

# **Diagnostics**

**Gad Kimmel**

# Outline

- Introduction.
- Cross validation.
- Bootstrap method.
- ROC plot.

# Introduction

# Motivation

- Given a learning algorithm, evaluate its quality.
  - Let the algorithm be a function  $f_\theta: \mathcal{R}^n \rightarrow \mathcal{R}$ , where  $\theta$  is a parameter (or a set of parameters) of  $f$ .
  - What is the best parameter for  $f$ ?
  - How many parameters should we choose?
- Given data samples  $x_1, x_2, \dots, x_N$ , evaluate some estimator, say the average.
  - It is also important to know how good this estimator is. How can we evaluate it?

# Example

- The yield of the Dow Jones Index over the past two years is  $\sim 12\%$ .
- Your friend tells you he made a yield of  $18\%$ , just by picking specific stocks from the Dow Jones.
- Did your friend really achieve a higher yield than the Dow Jones Index, or is just by chance?
- Put it differently, what is the distribution of the yield of randomly chosen stocks?

# **Bootstrap Method**

# Evaluating Accuracy

- A simple method for accuracy estimation is to provide the bias or variance of the estimator.
- For example, suppose the samples are independently identically distributed (i.i.d.), with finite variance.
  - We know, by the central limit theorem, that

$$\frac{n^{1/2}(\bar{x}_n - \mu)}{\sigma} \rightarrow Z \sim N(0,1)$$

- Roughly speaking,  $\bar{x}_n$  is normally distributed with expectation  $\mu$  and variance  $\sigma^2/n$ .

# Assumptions Do Not Hold

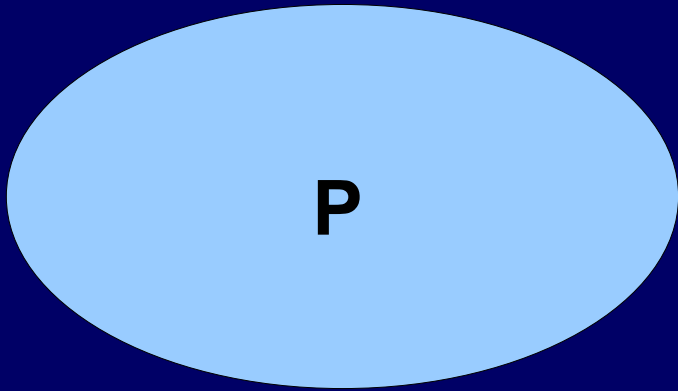
- What if the r.v. are **not** i.i.d. ?
- What if we want to evaluate another estimator (and not  $\bar{x}_n$  )?
- It would be nice to have many different samples of samples.
- In that case, one could calculate the estimator for each sample of samples, and infer its distribution.
- But... we don't have it.

# Solution - Bootstrap

- Repeated re-estimation of a parameter using random samples with replacement from the original data.

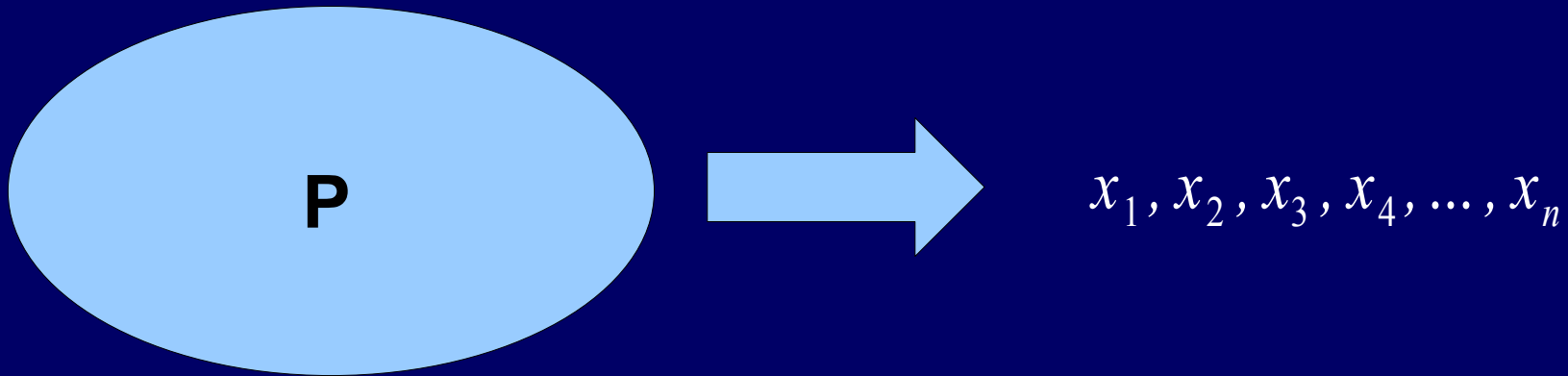
# Bootstrap - Illustration

- Goal: Sampling from P.



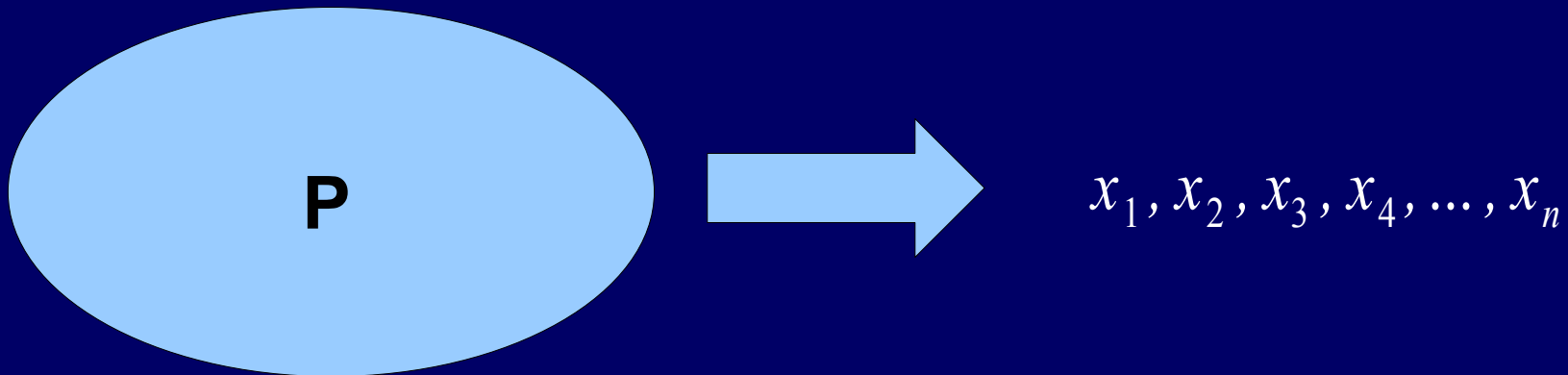
# Bootstrap - Illustration

- Goal: Sampling from P.



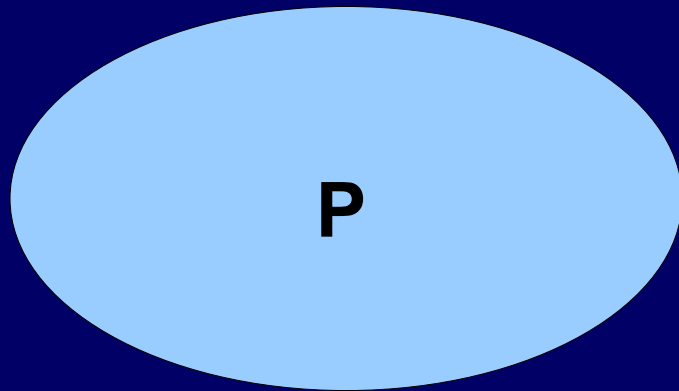
# Bootstrap - Illustration

- Goal: Sampling from P.



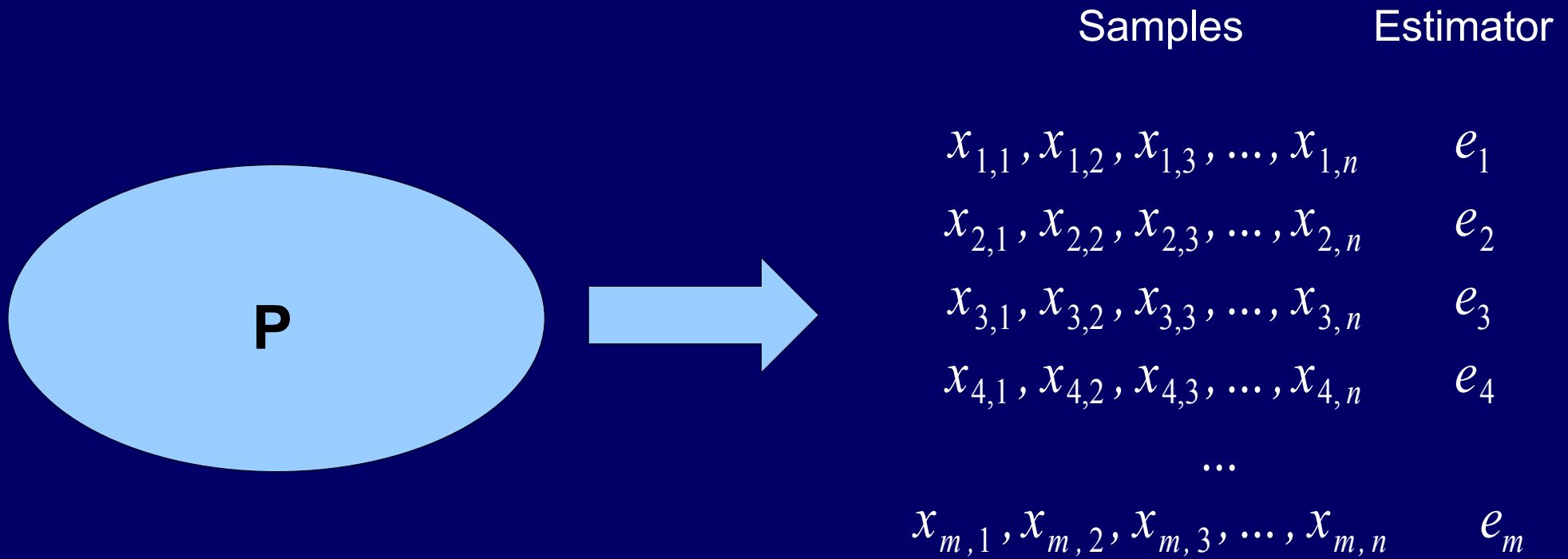
... in order to estimate the variance of an estimator.  
E.g., estimator is the mean of the elements in the sample.

# Bootstrap - Illustration



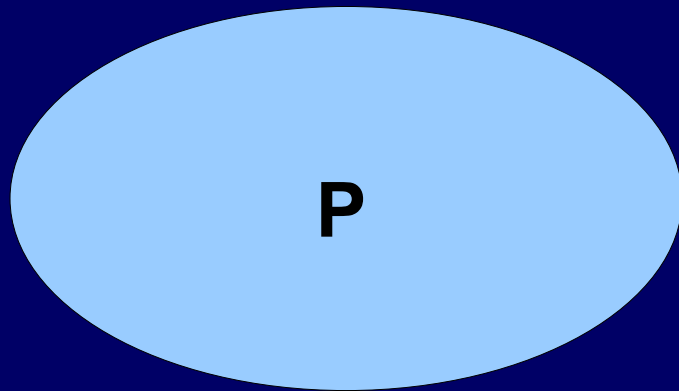
Samples	Estimator
$x_{1,1}, x_{1,2}, x_{1,3}, \dots, x_{1,n}$	$e_1$
$x_{2,1}, x_{2,2}, x_{2,3}, \dots, x_{2,n}$	$e_2$
$x_{3,1}, x_{3,2}, x_{3,3}, \dots, x_{3,n}$	$e_3$
$x_{4,1}, x_{4,2}, x_{4,3}, \dots, x_{4,n}$	$e_4$
...	
$x_{m,1}, x_{m,2}, x_{m,3}, \dots, x_{m,n}$	$e_m$

# Bootstrap - Illustration



- What is the variance of  $e$  ?

