Diagnostics

Gad Kimmel

Outline

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

Introduction

Motivation

- Estimating properties of an estimator (an estimator is a function of input points).
  - Given data samples $x_1, x_2, \ldots, x_N$, evaluate some estimator, say the average:

$$\frac{\sum x_i}{N}$$

  - How can we estimate its properties (e.g., its variance)?

$$\mathrm{var}\left(\frac{\sum x_i}{N}\right) = \frac{1}{N^2}\mathrm{var}\left(\sum x_i\right)$$

- Model selection.
  - How many parameters should we use?

Bootstrap Method

Evaluating Accuracy

- A simple approach for accuracy estimation is to provide the bias or variance of the estimator.
- 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 μ and variance σ^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 $\bar{x}_n$)?

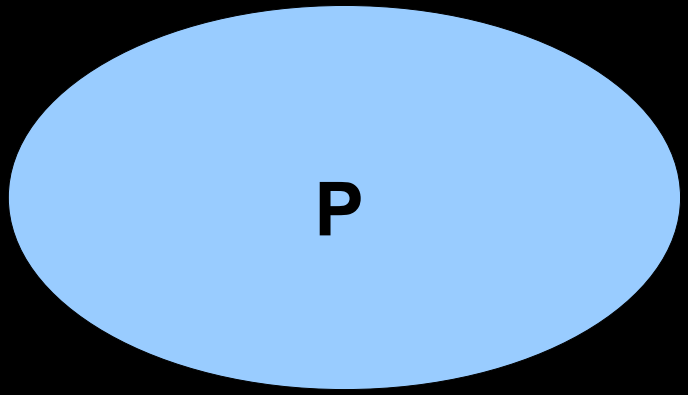
- It would be nice to have many different samples of samples.
- In that case, one could calculate the estimator for each sample of samples, and infer its distribution.
- But... we don't have it.

Solution - Bootstrap

- Estimating the sampling distribution of an estimator by resampling with replacement from the original sample.
- Efron, *The Annals of Statistics*, '79.

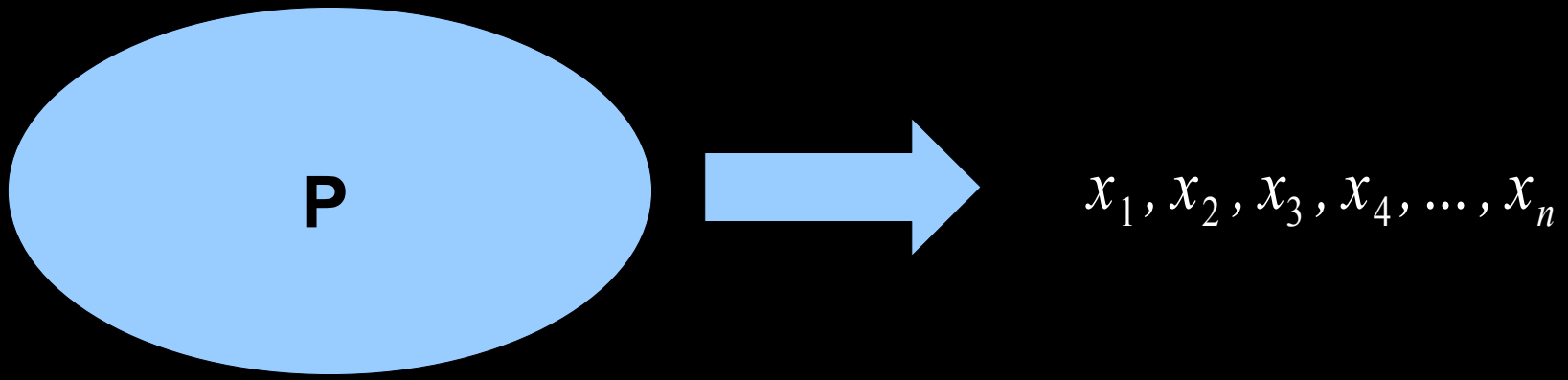
Bootstrap - Illustration

- Goal: Sampling from P.



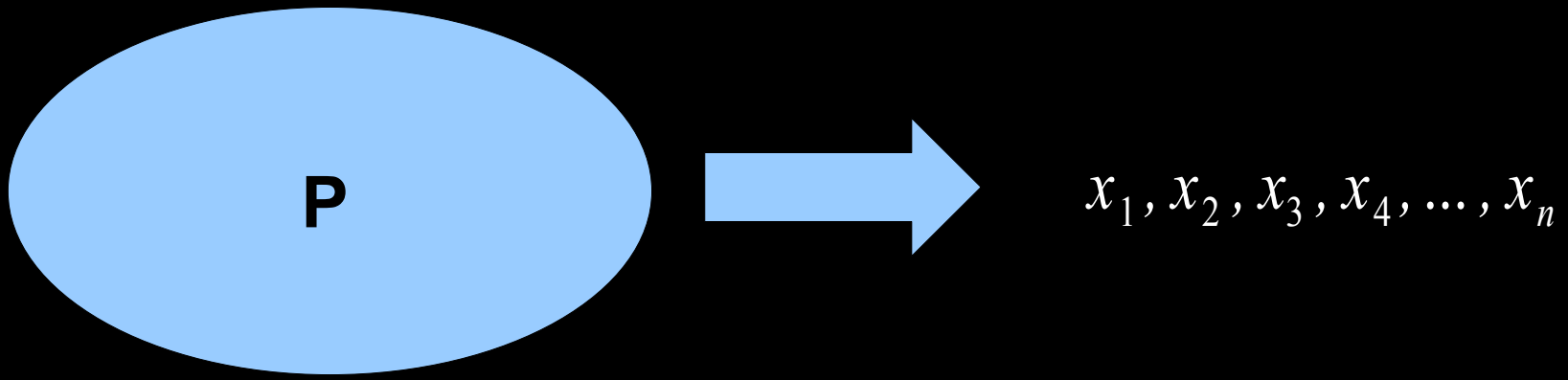
Bootstrap - Illustration

- Goal: Sampling from P.



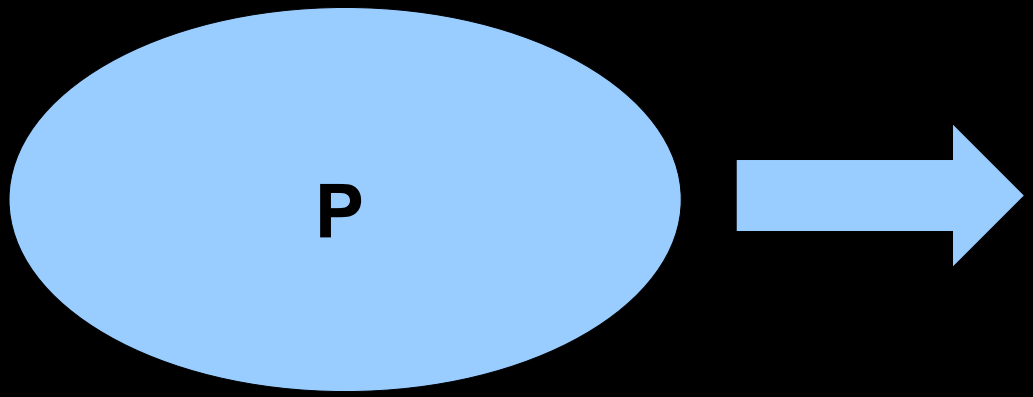
Bootstrap - Illustration

- Goal: Sampling from P.



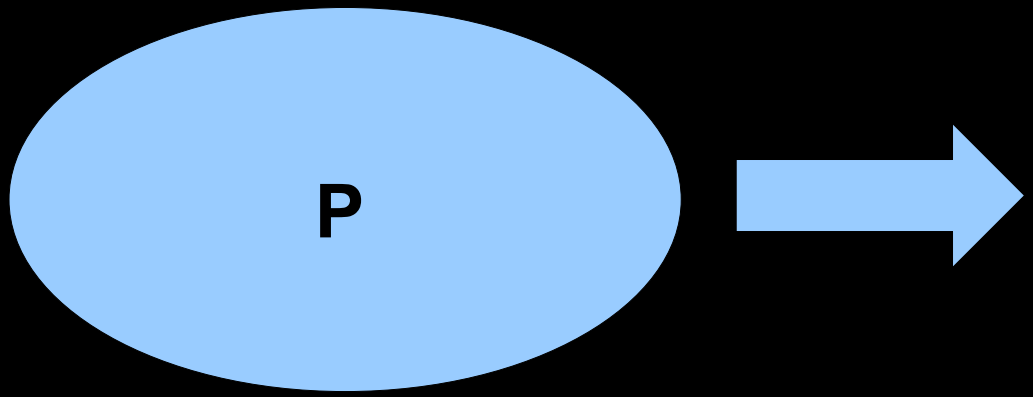
... in order to estimate the variance of an estimator.

Bootstrap - Illustration



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

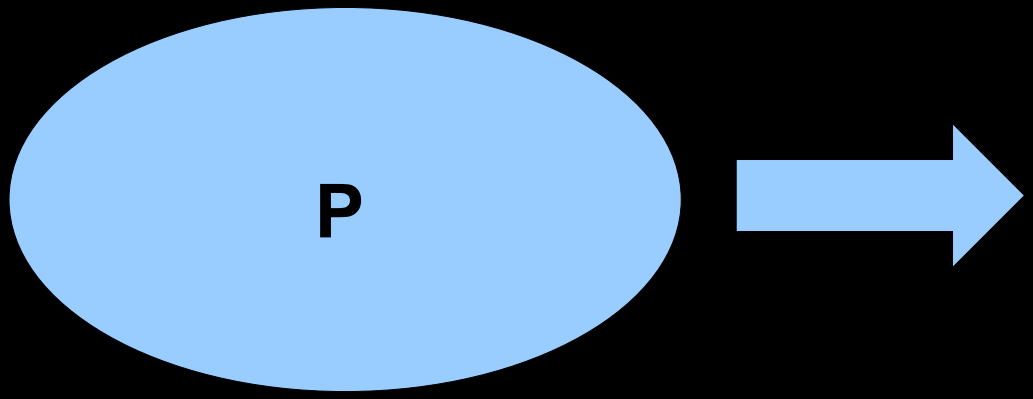
Bootstrap - Illustration



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

- What is the variance of e ?

