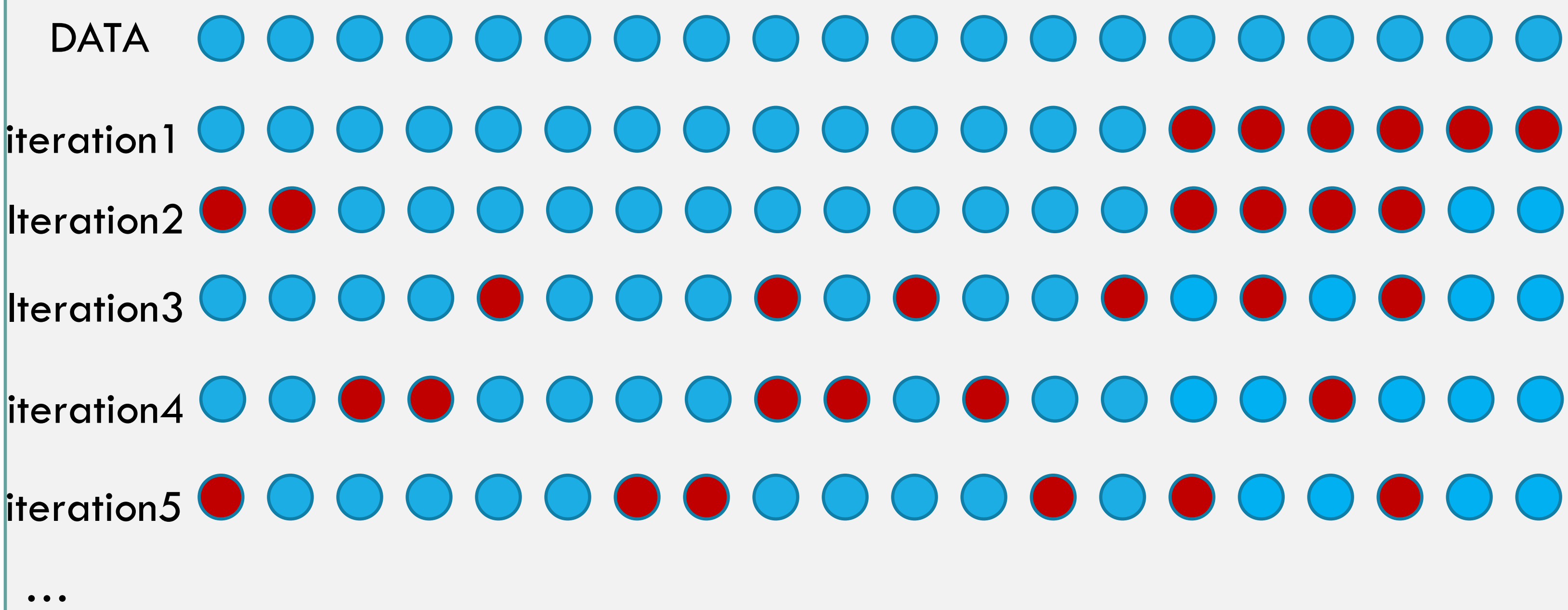
CROSS VALIDATION

1. Randomly split the data into a train set and a test set (70% /30%)
2. Learn a hypothesis on the training set
3. Test it on the test set and remember the performances
4. Repeat steps 1-3
5. Calculate the average performances

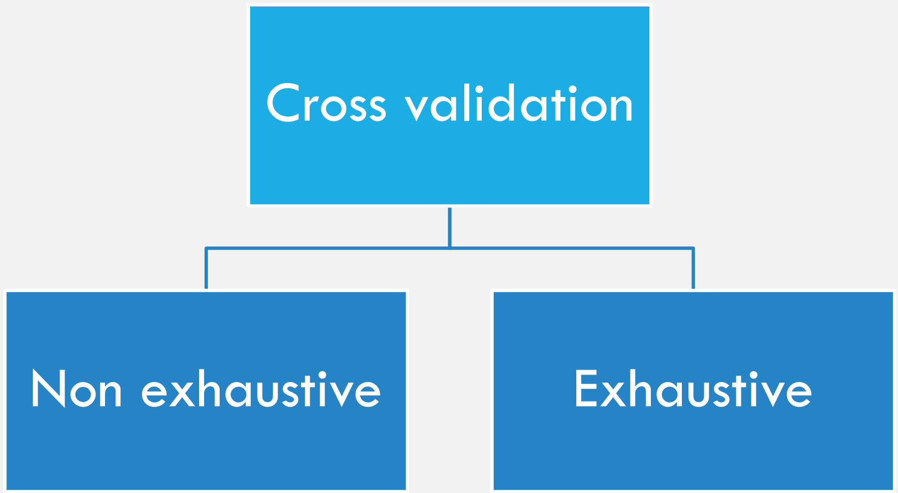
What does it remind you?