# Bootstrap - Illustration

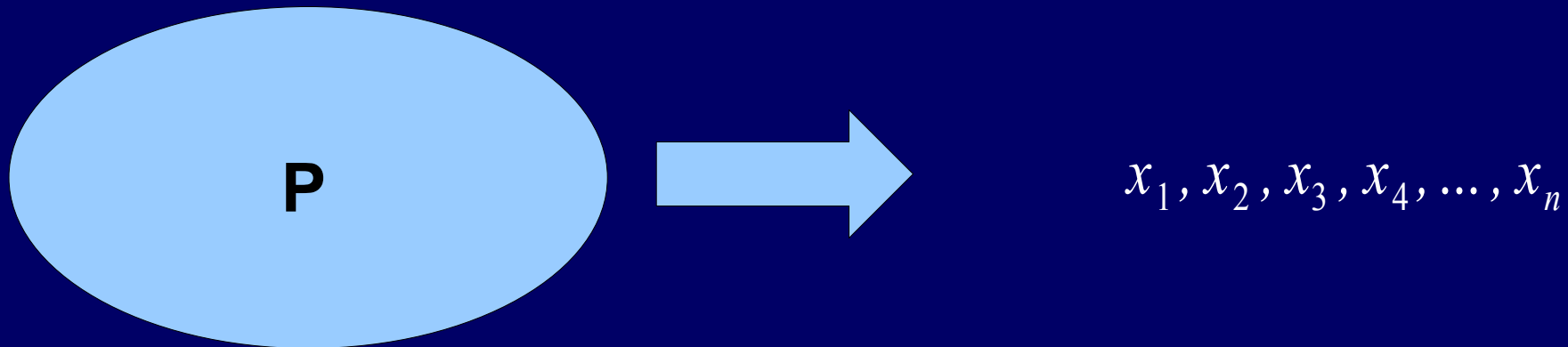


Samples	Estimator
$x_{1,1}, x_{1,2}, x_{1,3}, \dots, x_{1,n}$	$e_1$
$x_{2,1}, x_{2,2}, x_{2,3}, \dots, x_{2,n}$	$e_2$
$x_{3,1}, x_{3,2}, x_{3,3}, \dots, x_{3,n}$	$e_3$
$x_{4,1}, x_{4,2}, x_{4,3}, \dots, x_{4,n}$	$e_4$
...	
$x_{m,1}, x_{m,2}, x_{m,3}, \dots, x_{m,n}$	$e_m$

- Estimate the variance by  $\text{var}(e) = \frac{1}{m} \sum_{i=1}^m (e_i - \hat{\mu})^2$

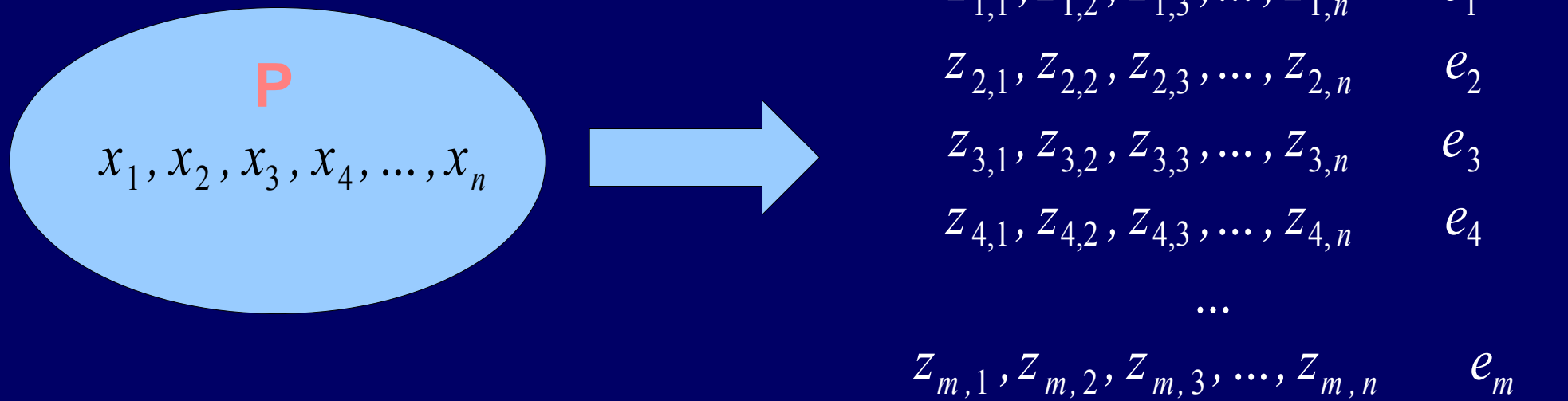
# Bootstrap - Illustration

- We only have 1 sample:



# Bootstrap - Illustration

- Sampling is done from the **empirical distribution**.



# Formalization

- The data is  $(x_1, x_2, \dots, x_n) \sim P$ . Note that the distribution function  $P$  is unknown.
- We sample  $m$  samples  $Y_1, Y_2, \dots, Y_m$ .  
 $Y_i = (z_{i,1}, z_{i,2}, \dots, z_{i,n})$  contains  $n$  samples drawn from the empirical distribution of the data:

$$\Pr[z_{j,k} = x_i] = \frac{1}{\#x_i}$$

Where  $\#x_i$  is the number of times  $x_i$  appears in the original data.

# The Main Idea

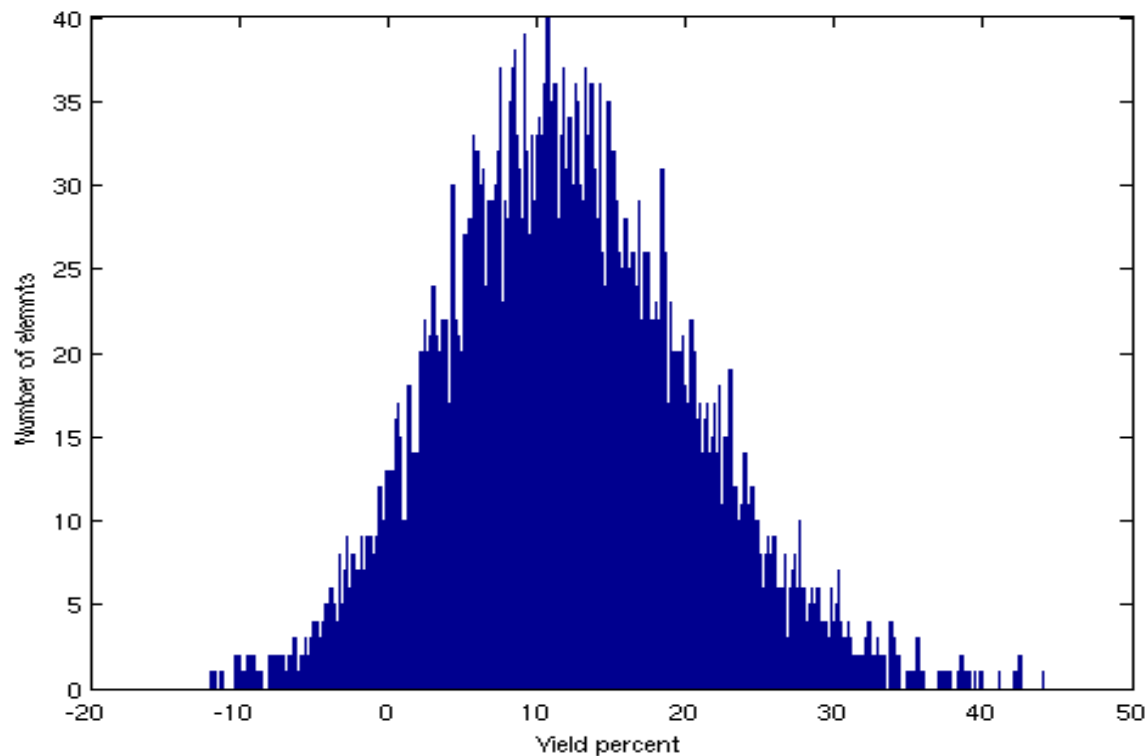
- $Y_i \sim \hat{P}$  .
- We wish that  $P = \hat{P}$  . Is it (always) true? NO.
- Rather,  $\hat{P}$  is an approximation of  $P$  .

# Example 1

- The yield of the Dow Jones Index over the past two years is  $\sim 12\%$ .
- Your friend tells you he made a yield of  $18\%$ , just by picking specific stocks from the Dow Jones.
- Let  $x$  be a r.v. that represents the yield of randomly selected stocks.
- Do we know the distribution of  $x$ ?

# Example 1 (cont.)

- Prepare a sample  $x_1, x_2, \dots, x_{10,000}$ , where each  $x_i$  is the yield of randomly selected stocks.
- Approximate the distribution of  $x$  using this sample.

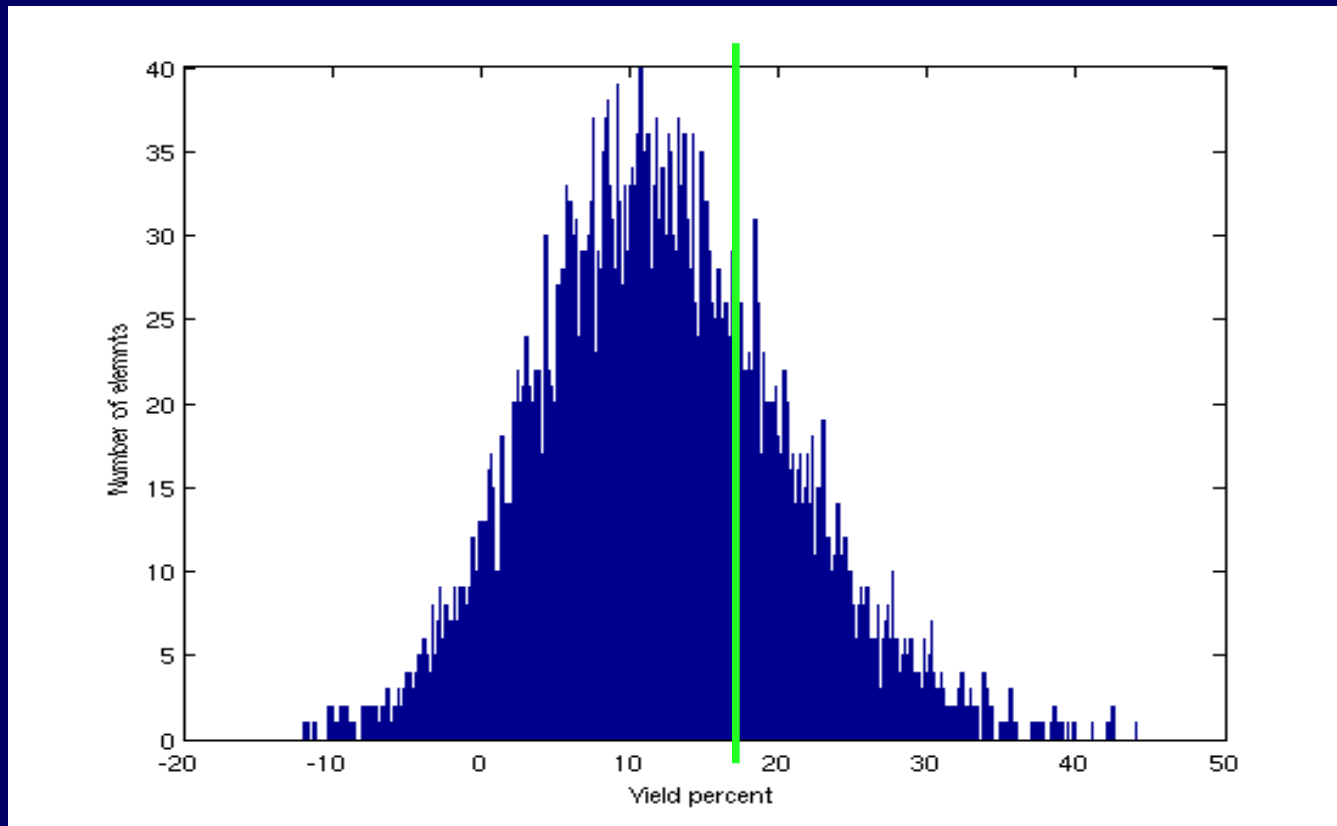


# Evaluation of Estimators

- Using the approximate distribution, we can evaluate estimators. E.g.:
  - Variance of the mean.
  - Confidence intervals.

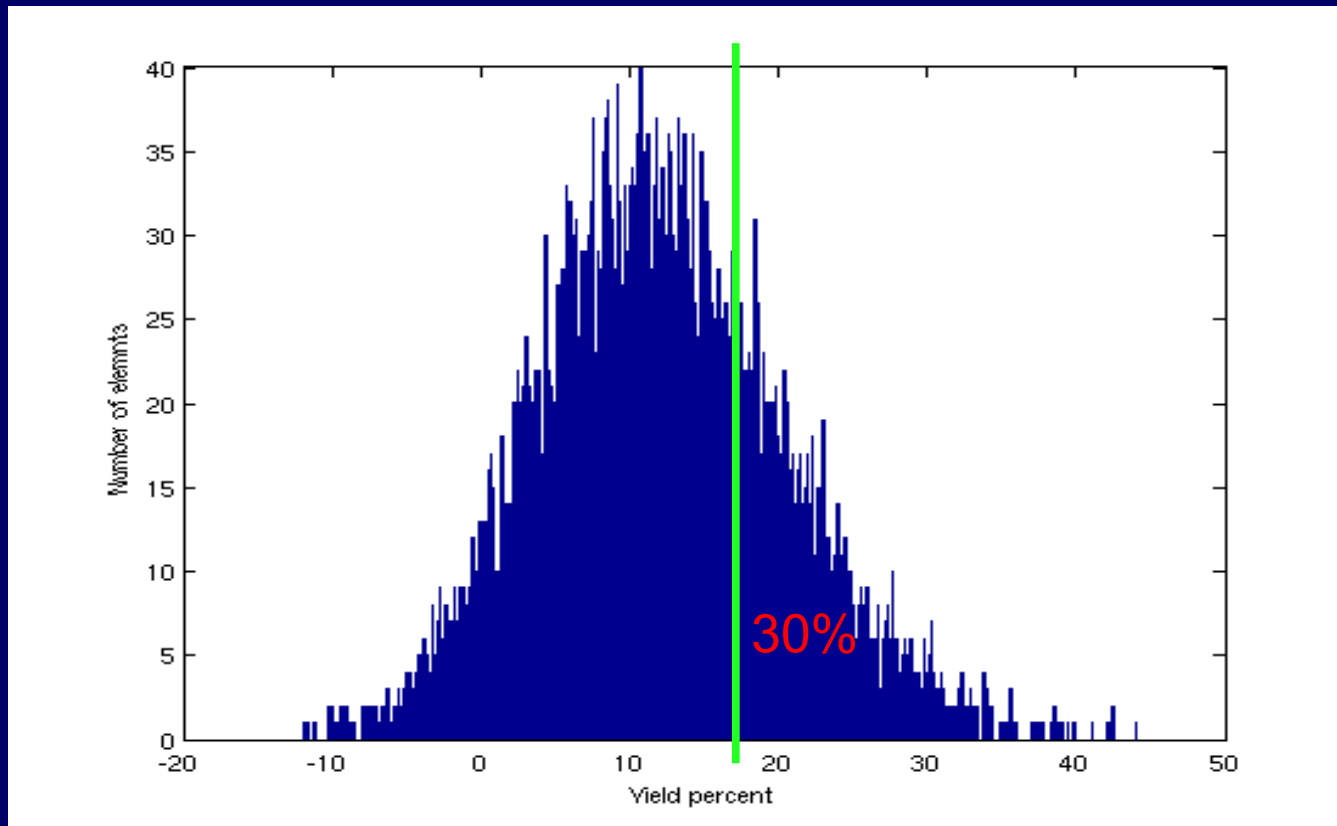
# Example 1 (cont.)