Bootstrap - Illustration

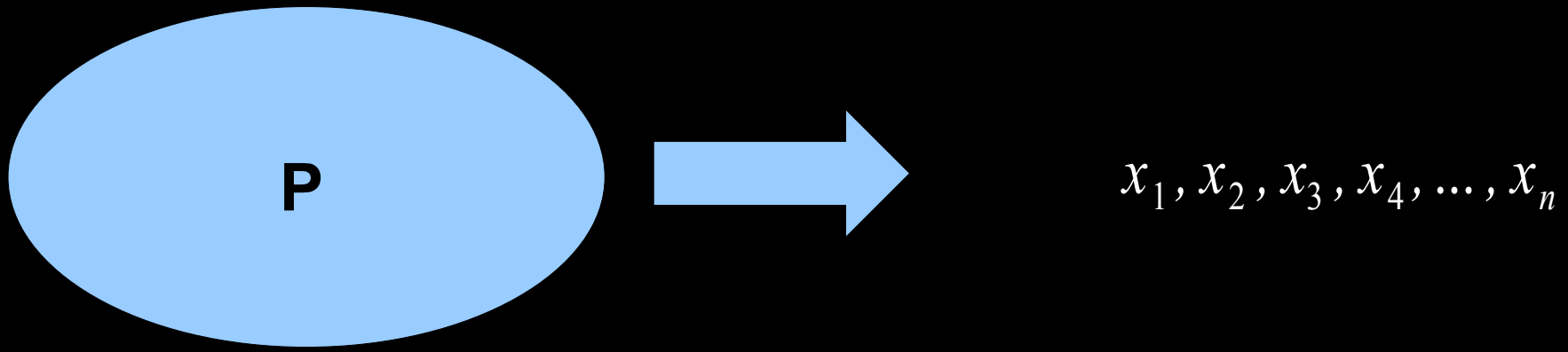


Samples	Estimator
$x_{1,1}, x_{1,2}, x_{1,3}, \ldots, x_{1,n}$	e_1
$x_{2,1}, x_{2,2}, x_{2,3}, \ldots, x_{2,n}$	e_2
$x_{3,1}, x_{3,2}, x_{3,3}, \ldots, x_{3,n}$	e_3
$x_{4,1}, x_{4,2}, x_{4,3}, \ldots, x_{4,n}$	e_4
...	
$x_{m,1}, x_{m,2}, x_{m,3}, \ldots, x_{m,n}$	e_m

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

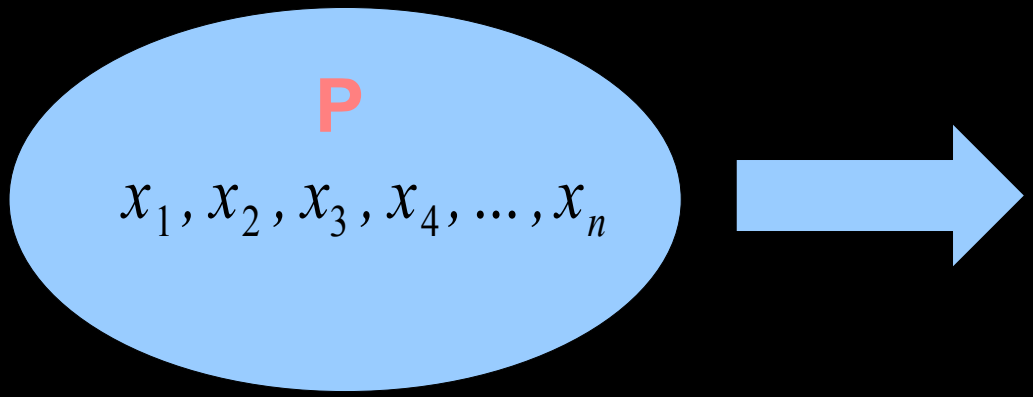
Bootstrap - Illustration

- We only have 1 sample:



Bootstrap - Illustration

- Sampling is done from the empirical distribution.



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

Formalization

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

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

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

The Main Idea

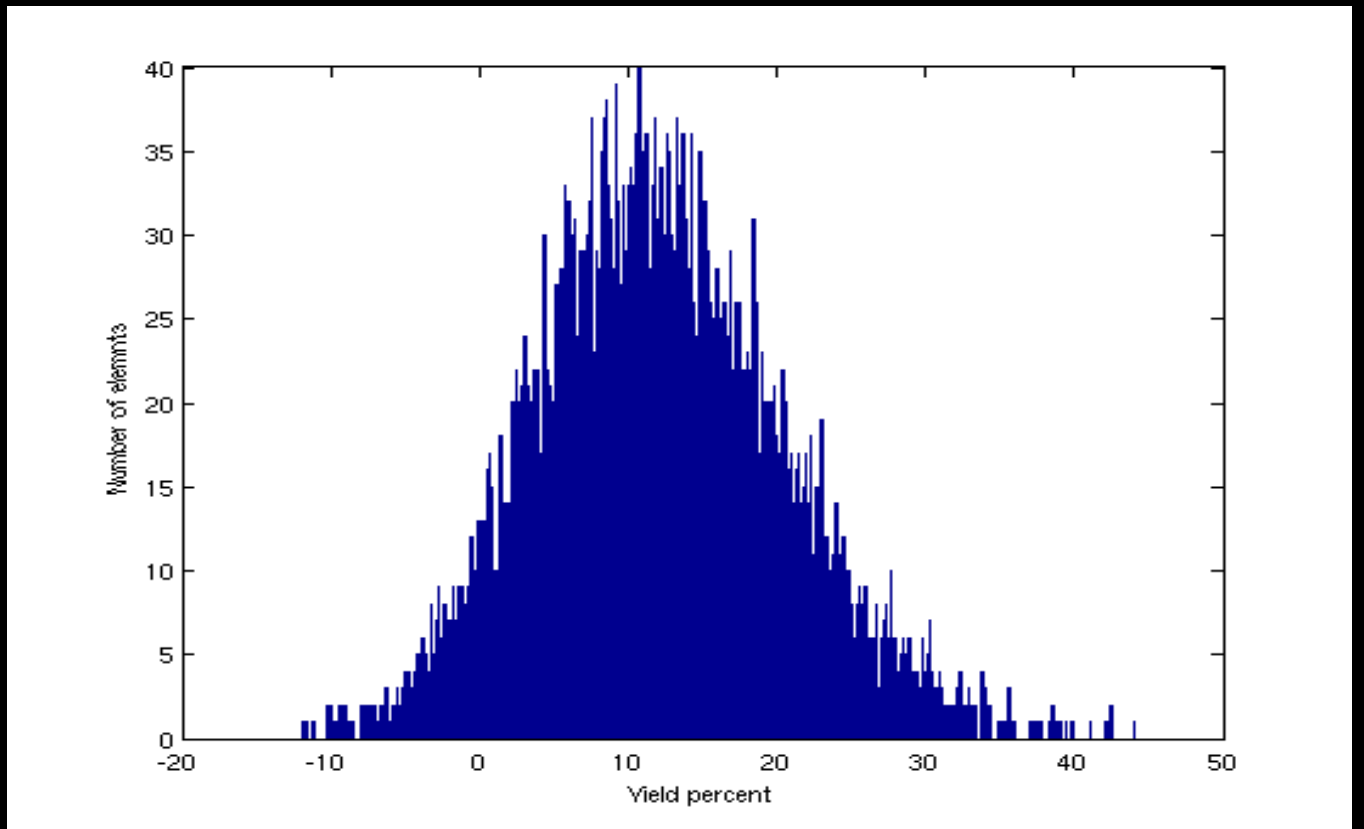
- $Y_i \sim \hat{P}$.
- We wish that $P = \hat{P}$. Is it (always) true? NO.
- Rather, $\hat{P}$ is an approximation of P .

Example 1

- The yield of the Dow Jones Index over the past two years is ~12%.
- You are considering a broker that had a yield of 25%, 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

- Prepare a sample $x_{1,}x_{2,}\cdots,x_{10,000}$, where each x_i is the yield of randomly selected stocks.
- Approximate the distribution of x using this sample.