Random forests

CROSS VALIDATION



TYPES OF CROSS VALIDATION



TYPES OF CROSS VALIDATION

	Non exhaustive	Exhaustive
Number of partitions	Pre-defined (less than all possible partitions)	All possible partitions
Accurate?		
Time		

TYPES OF CROSS VALIDATION

	Non exhaustive	Exhaustive
Number of partitions	Pre-defined (less than all possible partitions)	All possible partitions
Accurate?	Less accurate	Very accurate
Time		

TYPES OF CROSS VALIDATION

	Non exhaustive	Exhaustive
Number of partitions	Pre-defined (less than all possible partitions)	All possible partitions
Accurate?	Less accurate	Very accurate
Time	rapid	slow

EXHAUSTIVE CROSS VALIDATION CAN BE EXPENSIVE…

- Say we have 1000 samples in the initial set
- We want to do an exhaustive cross validation
- 80% - 20%
- How many ways to choose 200 (test) out of 1000 ?

$$\binom{1000}{200} = \frac{1000!}{200! * 800!} = \frac{1000 * 999 * 998 * \ldots * 2 * 1}{200 * 199 * \ldots * 2 * 800 * 799 * \ldots * 1}$$

- $=$ 6.6172e+215 …

EXHAUSTIVE CROSS VALIDATION

- Leave one out cross validation (LOOCV)
  - The size of the test set it one ☺
  - Each iteration:
    - take one sample out
    - learn on the rest of the m-1 samples
    - Test on the sample taken out
  - Very common and efficient (and even fast)
  - Number of iterations = m (set size)

- Leave p-out cross validation
  - Every iteration, take p samples out, learn and test them
  - LOOCV is a special case when p=1
  - Can be slow (m over p iterations)

TODAY'S LECTURE

Evaluation approaches

TODAY'S LECTURE

Evaluation approaches

Numerical evaluation of model

ML EVALUATION PROCESS USING THE CLASSICAL APPROACH –(TWO CLASSES)

- Given a model
- We train it on the training set
- Then we want to test it on the test set
- How?

ML EVALUATION PROCESS USING THE CLASSICAL APPROACH –(TWO CLASSES)

$err(h_\theta(x),y) = 1$ if:

$h_\theta(x)=1, y=0$
$h_\theta(x)=0, y=1$

0 otherwise

$$Test\ error = \sum_{i=1}^{m\ of\ test} err(h\theta(x^{(i)}, y^{(i)}))$$

Test error rate = test error / m

ACCURACY

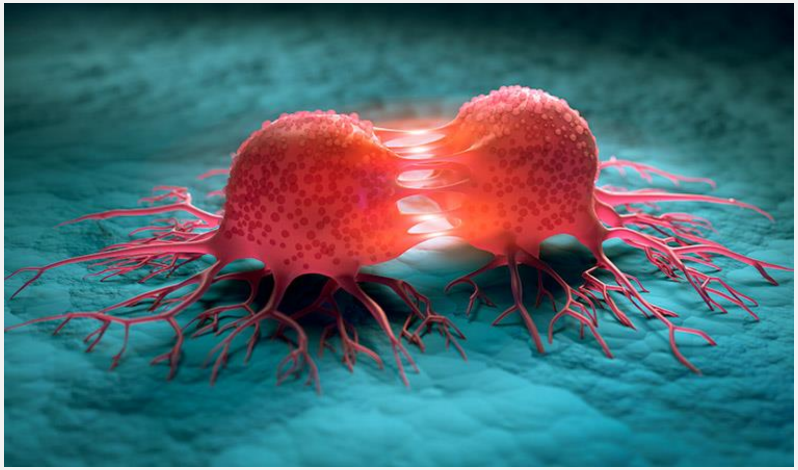
- Accuracy is a basic evaluation measure
- It means how well are the predictions
- Accuracy = 100% - test error rate

OTHER EVALUATION APPROACHES

Say we build a classifier to detect cancer

The classifier error rate is 1%

Is this a good classifier?