- What is the probability to obtain yield larger than 18% (p-value)?



# Example 1 (cont.)

- What is the probability to obtain yield larger than 18% (p-value)?



# Example 2 - Decision tree

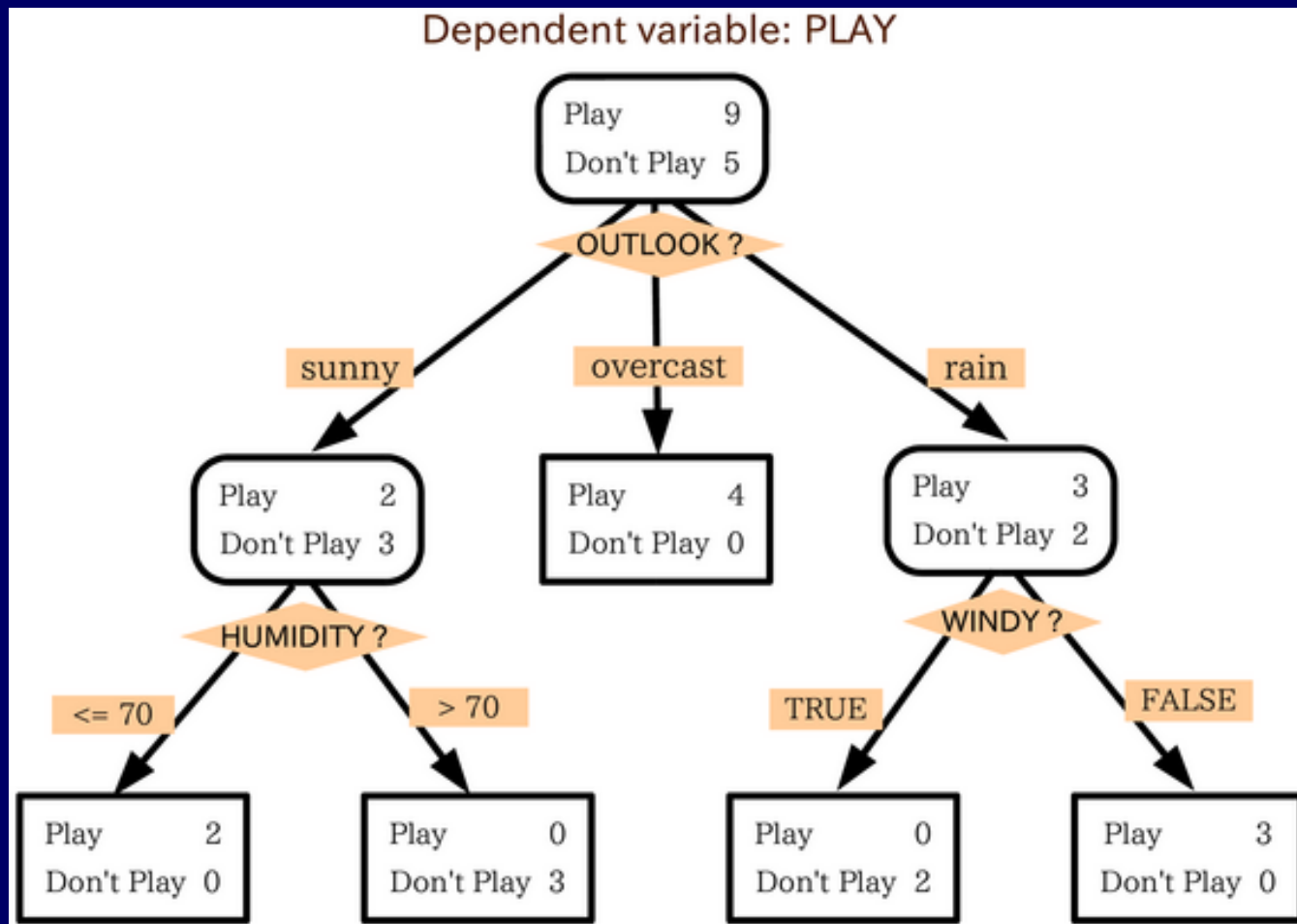
- Decision tree - short introduction.

Play golf dataset

Independent variables				Dep. var
OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY
sunny	85	85	FALSE	Don't Play
sunny	80	90	TRUE	Don't Play
overcast	83	78	FALSE	Play
rain	70	96	FALSE	Play
rain	68	80	FALSE	Play
rain	65	70	TRUE	Don't Play
overcast	64	65	TRUE	Play
sunny	72	95	FALSE	Don't Play
sunny	69	70	FALSE	Play
rain	75	80	FALSE	Play
sunny	75	70	TRUE	Play
overcast	72	90	TRUE	Play
overcast	81	75	FALSE	Play
rain	71	80	TRUE	Don't Play

# Example 2

- Building a decision tree.



## Example 2

- Many other trees can be built, using different algorithms.
- For a specific tree one can calculate prediction accuracy:

$$\frac{\text{\# of elements classified correctly}}{\text{total \# of elements}}$$

## Example 2

- Many other trees can be built, using different algorithms.
- For a specific tree one can calculate prediction accuracy:

$$\frac{\text{\# of elements classified correctly}}{\text{total \# of elements}}$$

- For calculating error bars for this value, we need to sample more, apply the algorithm many times, and each time evaluate the prediction.

# Example 2 - Applying Bootstrap

Build decision tree for each sample.

```
graph TD; A[Build decision tree for each sample.] --> B[Calculate prediction for each tree.]; B --> C[Evaluate error bars based on predictions.]
```

Calculate prediction for each tree.

Evaluate error bars based on predictions.

# Example 2 - Applying Bootstrap

$T_1, T_2, \dots, T_n$

$p_1, p_2, \dots, p_n$

$\mu \pm 1.96 \text{ STD}(p_1, p_2, \dots, p_n)$

Build decision tree for each sample.

Calculate prediction for each tree.

Evaluate error bars based on predictions.

# Example 2 - Applying Bootstrap

But we have  
only **one** data  
set !

Build decision  
tree for each  
sample.

Calculate prediction  
for each tree.

Evaluate error bars  
based on predictions.

# Example 2 - Applying Bootstrap

Use bootstrap to prepare many samples.



Build decision tree for each sample.



Calculate prediction for each tree.



Evaluate error bars based on predictions.

# Cross Validation

# Objective

- Model selection.

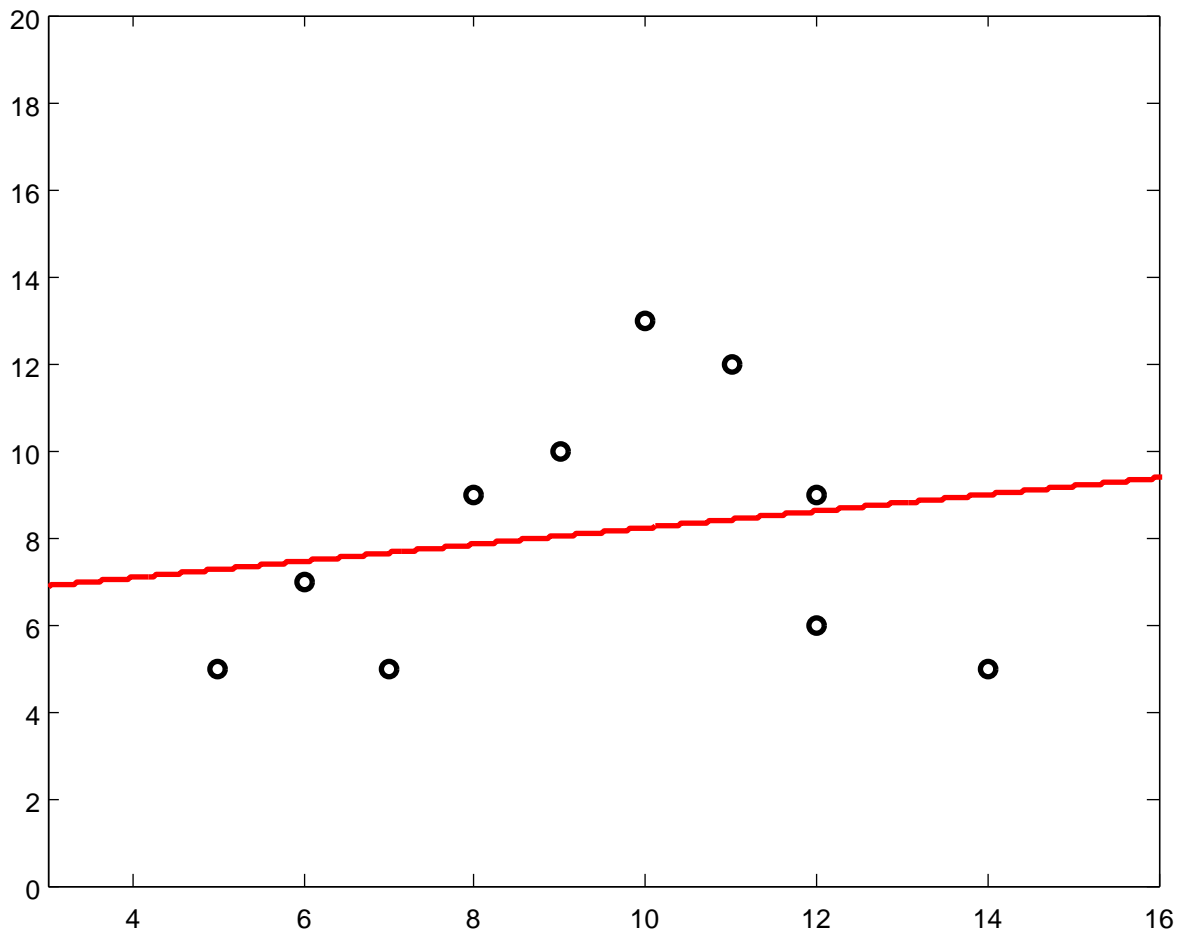
# Formalization

- Let  $(x, y)$  drawn from distribution  $P$ . Where  $x \in \mathcal{X}^n$  and  $y \in \mathcal{Y}$
- Let  $f_\theta: \mathcal{X}^n \rightarrow \mathcal{Y}$  be a learning algorithm, with parameter(s)  $\theta$ .

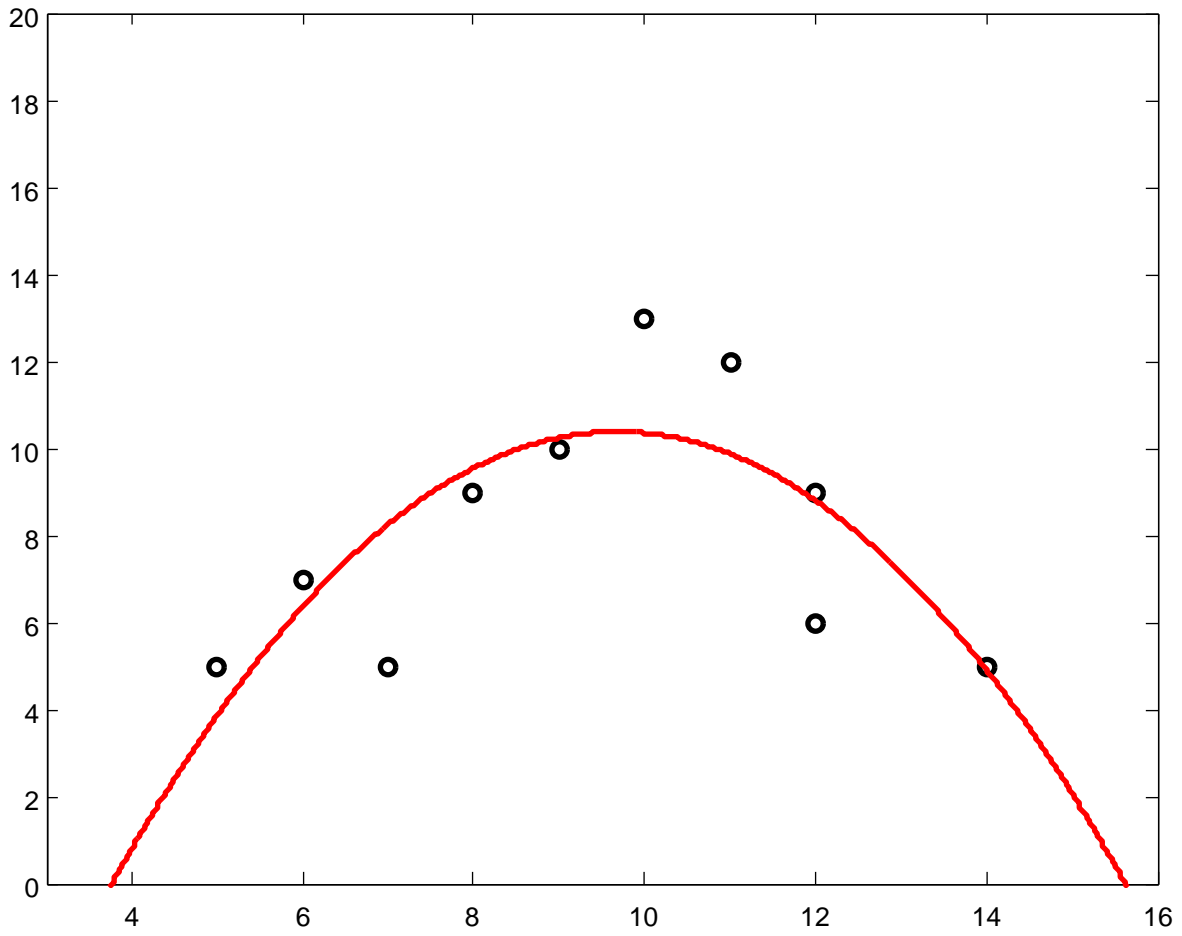
# Example

- Regression model.