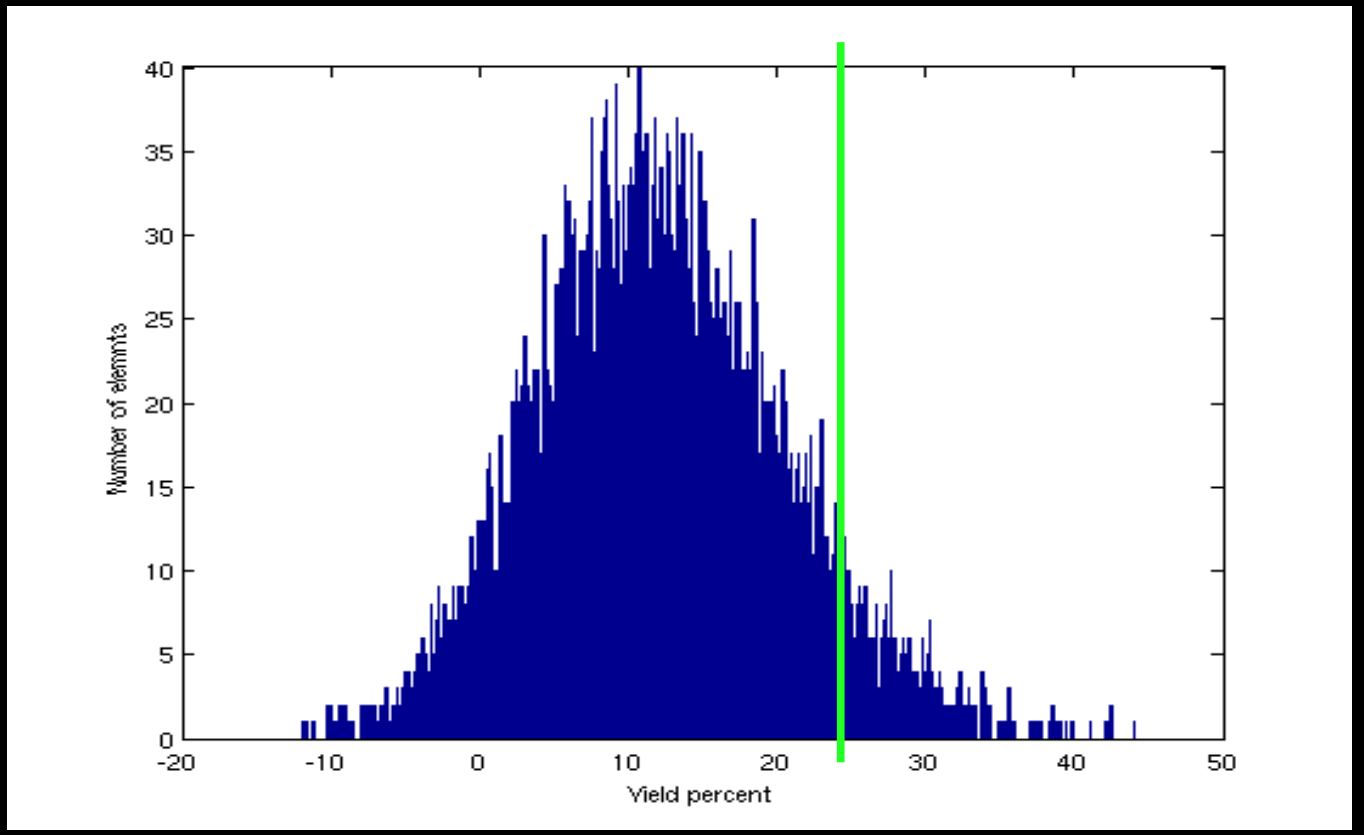


Evaluation of Estimators

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

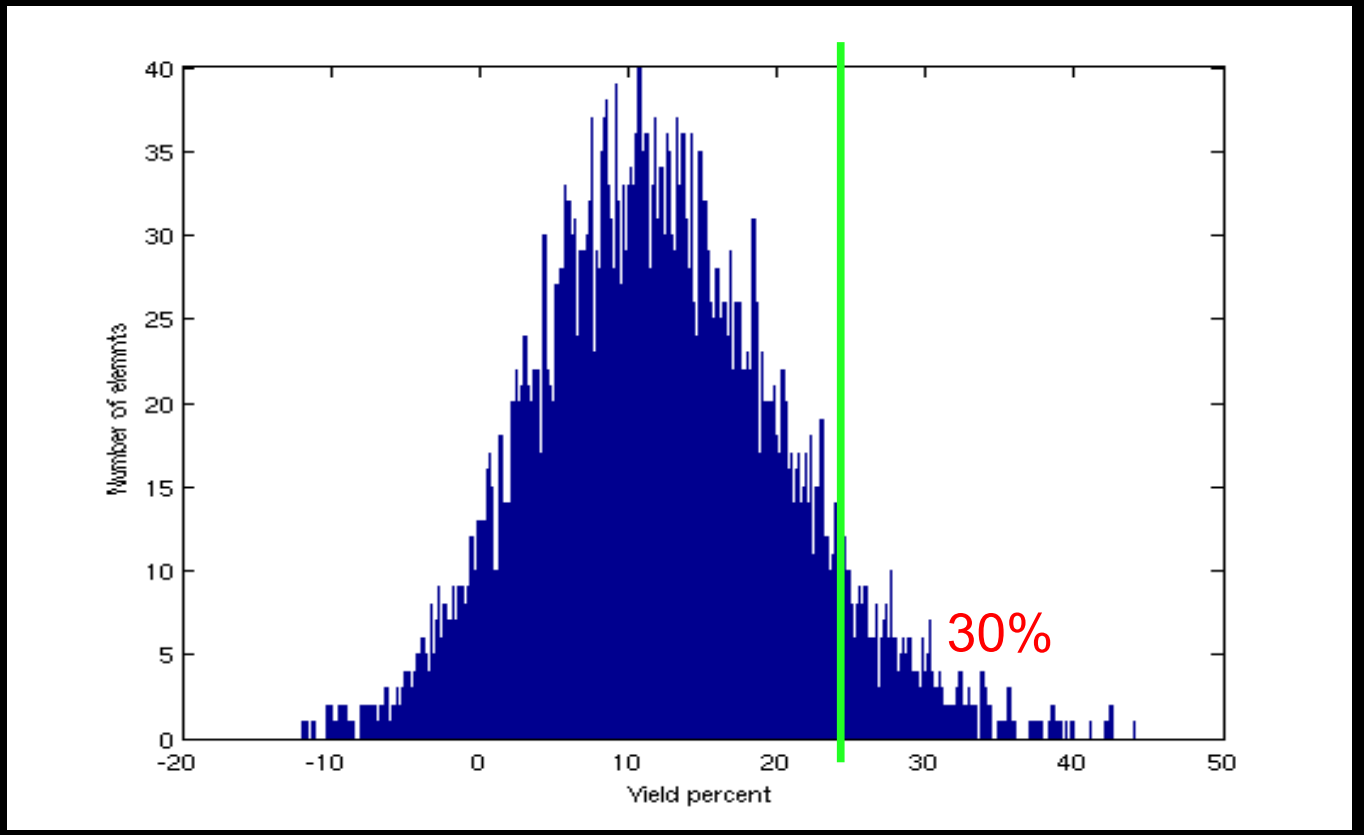
Example 1

- What is the probability to obtain yield larger than 25% (p-value)?



Example 1

- What is the probability to obtain yield larger than 25% (p-value)?



Example 2 - Decision tree

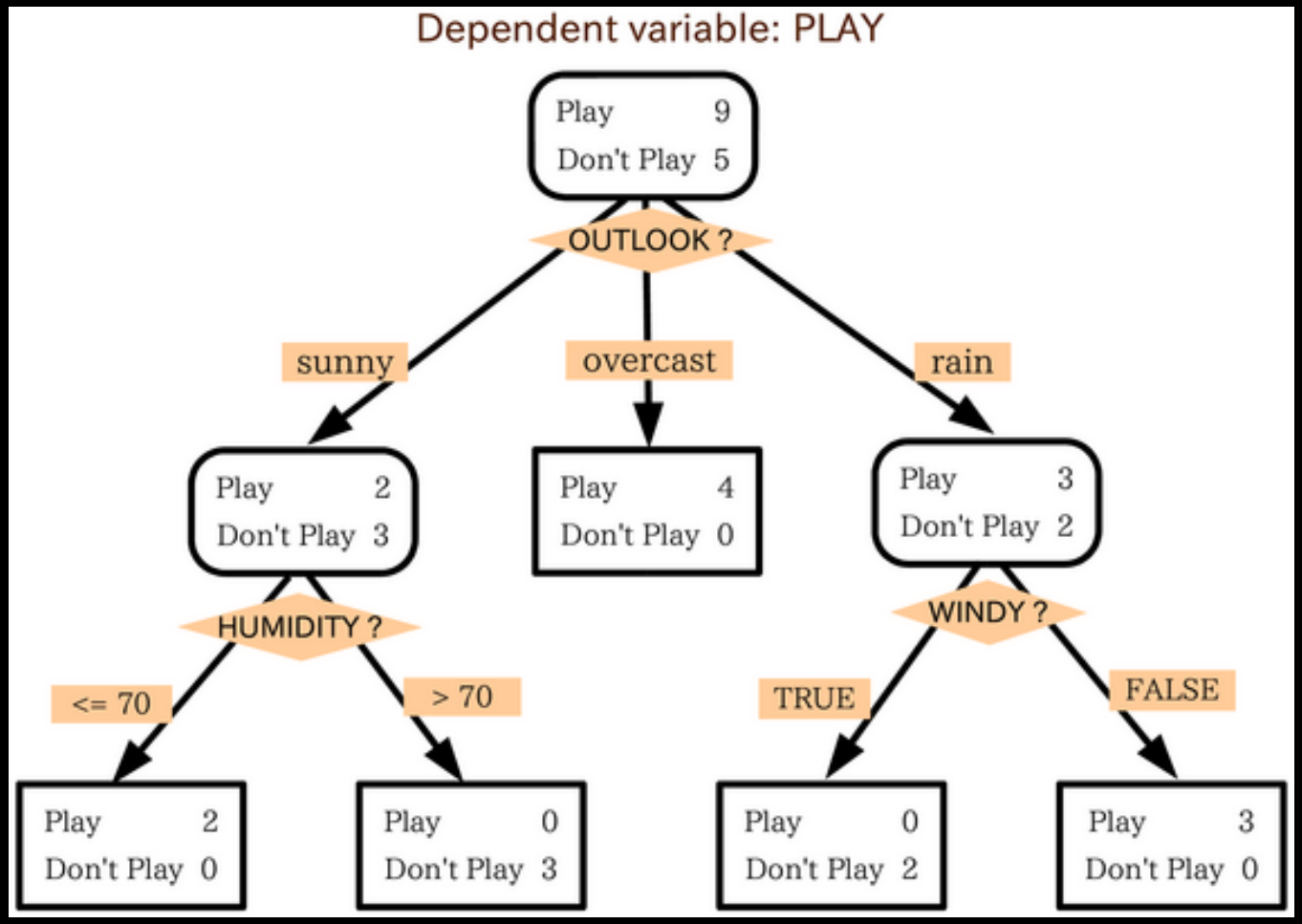
- Decision tree - short introduction.

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:

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

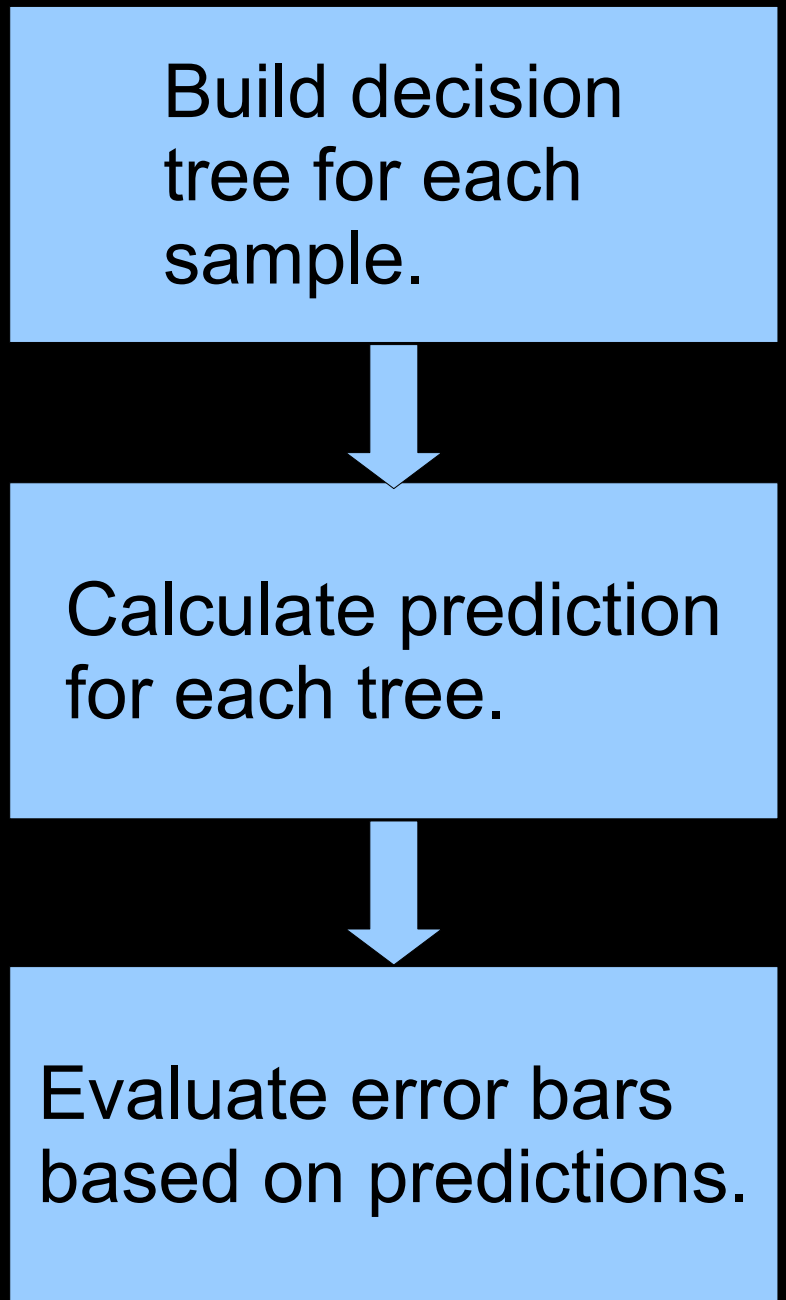
Example 2

- 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



Example 2 - Applying Bootstrap

$T_1, T_2, \ldots, T_n$

Build decision tree for each sample.