THE SKEWED CLASSES EXAMPLE

What if only 1% of the population have cancer?

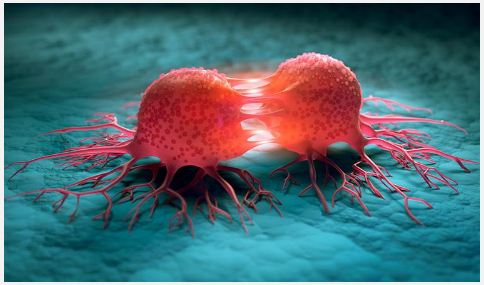
This is the skewed (unbalanced) classes problem

Theoretically we can use the following algorithm:

for each x_i:

$y_i = 0$

Is this a good classifier?



CONFUSION MATRIX

Actual

Predicted

	1	0
1	**True positive**	**False Positive**
0	**False Negative**	**True Negative**

PRECISION

- Of all patients for whom we predicted True, what fraction actually have cancer?

- $\frac{\text{True positive}}{\text{\# predicted positive}} = \frac{TP}{TP+FP}$

$0 \leq \text{Precision} \leq 1$

		Actual	
Predicted		1	0
	1	TP	FP
	0	FN	TN

Dr. Zohar Barnett-Itzhaki

RECALL

- Of all patients for that actually have cancer, what is the fraction we correctly detected as having cancer ?

- $\frac{\text{True positive}}{\text{\# actual positive}} = \frac{TP}{TP+FN}$

$0 \leq \text{Recall} \leq 1$

		Actual	
		1	0
Predicted	1	TP	FP
	0	FN	TN

RECALL AND PRECISION

- What is the precision and recall of our 1% (always 0) algorithm?

- TP = 0 so both recall and precision are 0

- When both recall and precision are high – it means that we are on our right way, and that the algorithm is probably good.

Actual

Predicted

	1	0
1	TP	FP
0	FN	TN

RECALL AND PRECISION EXAMPLES

- Say I want to predict if today will be rainy (1) or dry (0)
- I examine a 100 days data and create a model based on 70% of the days
- Then I test it on the remaining 30% days
- I get the following results:

- What is the precision?
- P = TP/(TP+FP) = 12/(12+3) = 0.8

- What is the recall?
- R = TP/(TP+FN) = 12/(12+5) = 0.71

Actual days of rain

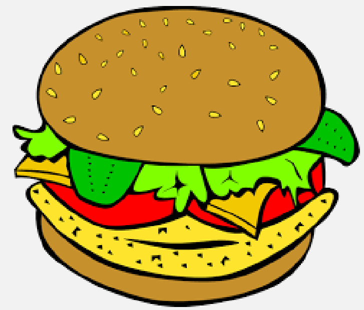
Predicted rain

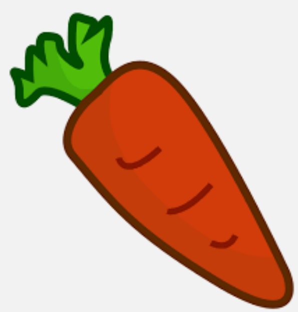
	Rain (1)	Dry (0)
Rain (1)	12	3
Dry (0)	5	10

Dr. Zohar Barnett-Itzhaki

RECALL AND PRECISION EXAMPLES

I want to develop an algorithm that predicts healthy food (1). Our test set is:





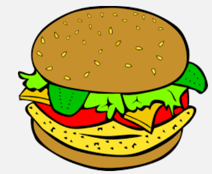
RECALL AND PRECISION EXAMPLES

We predicted the following as health:

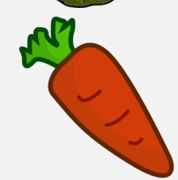

Actual

Predicted

	1	0
1	TP 2	FP 0
0	FN 1	TN 2







- What is the precision?
- P = TP/(TP+FP) = 2/(2+0) = 1

- What is the recall?
- R = TP/(TP+FN) = 2/(2+1) = 2/3

Dr. Zohar Barnett-Itzhaki

RECALL AND PRECISION EXAMPLES

We predicted the following as health:

- What is the precision?
- P = TP/(TP+FP) = 3/(3+1) = 3/4

- What is the recall?
- R = TP/(TP+FN) = 3/(3+0) = 1

F1 SCORE

$$\underline{\mathbf{F1}} = \frac{2PR}{P + R}$$

P = 0 / R = 0 → F1 = 0

P = 1 AND R = 1 → F1 = 1

$0 \leq F1 \leq 1$

TODAY'S LECTURE

- Evaluation approaches
- Numerical evaluation of model

TODAY'S LECTURE

Evaluation approaches

Numerical evaluation of model

ROC

EVALUATION APPROACHES

Leave X out cross validation average error rate

Basic measurements:

- Accuracy (% correct predictions)
- Recall
- Precision
- F1 score

ROC

RECEIVER OPERATING CHARACTERISTIC CURVE (ROC)

So far we learned what are:

- TP, FP, TN, FN
- Recall (also called TPR), Precision

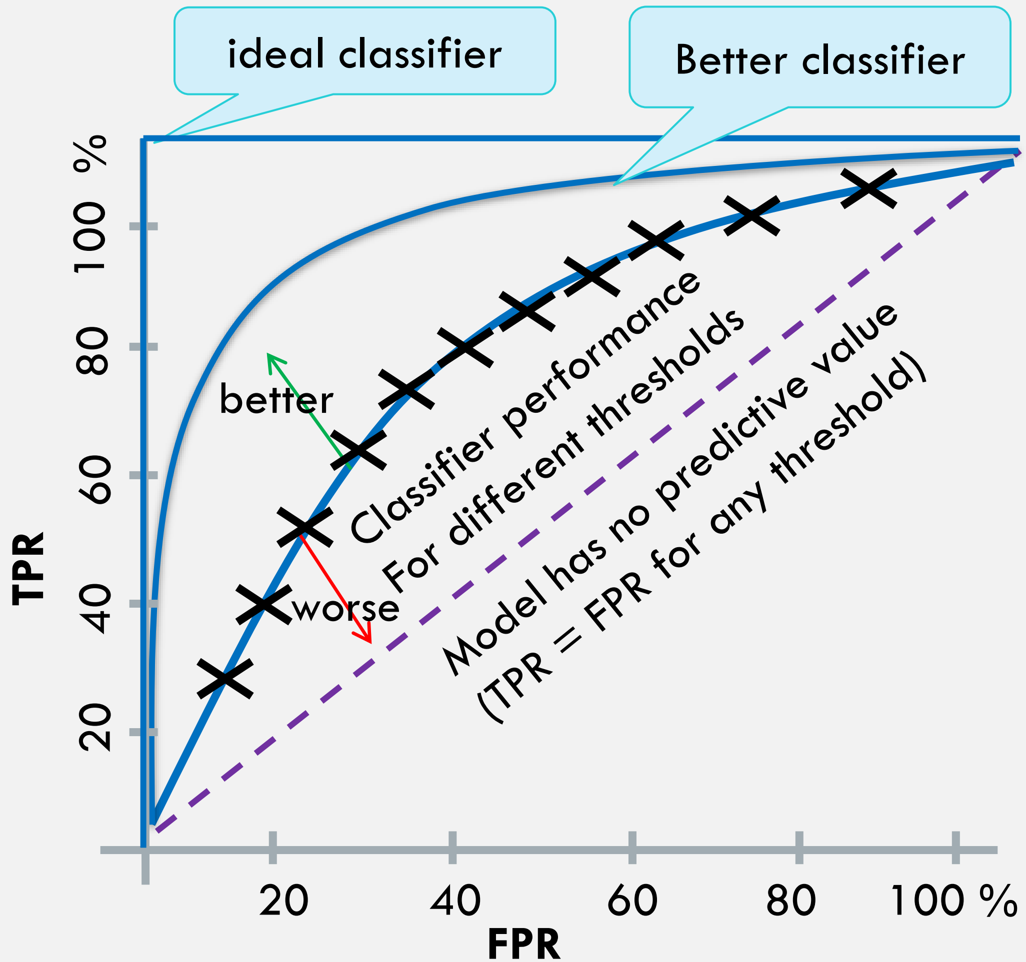
An additional metric:

- FPR = 1 - specificity = 1-(TN/N)

RECEIVER OPERATING CHARACTERISTIC CURVE (ROC)

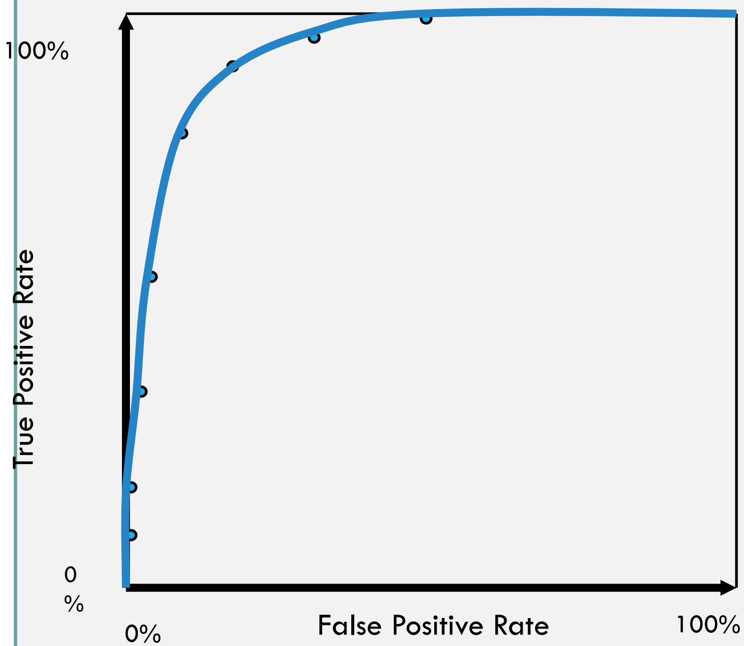
- This curve plots the dependencies between True Positive Rate (TPR) and the False Positive Rate (FPR) at various thresholds (threshold in Logistic regression etc…)
- The idea:
  - In random models, there is a linear relationship between TPR and FPR.
  - In good models we can achieve higher TPR with relatively lower FPR (more "true" and less "false")

RECEIVER OPERATING CHARACTERISTIC CURVE (ROC)