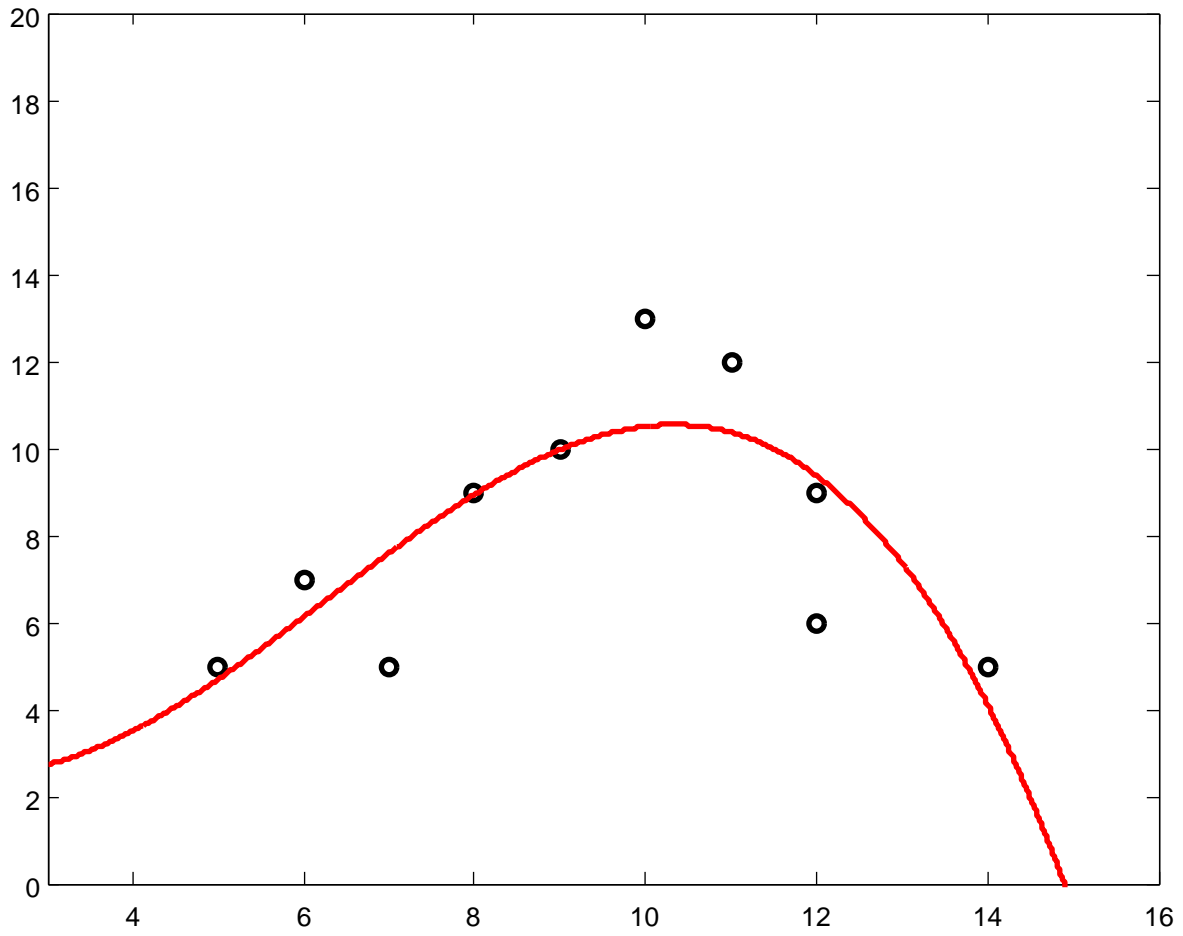
Regression – Order of 1 (Linear)



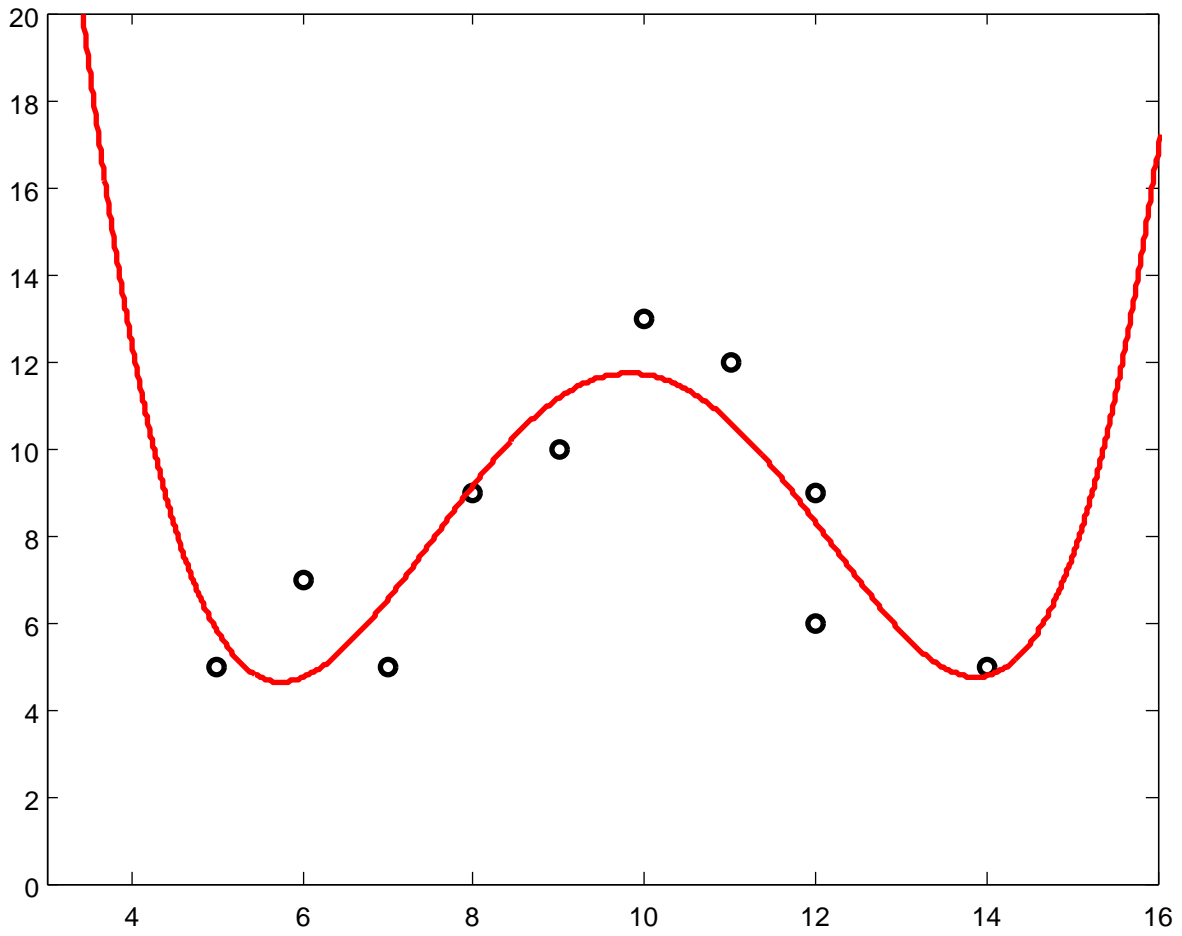
Regression – Order of 2



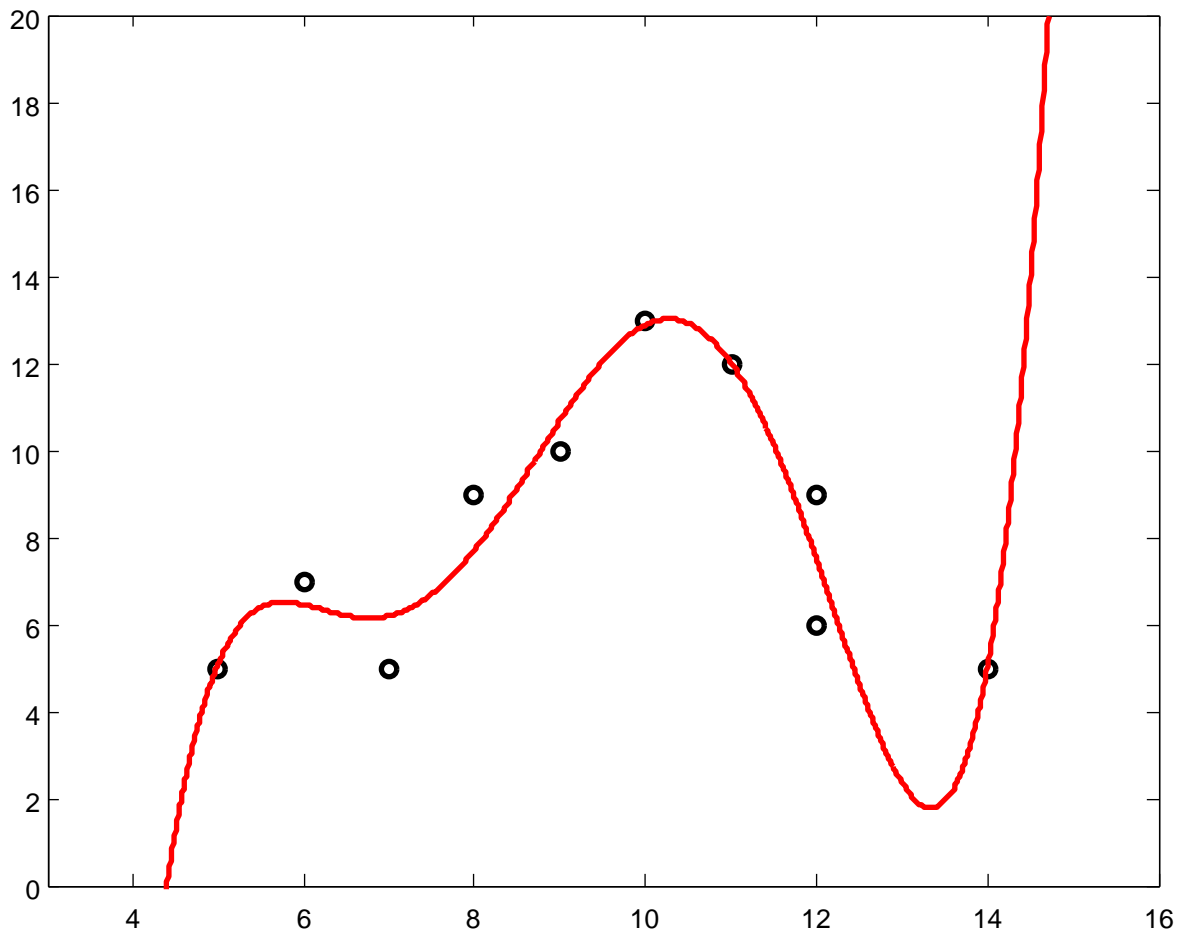
Regression – Order of 3



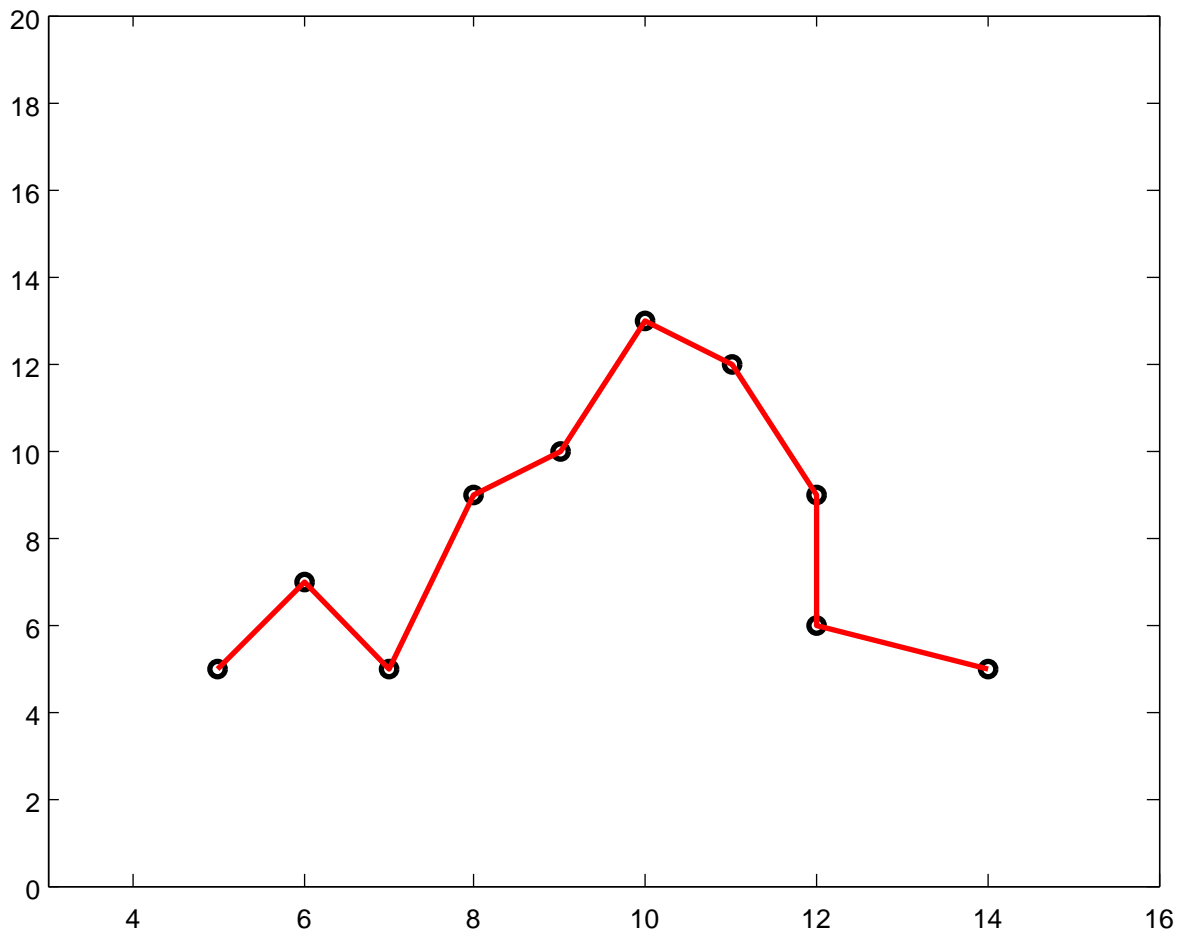
Regression – Order of 4



Regression – Order of 5



## Regression – Join the Dots



# What Do We Want?

- We want the method that is going to predict future data most accurately, assuming they are drawn from the distribution  $P$ .