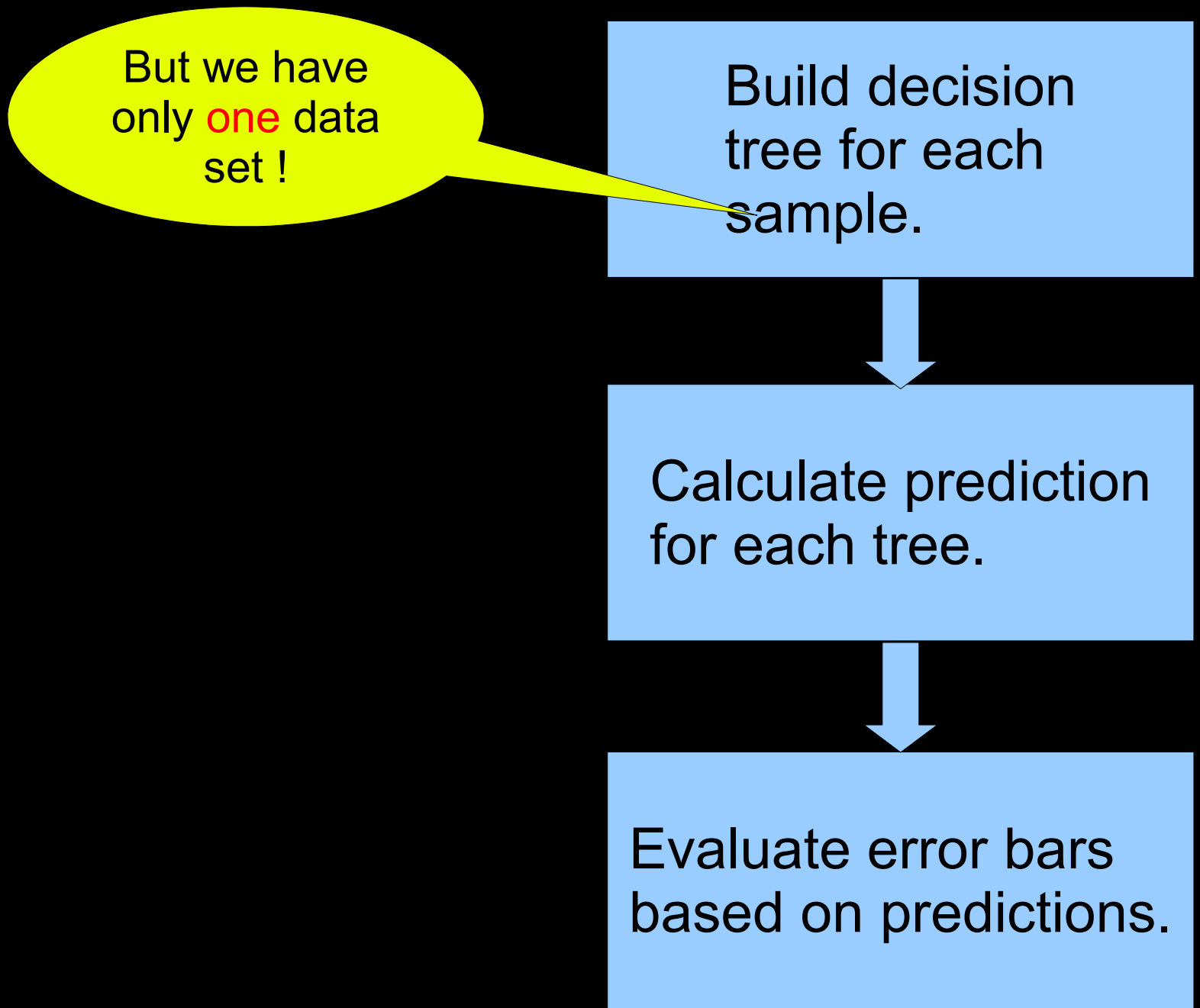
$p_1, p_2, \ldots, p_n$

Calculate prediction for each tree.

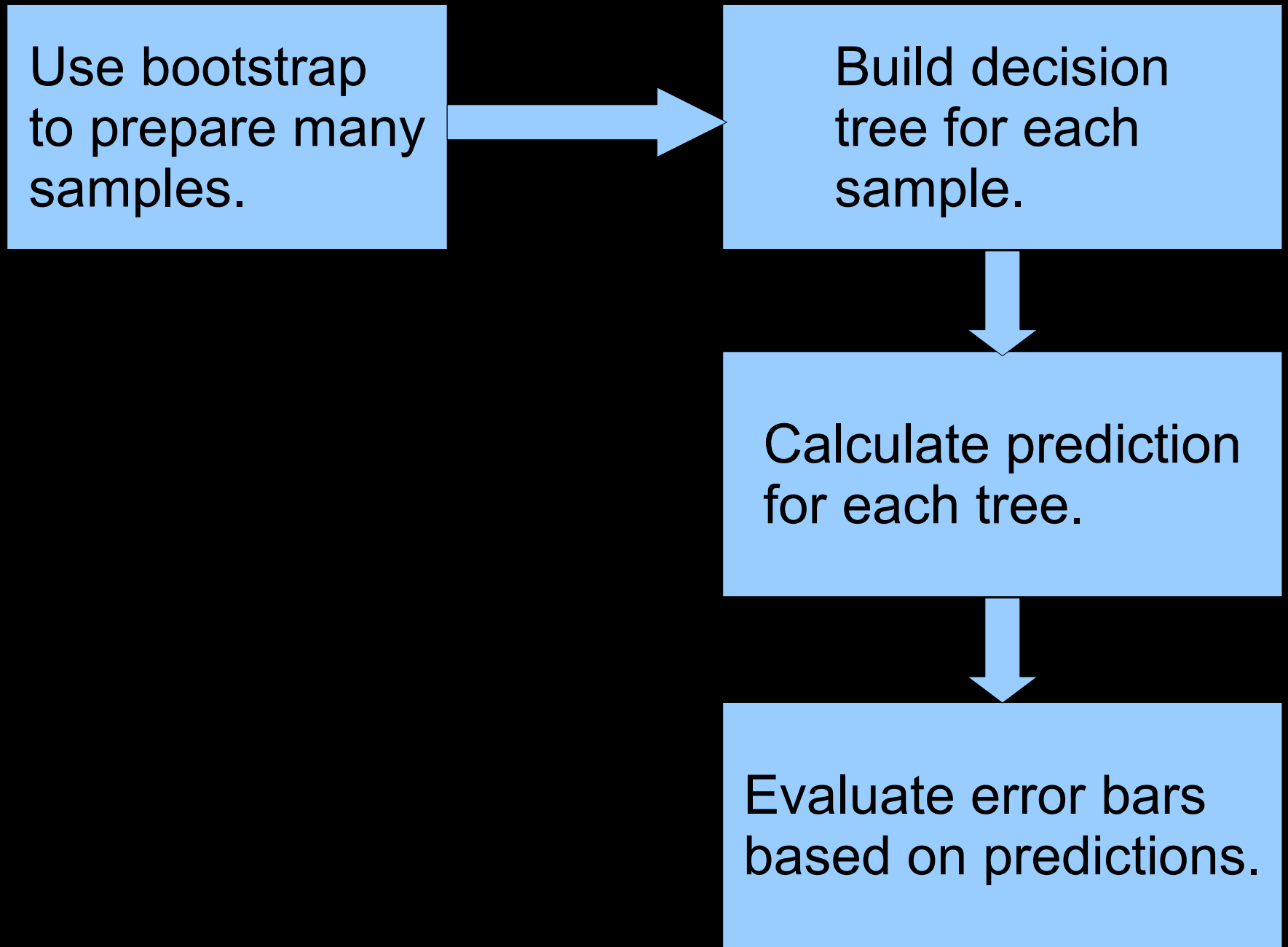
$\mu \pm 1.96\, STD(p_1, p_2, \ldots, p_n)$

Evaluate error bars based on predictions.

Example 2 - Applying Bootstrap



Example 2 - Applying Bootstrap



Cross Validation

Objective

- Model selection.

Formalization

- Let (x, y) drawn from distribution *P*. Where $x \in \Re^n$ and $y \in \Re$
- Let $f_\theta : \Re^n \to \Re$ be a learning algorithm, with parameter(s) θ.

Example

- Regression model.

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.

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*.
- Niels Bohr:

> "*It is very difficult to make an accurate prediction, especially about the future.*"

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:

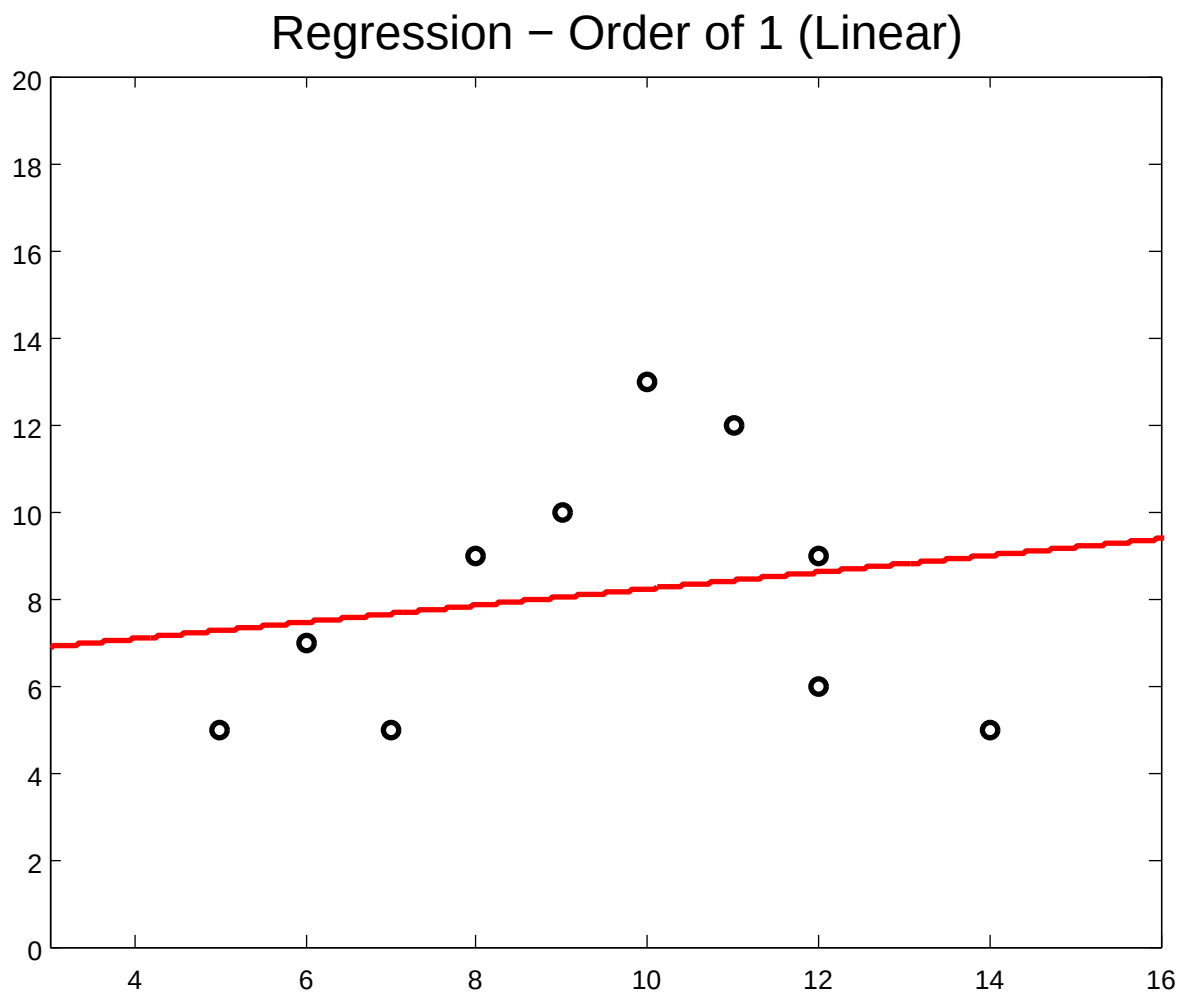
$$\mathrm{E}[(f_\theta(x) - y)^2]$$

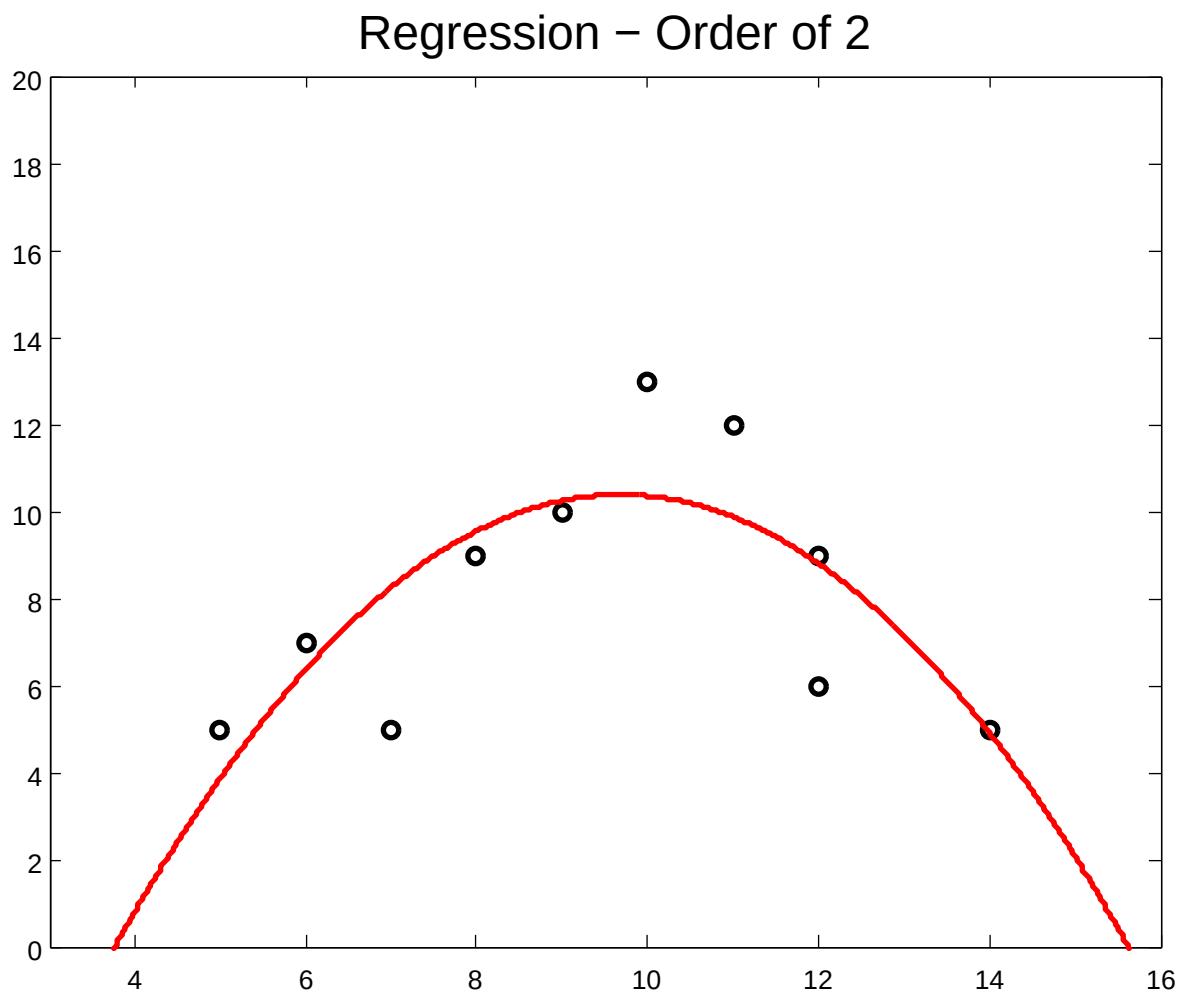
Choosing the Best Model (cont.)

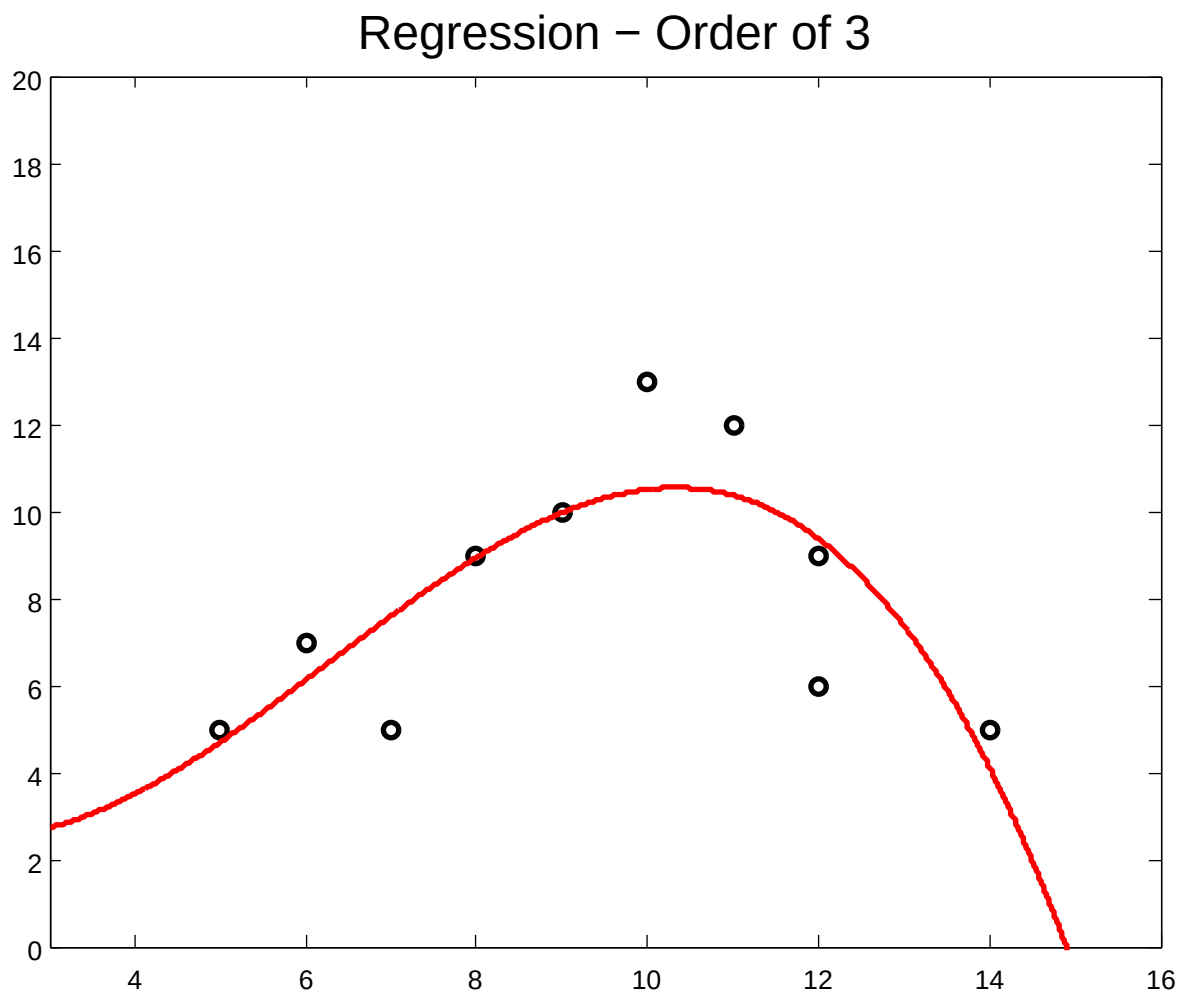
- Choose the parameter(s) θ :

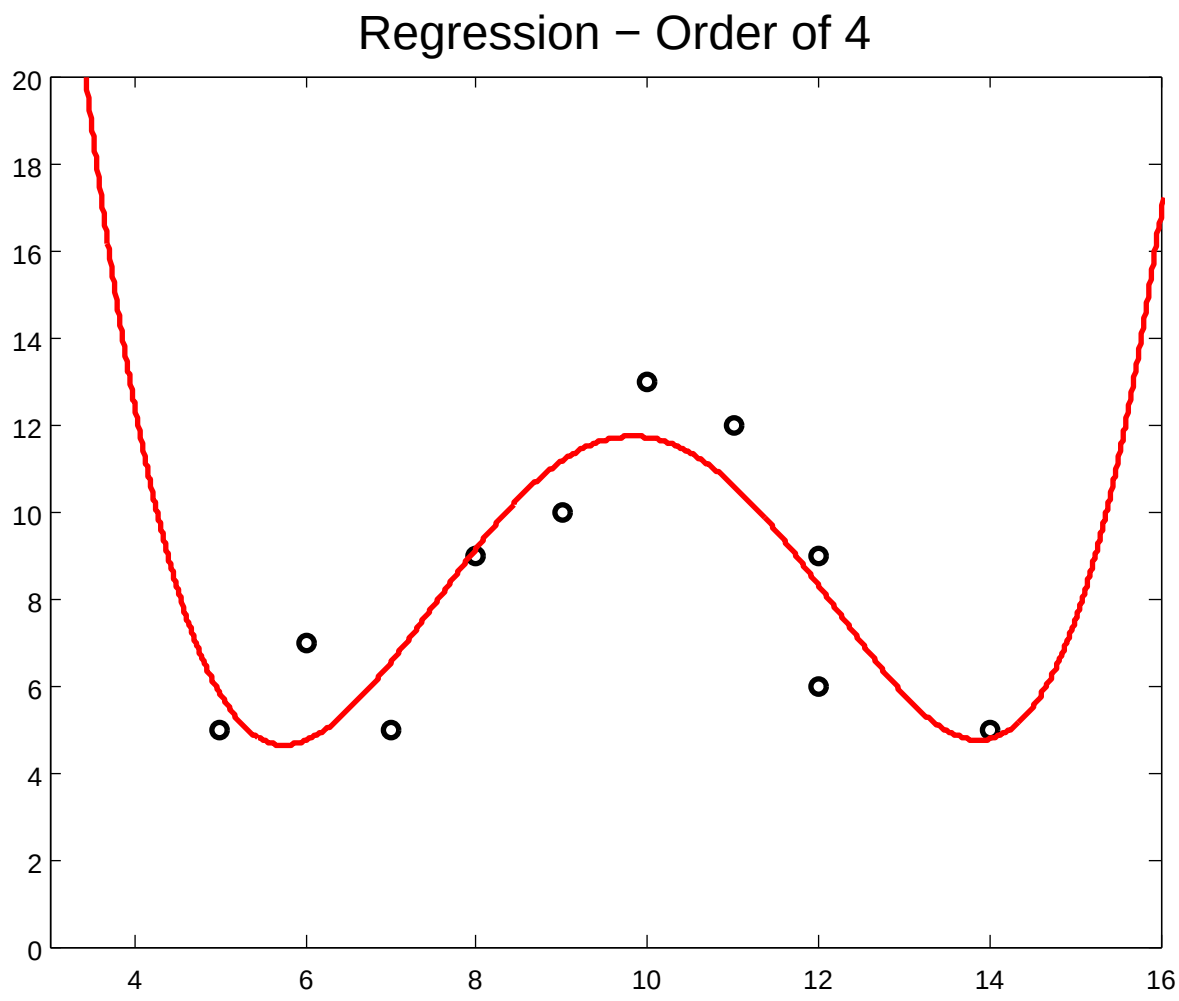
$$\operatorname{argmin}_\theta \mathrm{E}[(f_\theta(x) - y)^2]$$