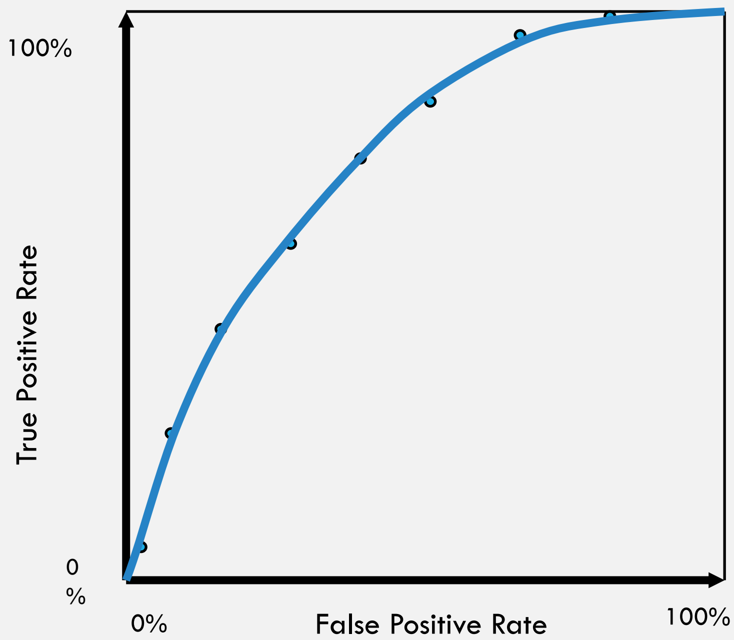


ROC CURVE COMPARISONS

A good performance



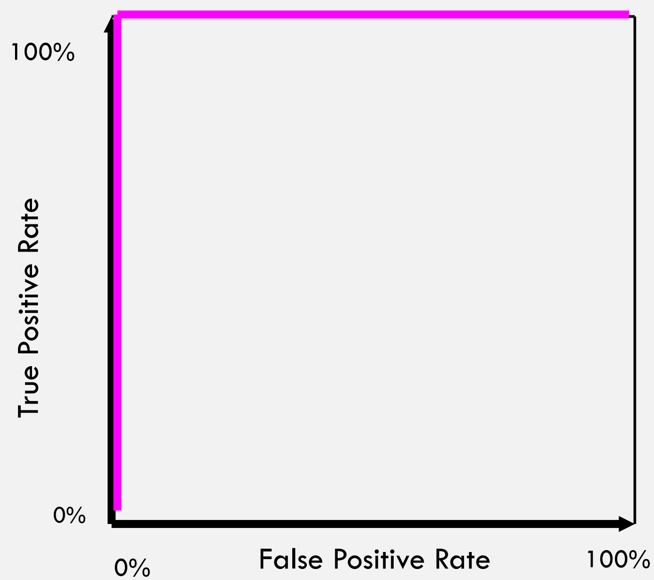
A poor performance



Dr. Zohar Barnett-Itzhaki

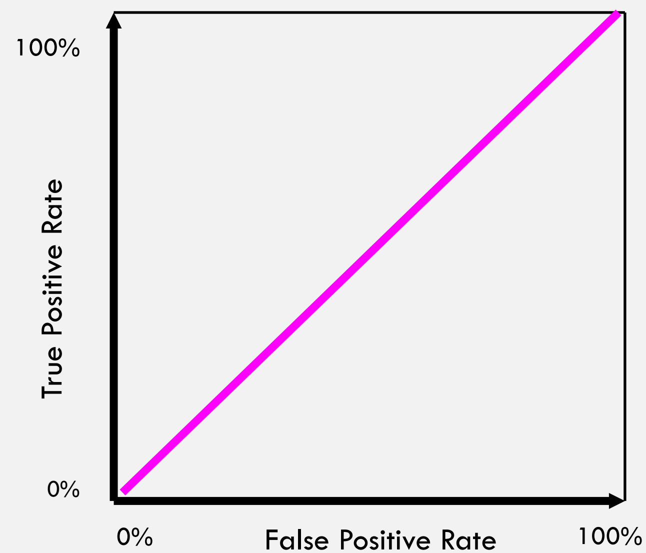
ROC CURVE COMPARISONS

Best Performance



The distributions don't overlap at all

Worst Performance



The distributions overlap completely

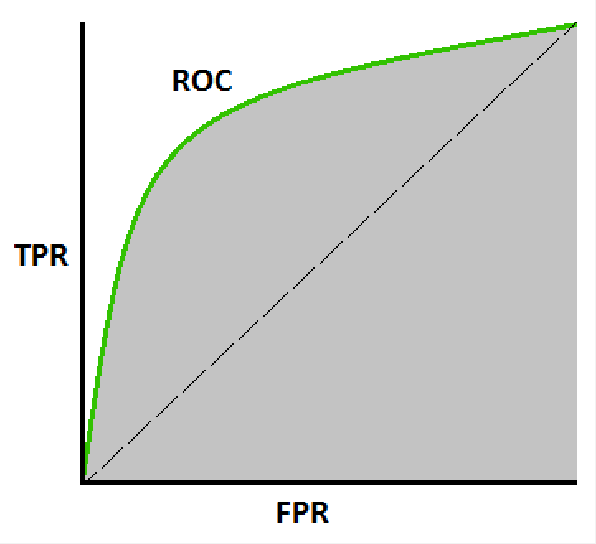
ROC AND AUC

We plot the ROC to understand the qualities of our model

The maximum area under the curve (AUC) is 1

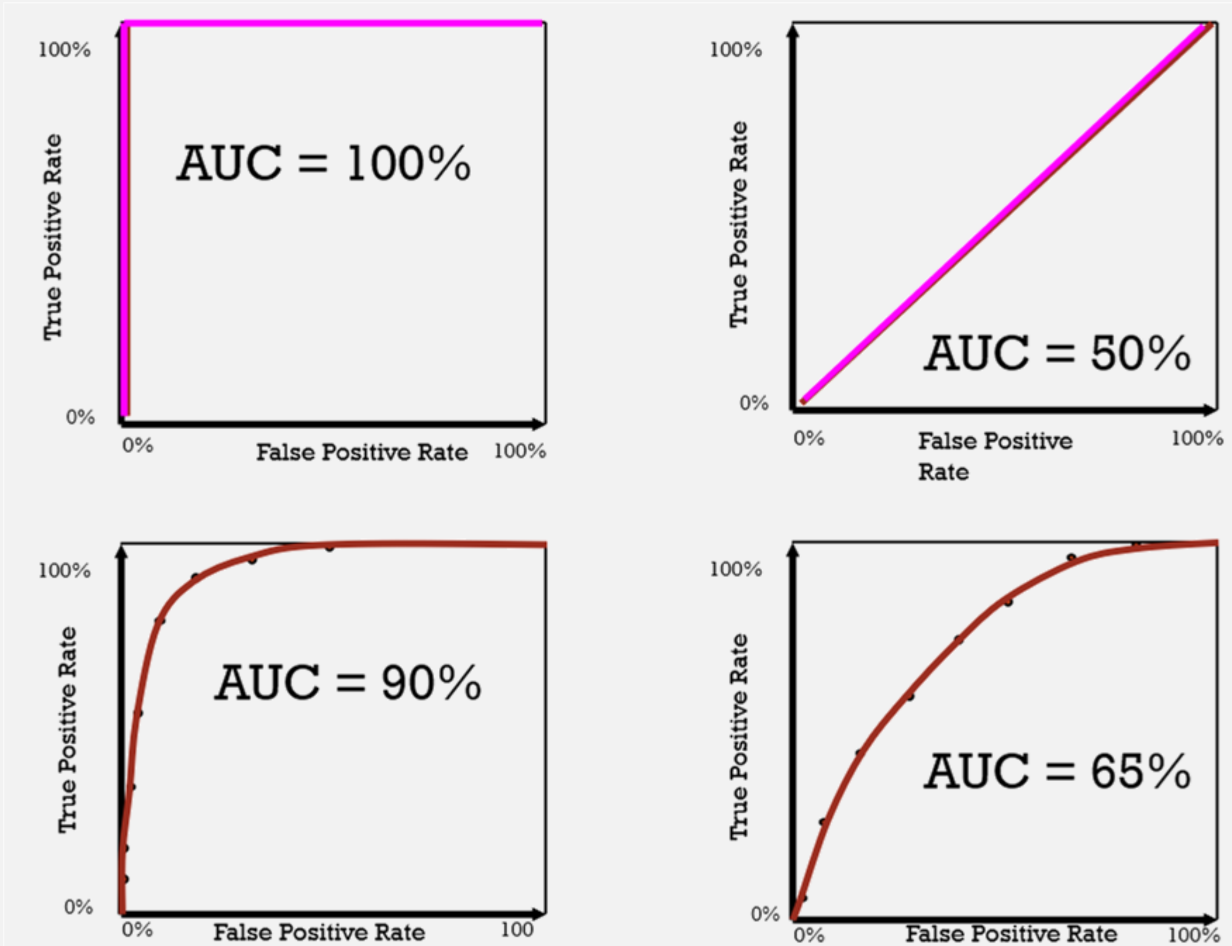
Random predictions have an AUC of 0.5 (half figure)

Better models will result in better AUC values (more TPR and less FPR)



Dr. Zohar Barnett-Itzhaki

ROC AND AUC



100%
True Positive Rate
AUC = 100%
0%
0%
False Positive Rate
100%
100%
True Positive Rate
AUC = 50%
0%
0%
False Positive Rate
100%
100%
True Positive Rate
AUC = 90%
0%
0%
False Positive Rate
100
100%
True Positive Rate
AUC = 65%
0%
0%
False Positive Rate
100%

SUMMARY

It is important to evaluate the ML performance

In order to correctly choose the correct model, use the 60%/20%/20% approach

Best evaluation methods are: calculating accuracy, recall, precision and F1score, in addition to ROC

Q