# Choosing the Best Model

- For a sample  $(x, y)$  which is drawn from the distribution function  $P$  :

$$(f_{\theta}(x) - y)^2$$

or

$$|(f_{\theta}(x) - y)|$$

- Since  $(x, y)$  is a r.v. we are usually interested in:

$$E[(f_{\theta}(x) - y)^2]$$

# Choosing the Best Model (cont.)

- Choose the parameter(s)  $\theta$  :

$$\operatorname{argmin}_{\theta} \mathbb{E}[(f_{\theta}(x) - y)^2]$$

- The problem is that we don't know to sample from  $\mathcal{P}$ .

# Solution - Cross Validation

- Partition the data to 2 sets:
  - Training set  $T$ .
  - Test set  $S$ .
- Calculate  $\theta$  using **only** the training set  $T$ .
- Given  $\theta$ , calculate

$$\frac{1}{|S|} \sum_{(x_i, y_i) \in S} (f_{\theta}(x_i) - y_i)^2$$

# Back to the Example

- In our case, we should try different orders for the regression (or different # of params).
- Each time apply the regression only on the training test, and calculate estimation error on the test set.
- The # of parameters will be the one minimizing the error.

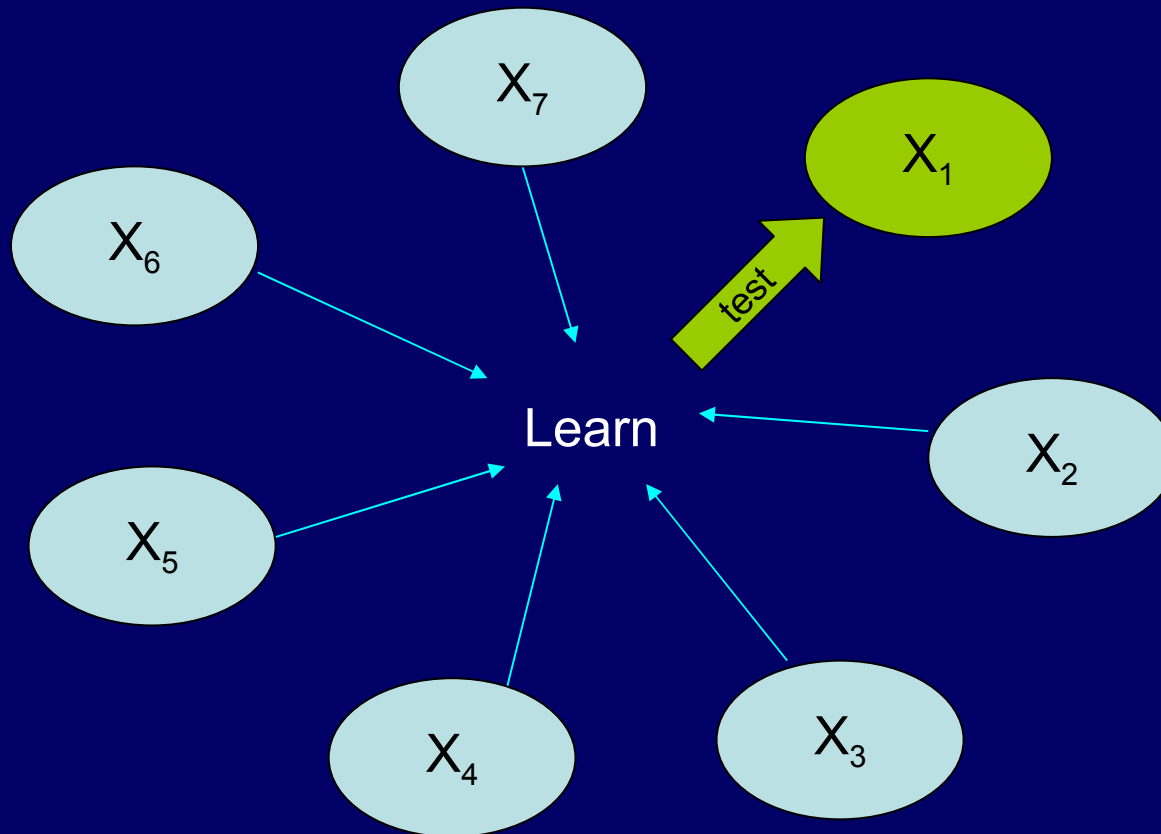
# Variants of Cross Validation

- Test - set.
- Leave one out.
- k-fold cross validation.

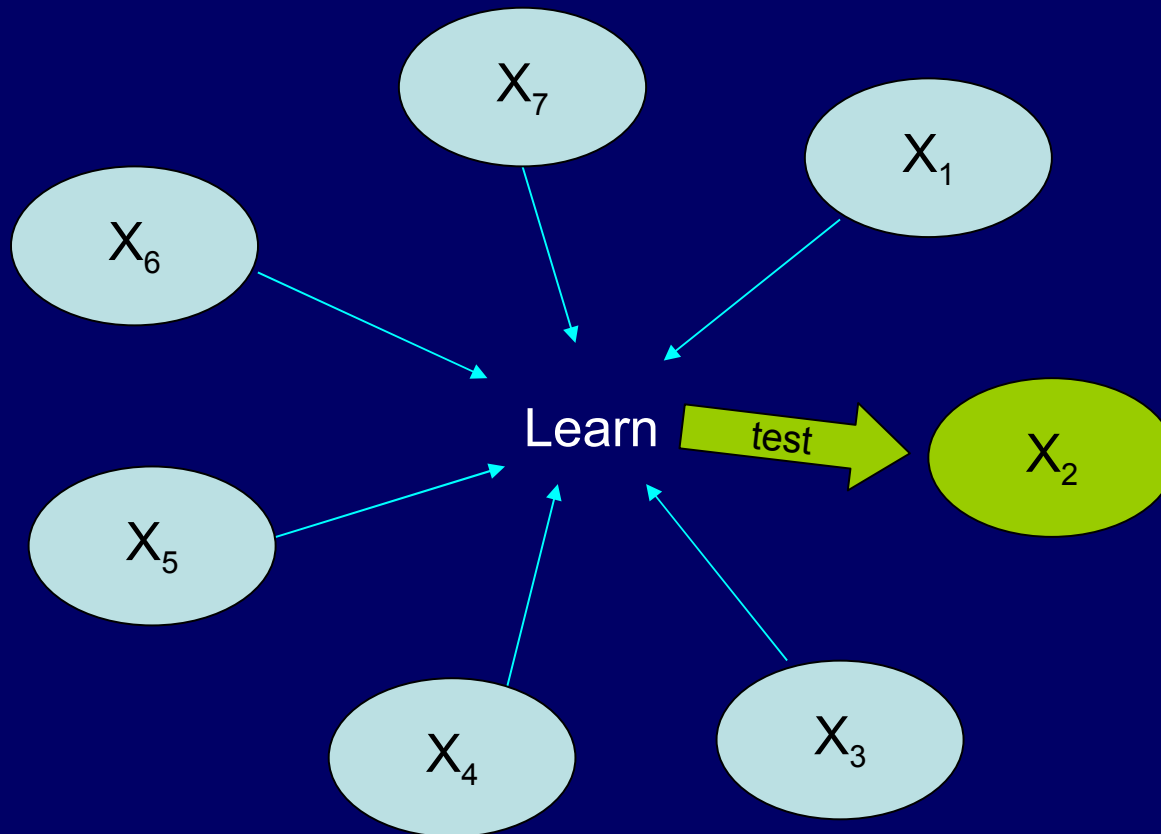
# K-fold Cross Validation - Illustration

- Example from last week lecture by Ben Blum.
- Divide the data to k groups  $X_1, X_2, X_3, \dots, X_k$
- Use each group as a validation set.

# K-fold Cross Validation - Illustration



# K-fold Cross Validation - Illustration



# ROC Plot

# Definitions

- Let  $f_\theta: \mathcal{R}^n \rightarrow \{0,1\}$  be a classifier function.

	Predicted positive	Predicted negative
Positive examples	True positives	False negatives
Negative examples	False positives	True negatives

# Example - Blood Pressure and Cardio Vascular Disease (CVD)

- Classifier: If a person has a mean blood pressure above  $t$ , he will have some CV event during 10 years. We have 100 samples.
- How do we choose  $t$  ?

$t = 0$

	Predicted positive	Predicted negative
Positive examples	70	0
Negative examples	30	0

**$t = 300$**

	Predicted positive	Predicted negative
Positive examples	0	70
Negative examples	0	30

**$t = 150$**

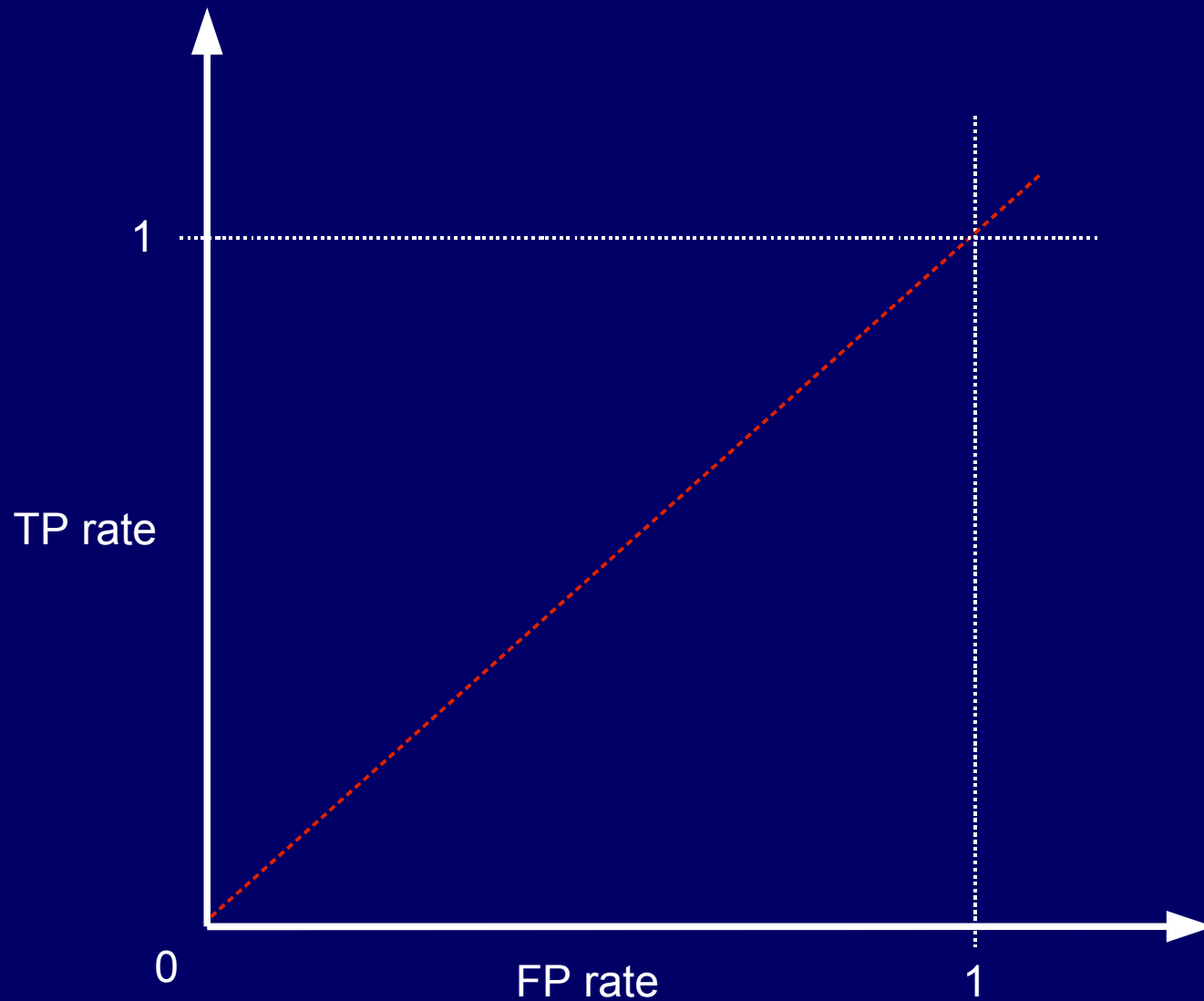
	Predicted positive	Predicted negative
Positive examples	30	40
Negative examples	10	20