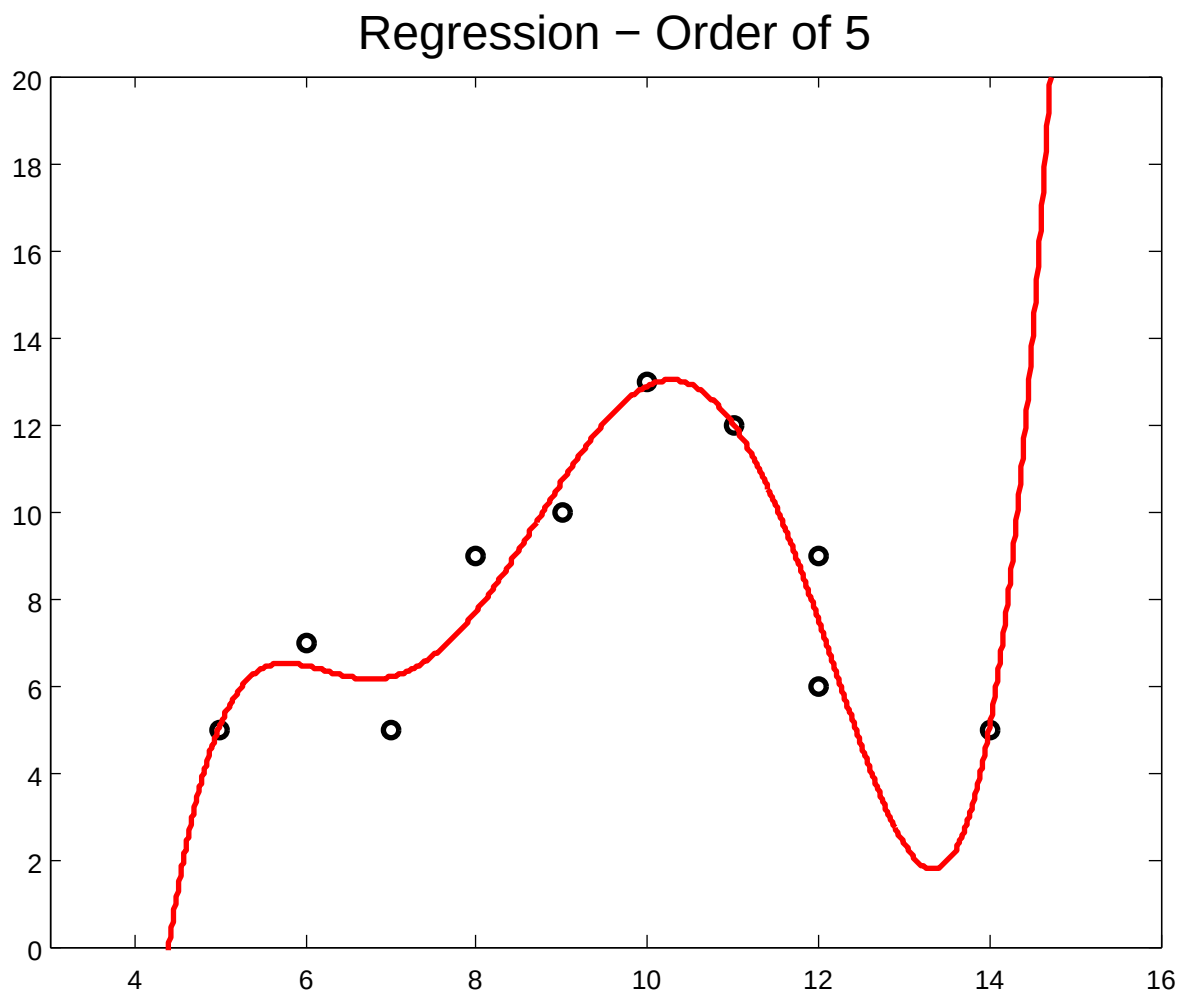
- The problem is that we don't know to sample from P.

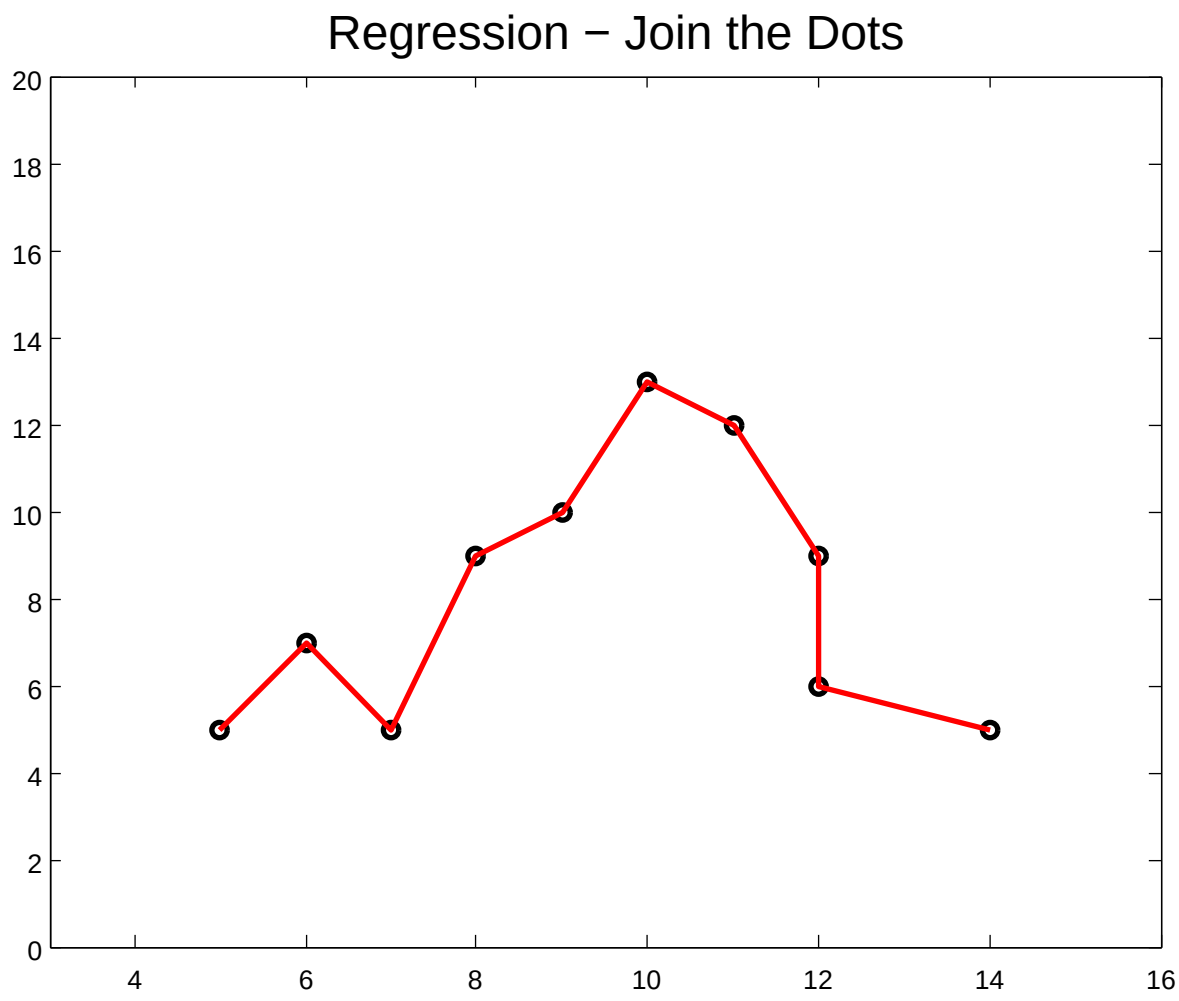

Regression − Order of 1 (Linear)
20
18
16
14
12
10
8
6
4
2
0
4
6
8
10
12
14
16


Regression − Order of 2
20
18
16
14
12
10
8
6
4
2
0
4
6
8
10
12
14
16


Regression − Order of 3
20
18
16
14
12
10
8
6
4
2
0
4
6
8
10
12
14
16


Regression − Order of 4
20
18
16
14
12
10
8
6
4
2
0
4
6
8
10
12
14
16


Regression − Order of 5
20
18
16
14
12
10
8
6
4
2
0
4
6
8
10
12
14
16


Regression − Join the Dots
20
18
16
14
12
10
8
6
4
2
0
4
6
8
10
12
14
16

Solution - Cross Validation

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

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

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 set, 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

K-fold Cross Validation

- We want to find a parameter that minimizes the cross validation estimate of prediction error:

$$\mathrm{CV}(\alpha) = \frac{1}{|N|} \sum L(y_i, f^{-k(i)}(x_i, \alpha))$$

K-fold Cross Validation

- How to choose K?
- K=N (= leave one out) - CV is unbiased for true prediction error, but can have high variance.
- When K increases - CV has lower variance, but bias could be a problem (depending on how the performance of the learning method varies with size of training set).