# More Definitions

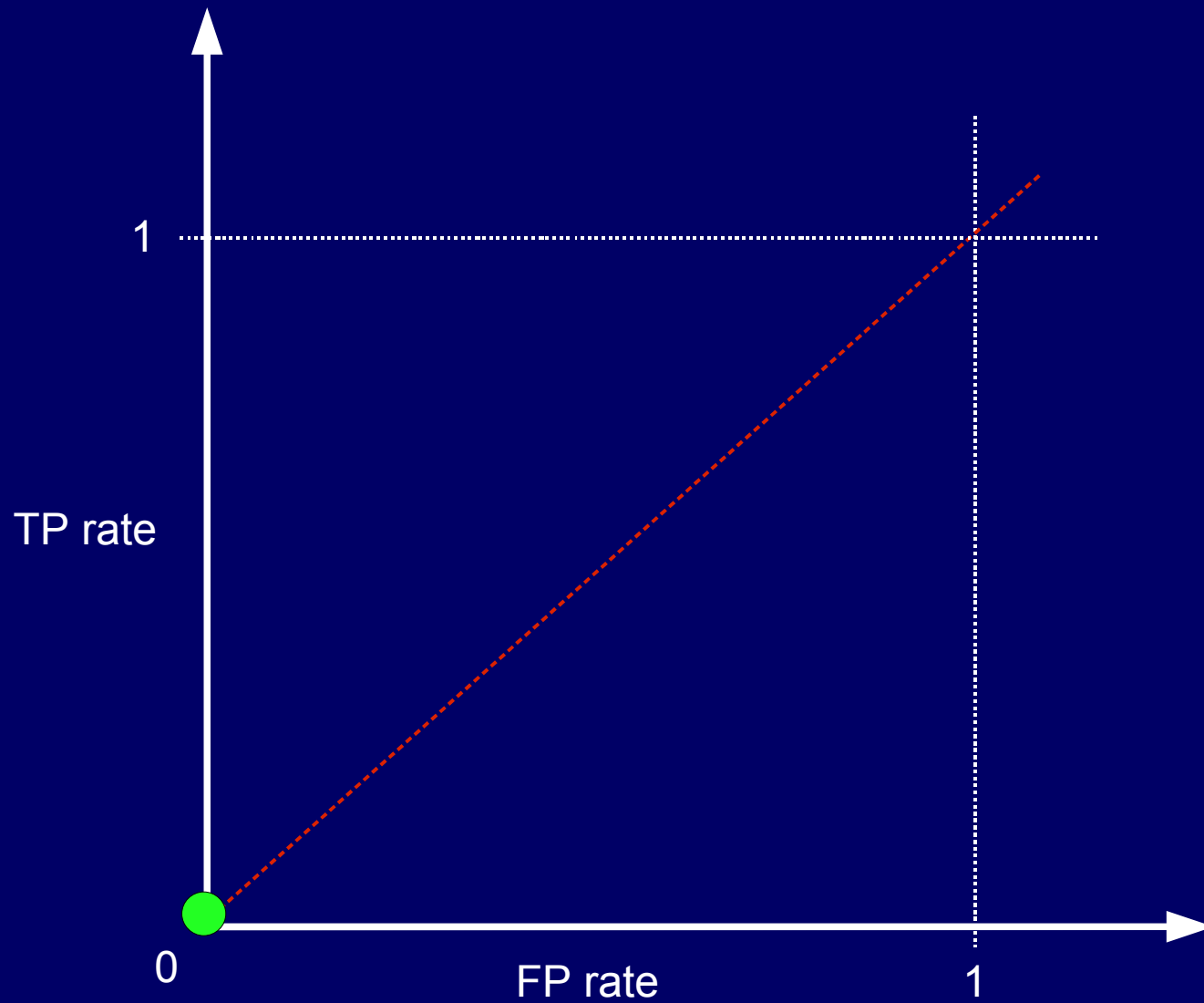
	Predicted positive	Predicted negative
Positive examples	TP	FN
Negative examples	FP	TN

- True positive rate =  $TP / (TP + FN)$
- False positive rate =  $FP / (FP + TN)$

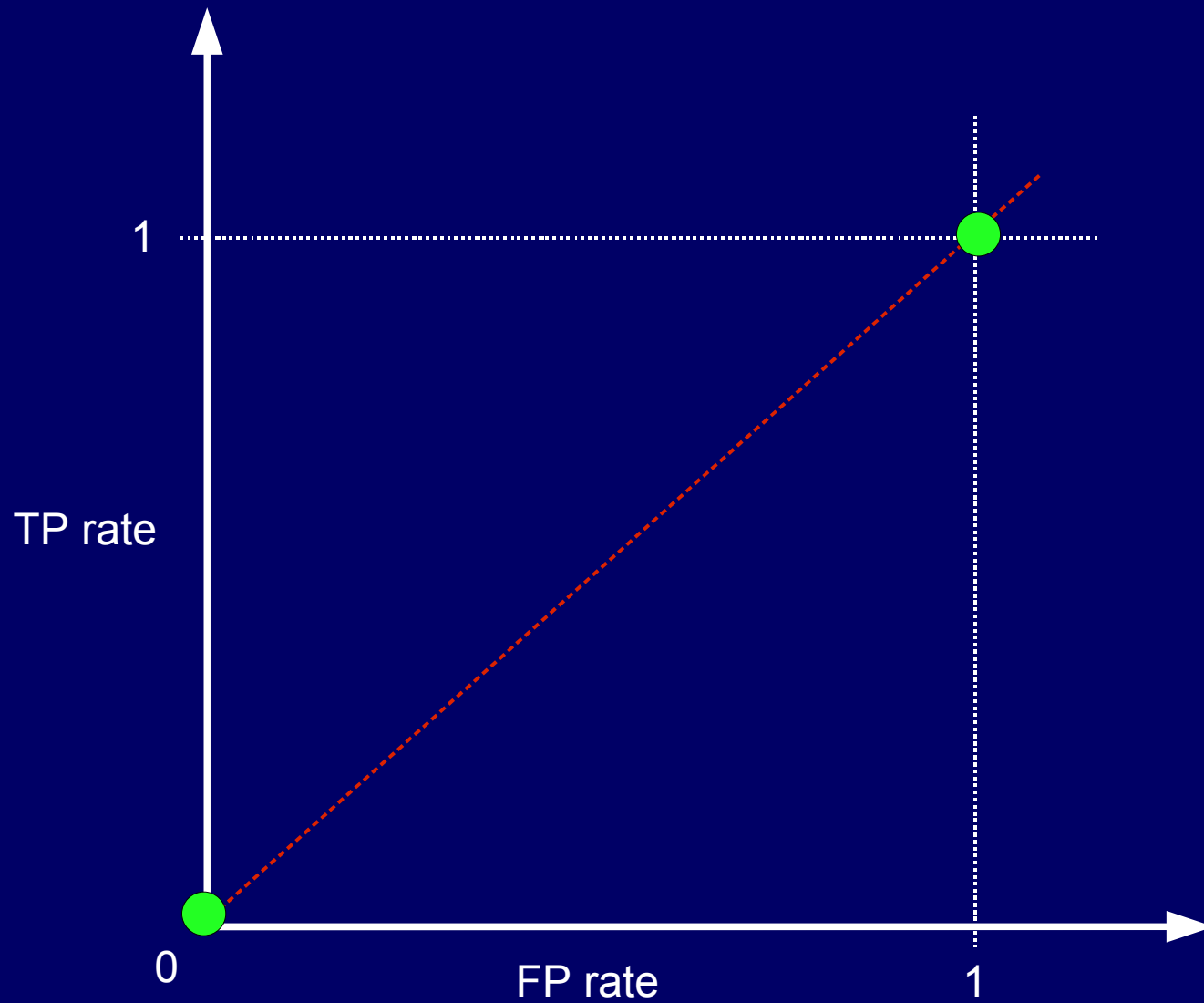
# ROC - Receiver Operating Characteristic Curve



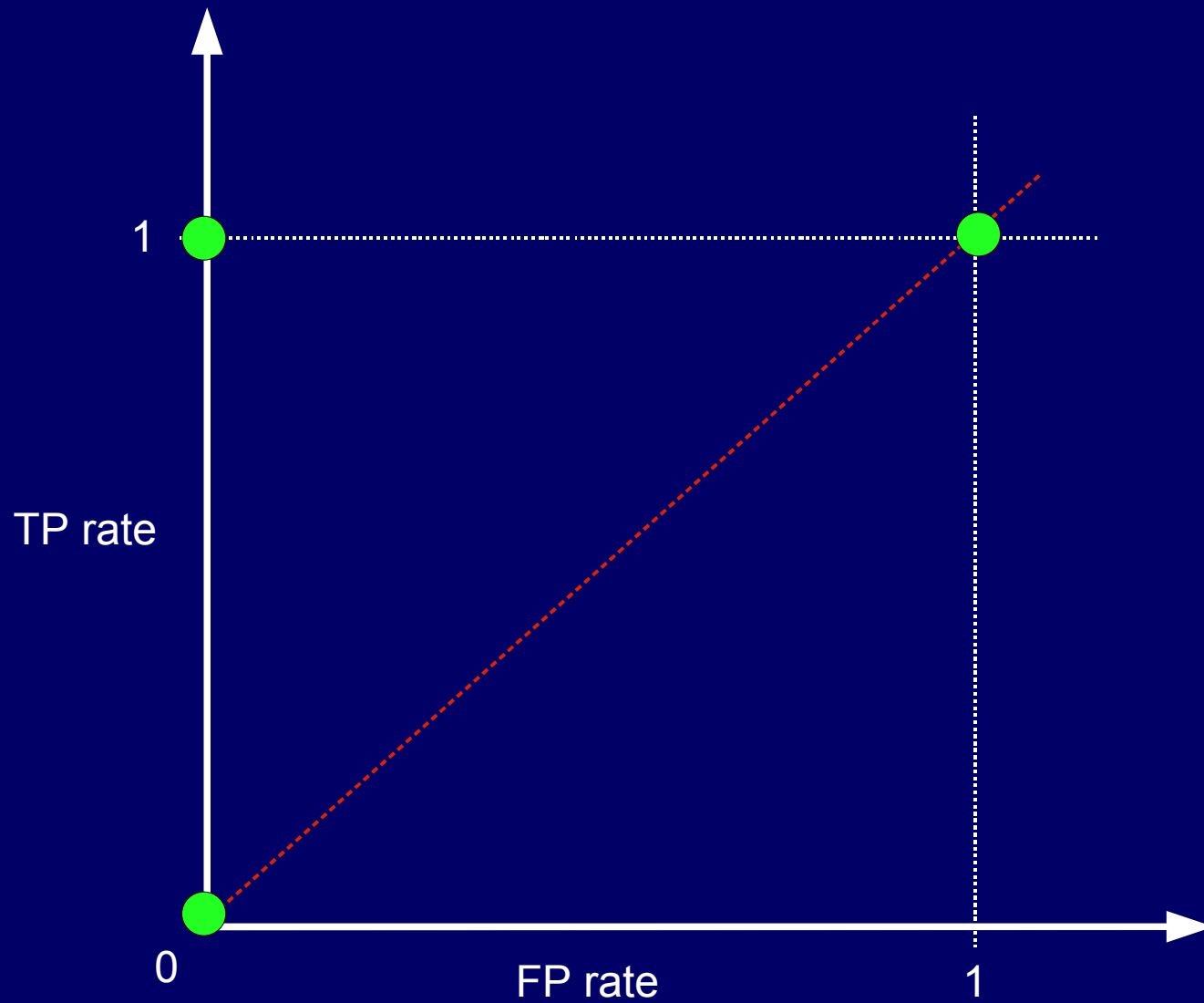
# ROC Curve



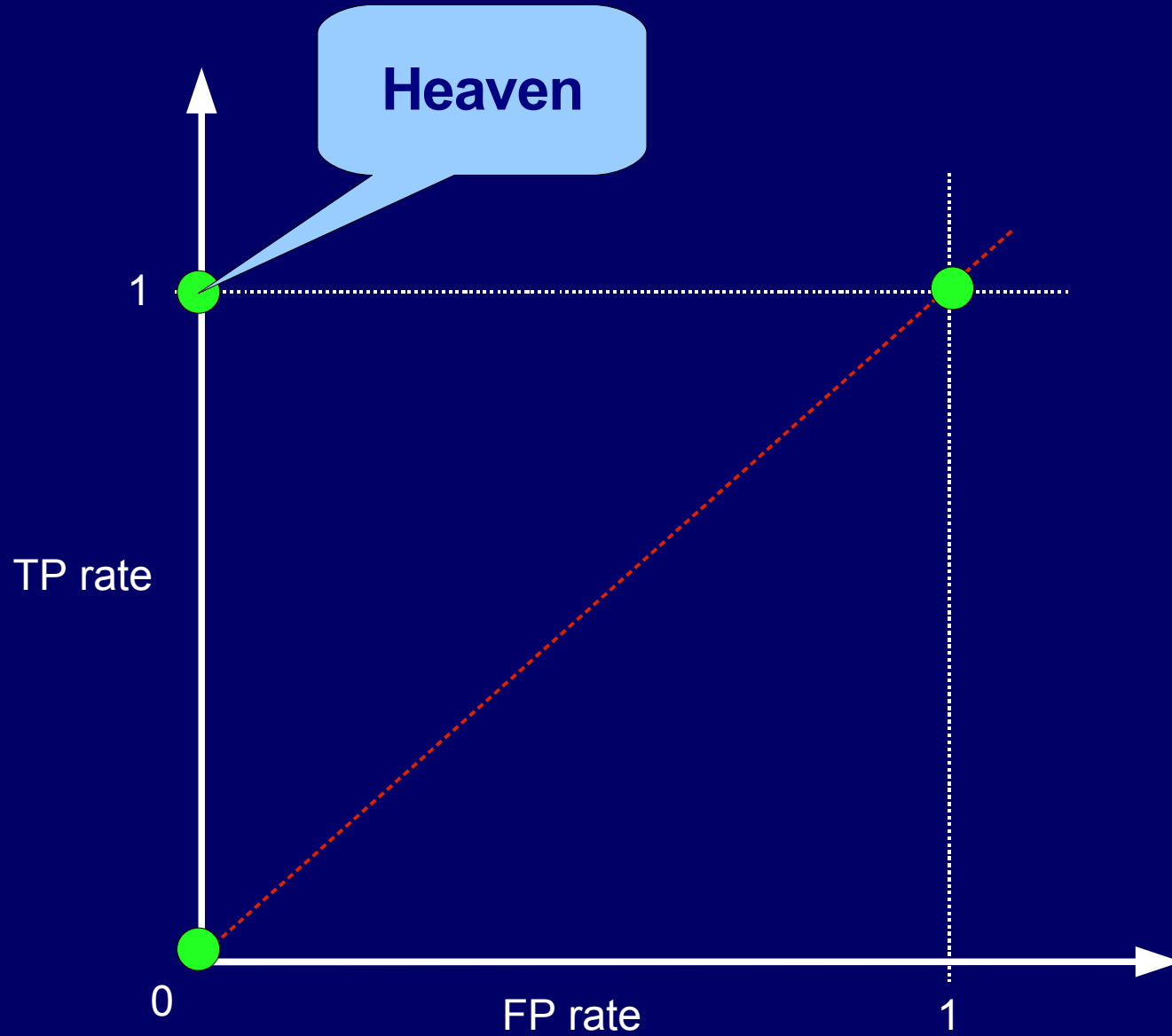
# ROC Curve



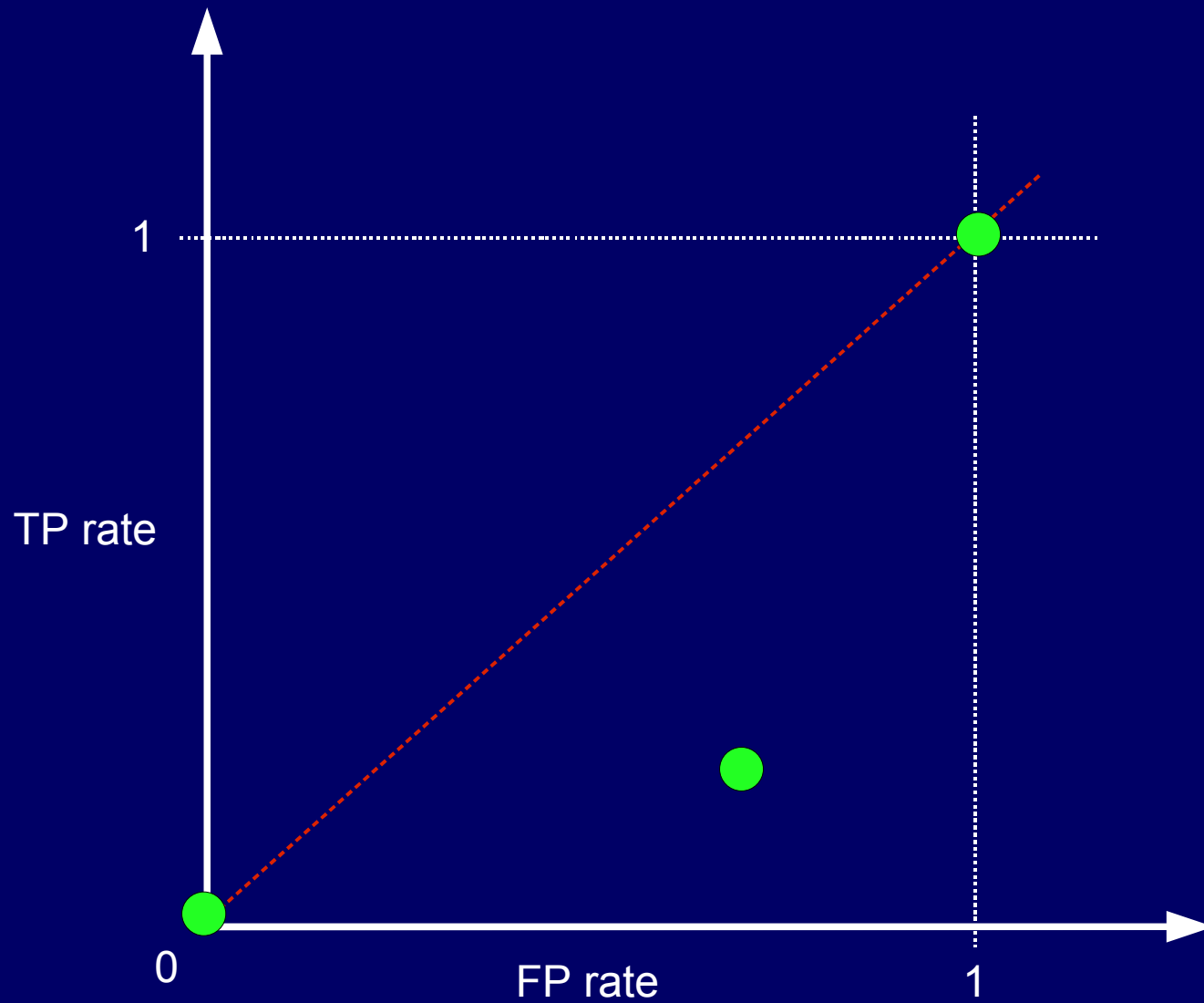
# ROC Curve



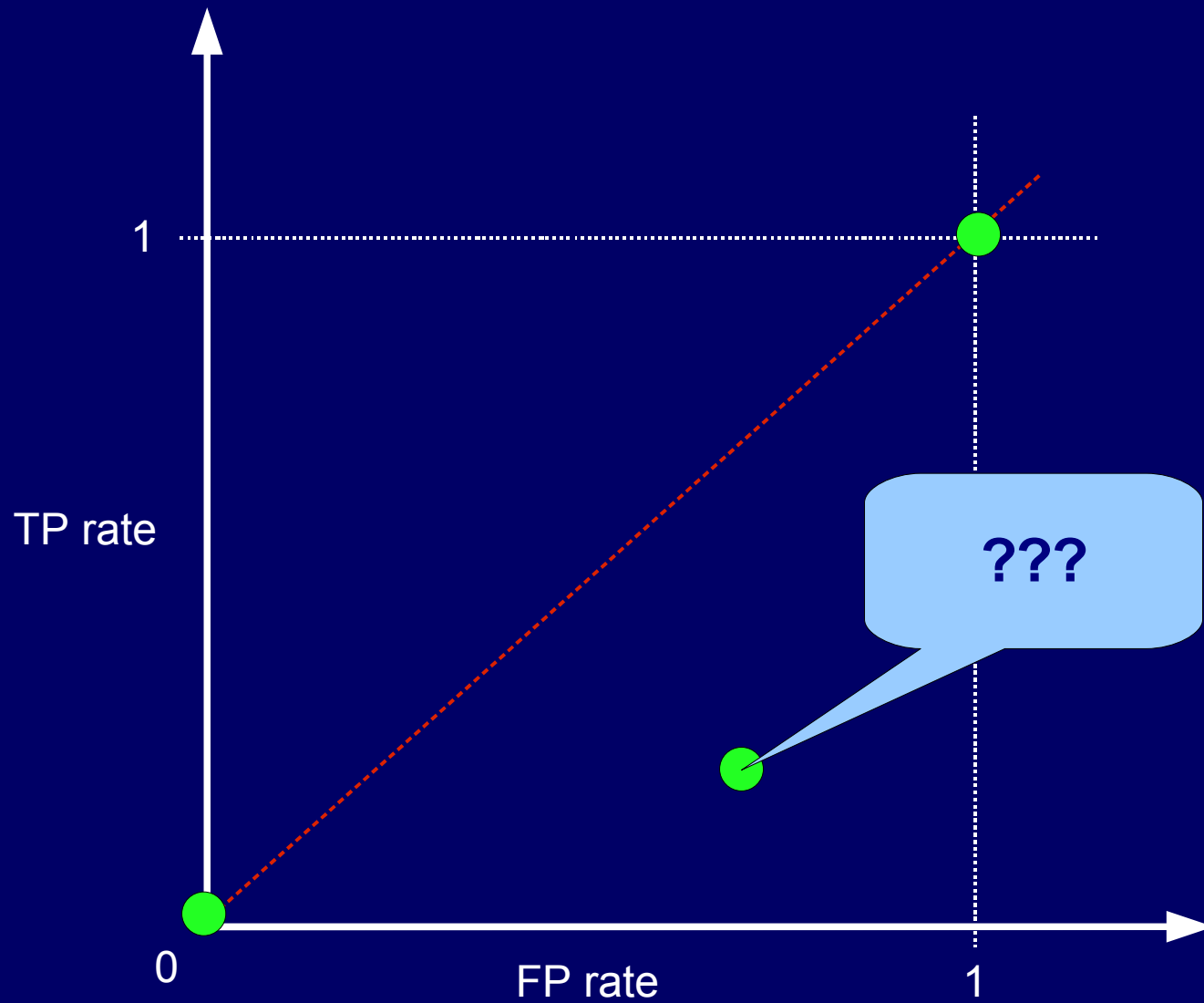
# ROC Curve



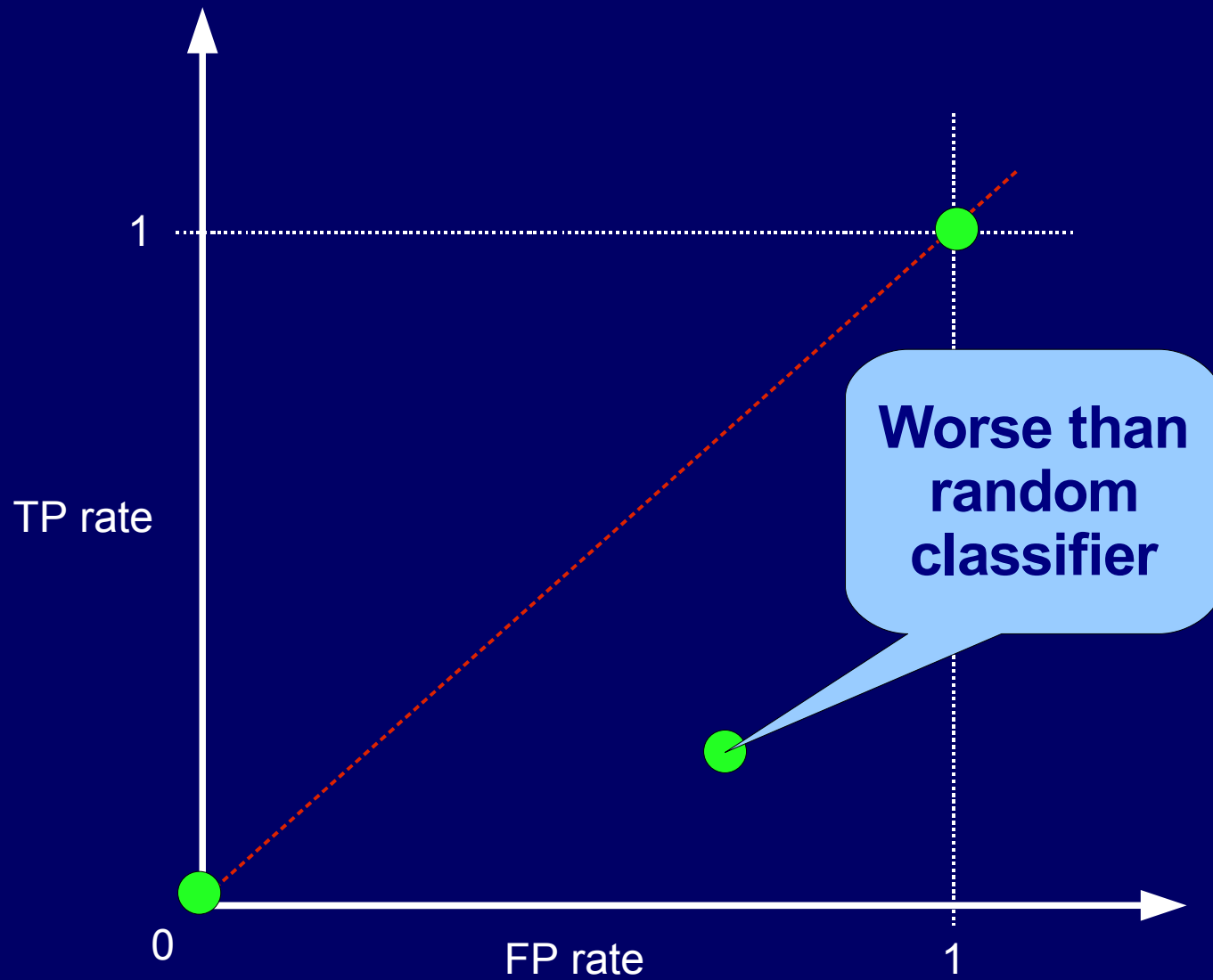
# ROC Curve



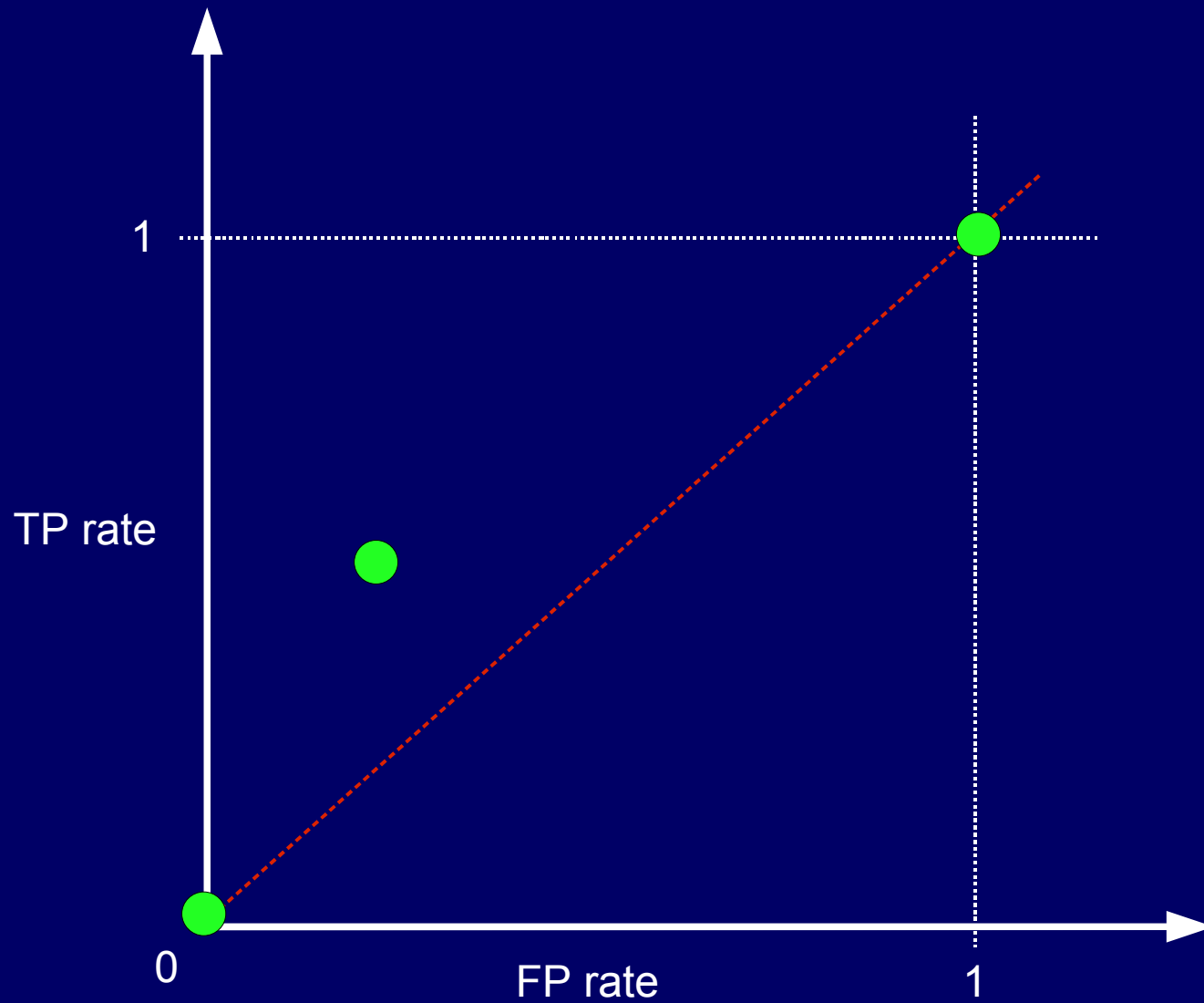
# ROC Curve



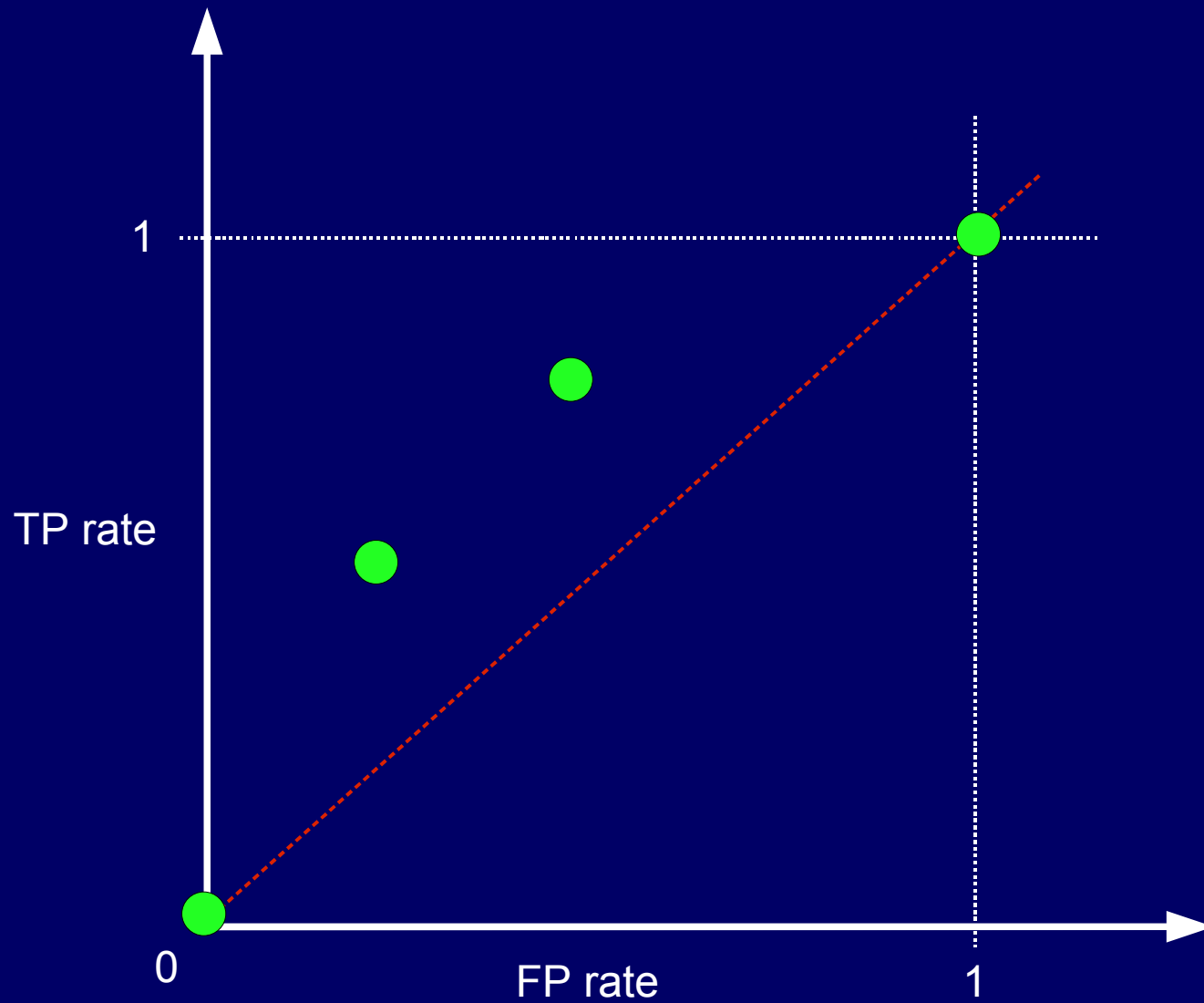