ROC Plot

(Receiver Operating Characteristic)

Definitions

- Let $f_\theta : \Re^n \rightarrow \{-1, 1\}$ 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

$t = 150$

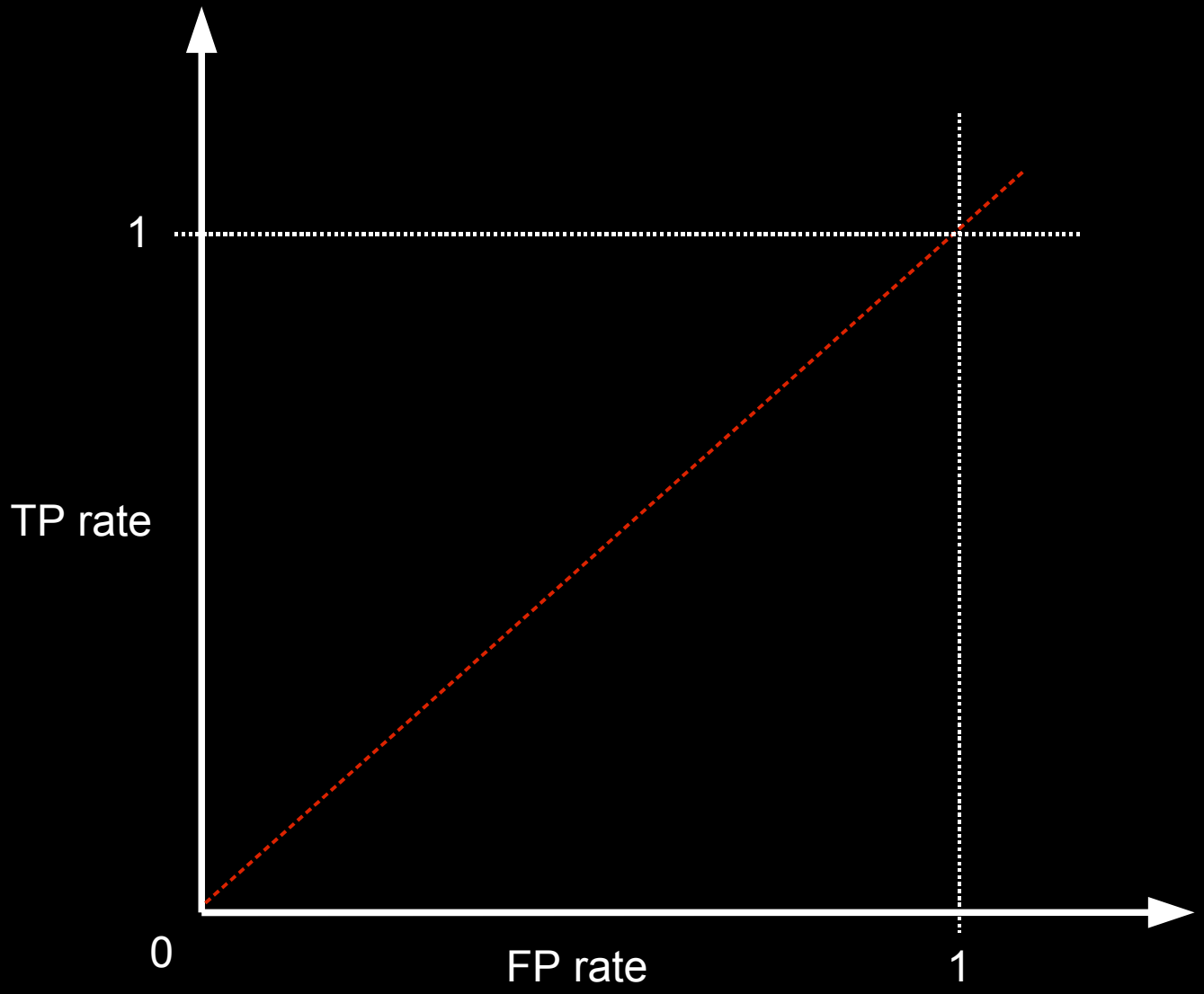
	Predicted positive	Predicted negative
Positive examples	30	40
Negative examples	10	20

More Definitions

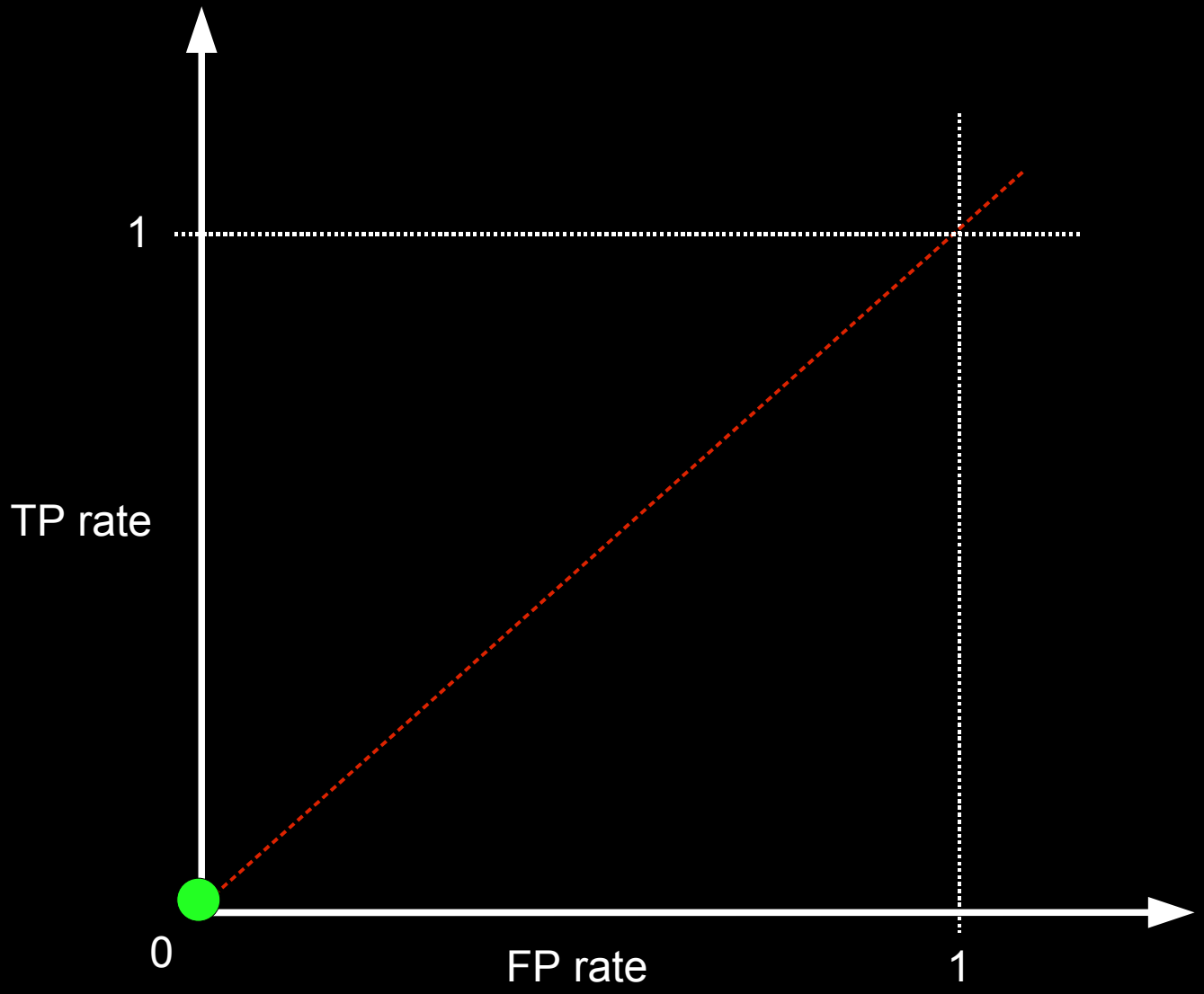
	Predicted positive	Predicted negative
Positive examples	TP	FN
Negative examples	FP	TN

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

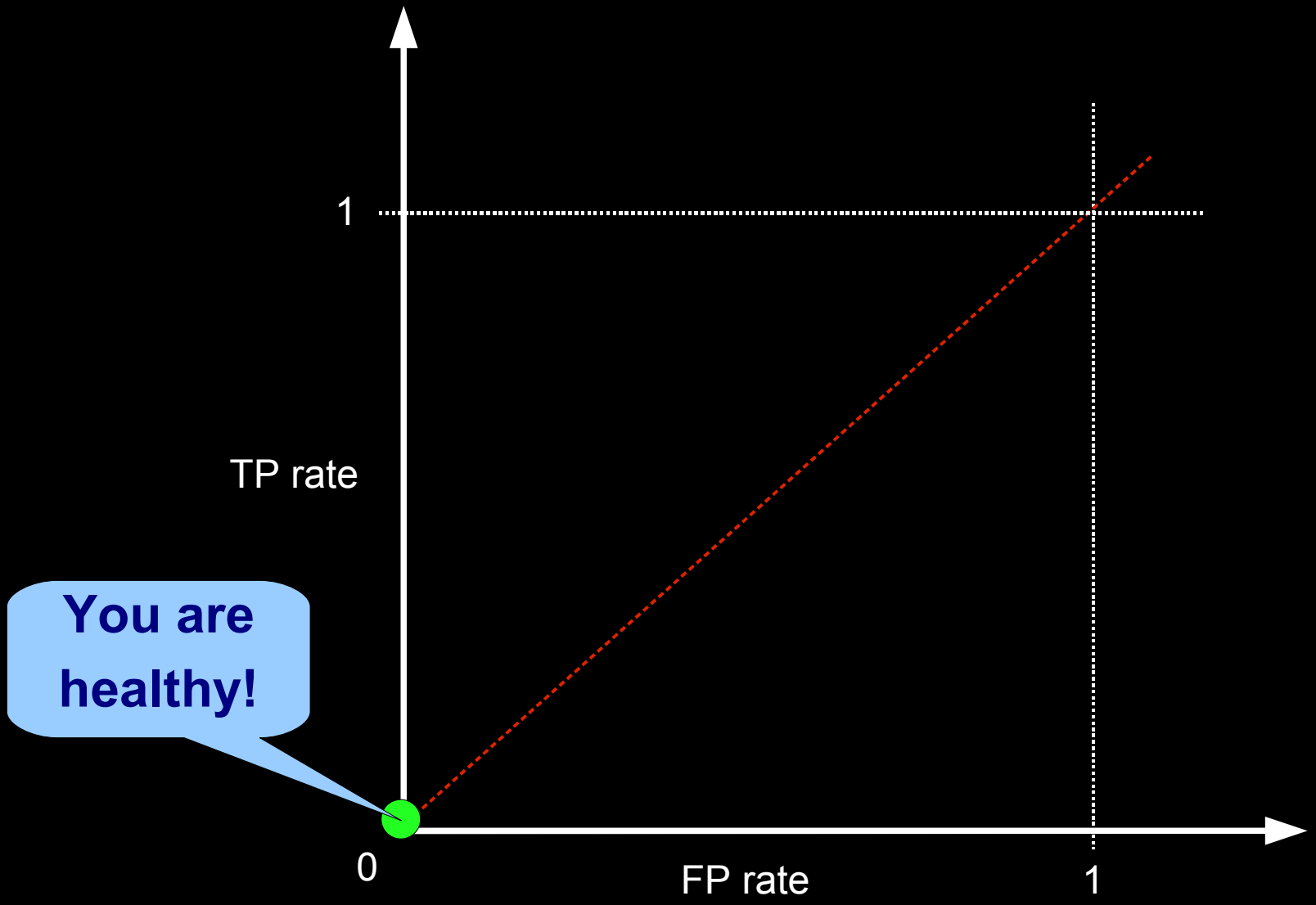
ROC - Receiver Operating Characteristic Curve