# ROC Curve



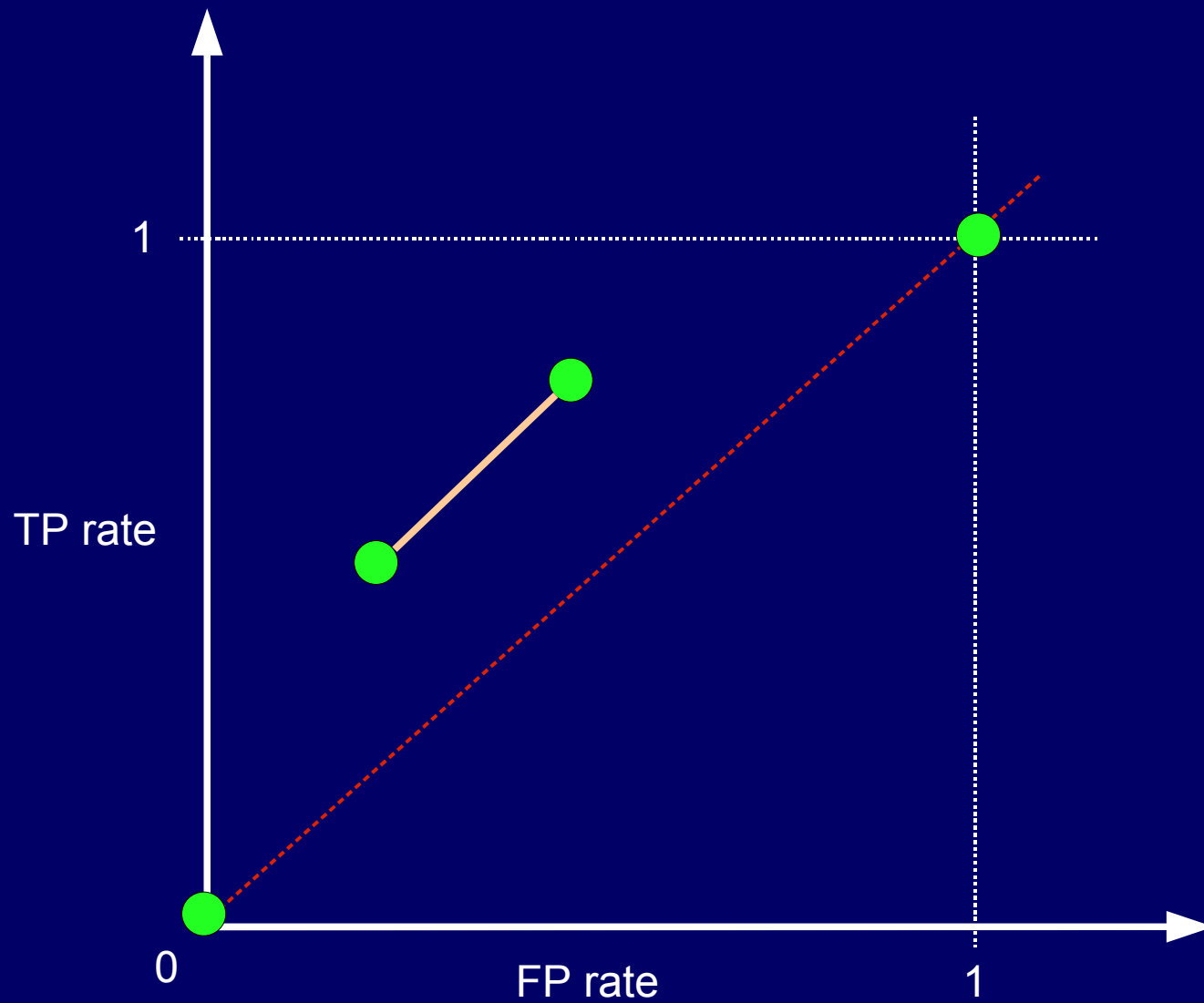
# ROC Curve



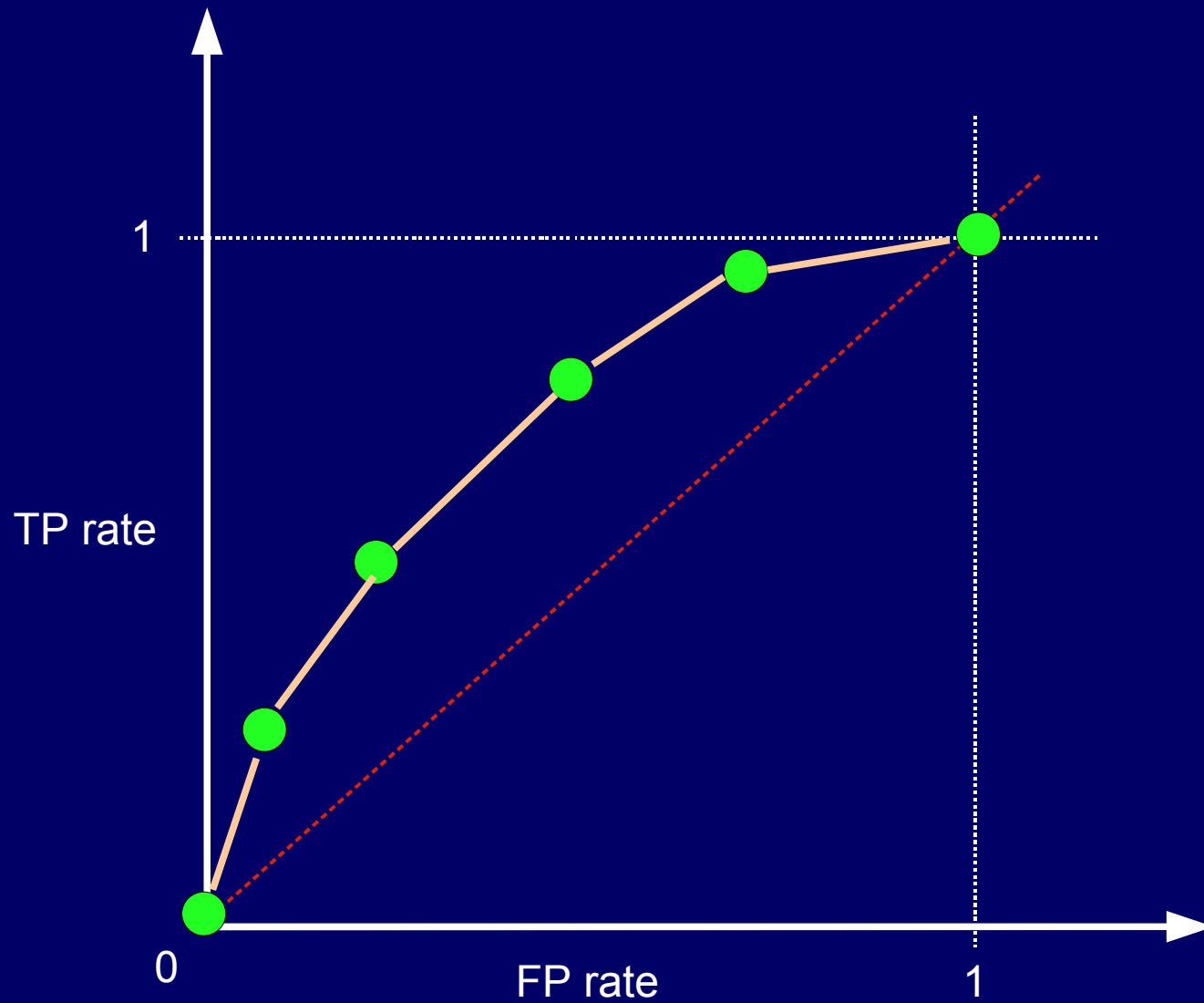
# ROC Curve



# ROC Curve



# ROC Curve



# The AUC Metric

- The Area Under the Curve (AUC) metric assesses the accuracy of the ranking in terms of separation of the classes.
- In random classifier (bad):  $AUC = 0.5$ .
- In perfect classifier (good):  $AUC = 1$ .