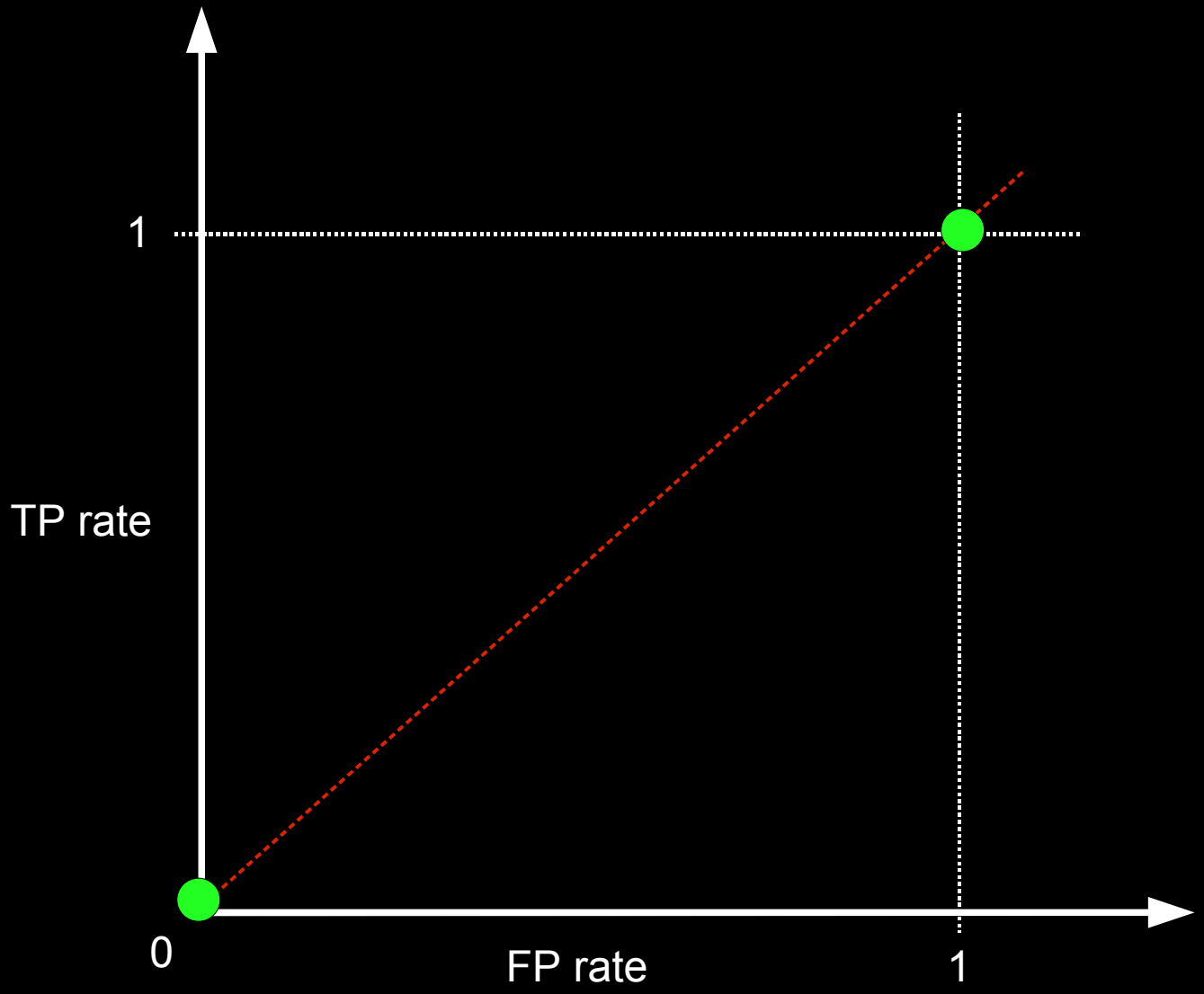
ROC Curve



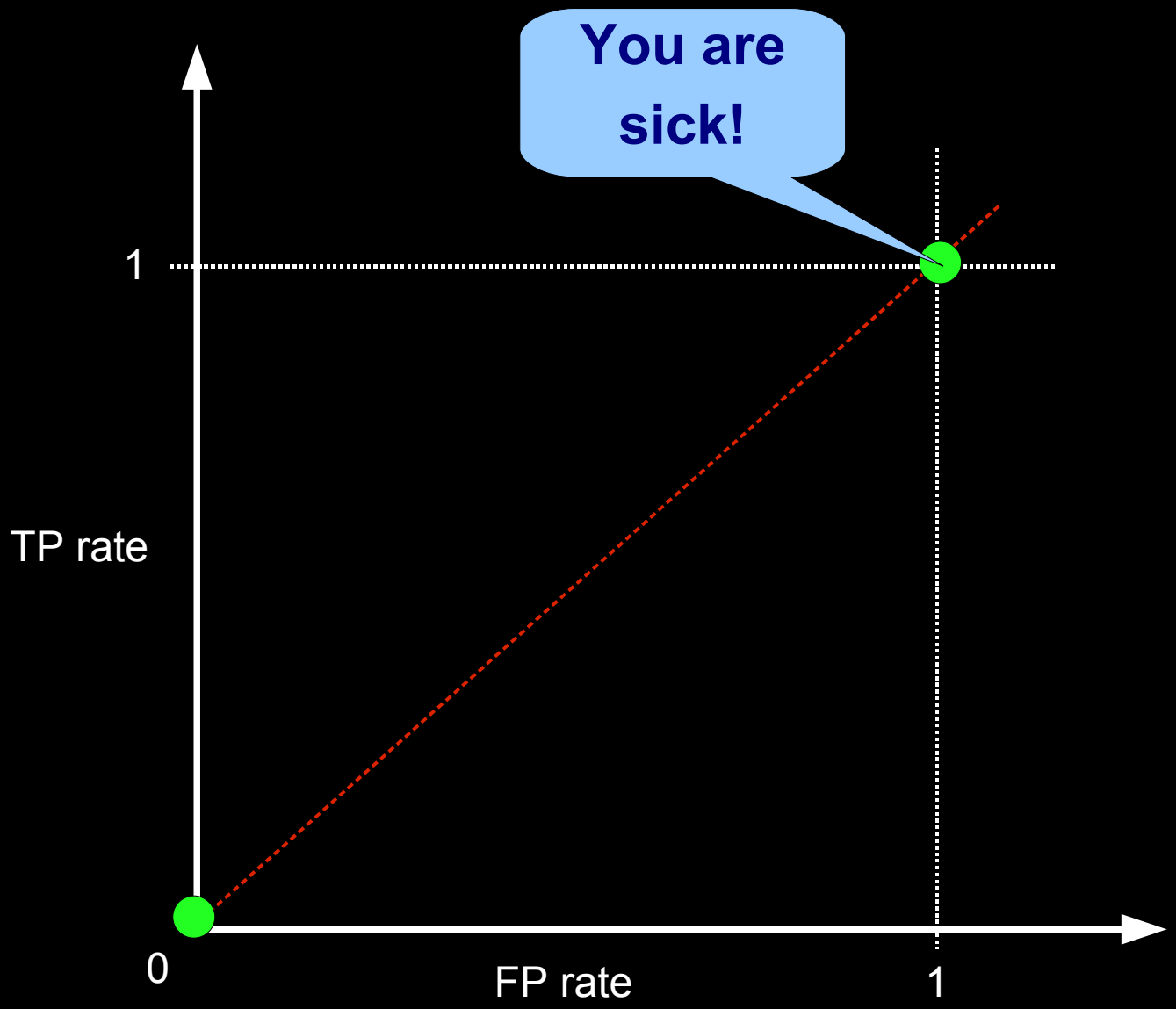
ROC Curve



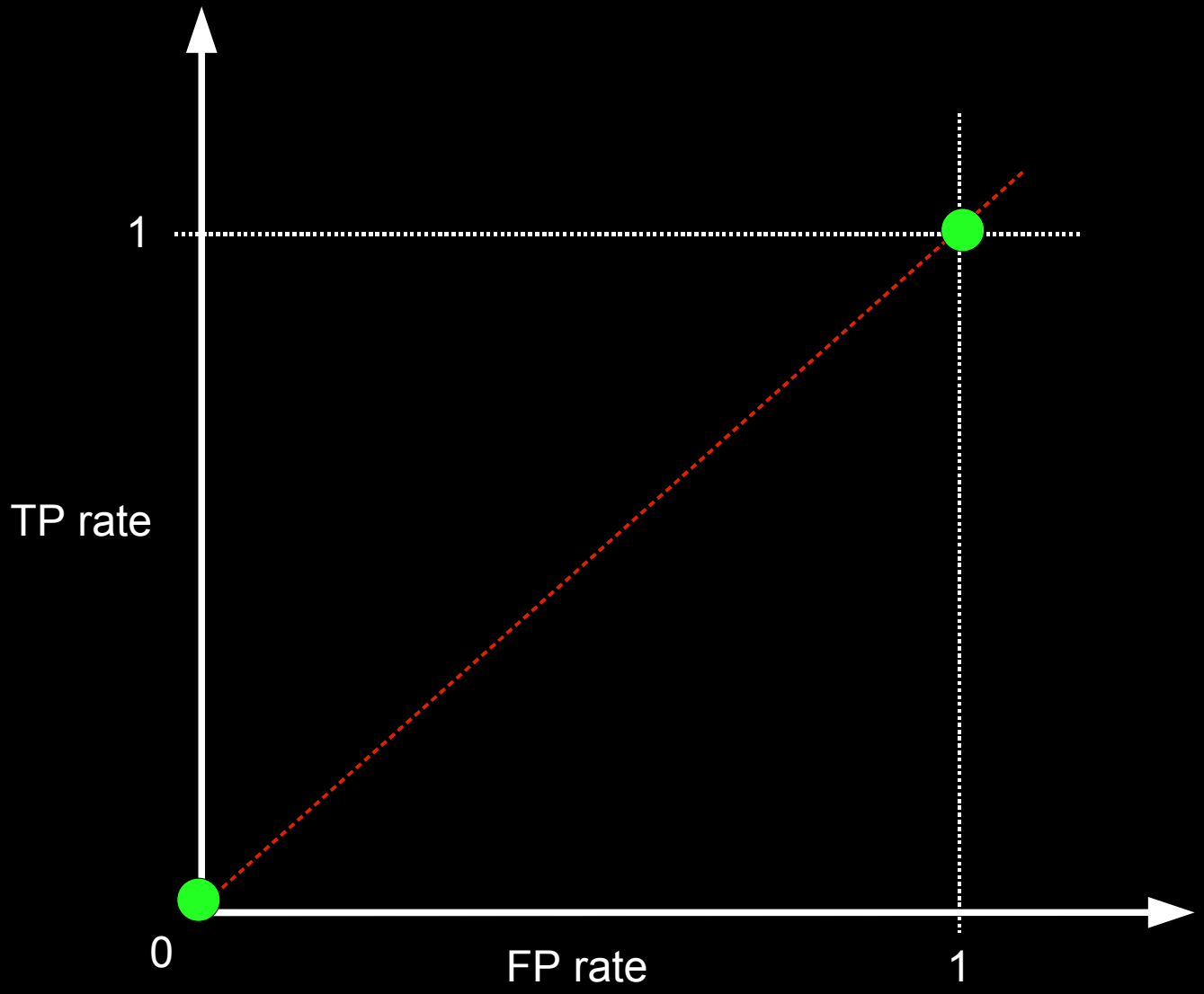
ROC Curve



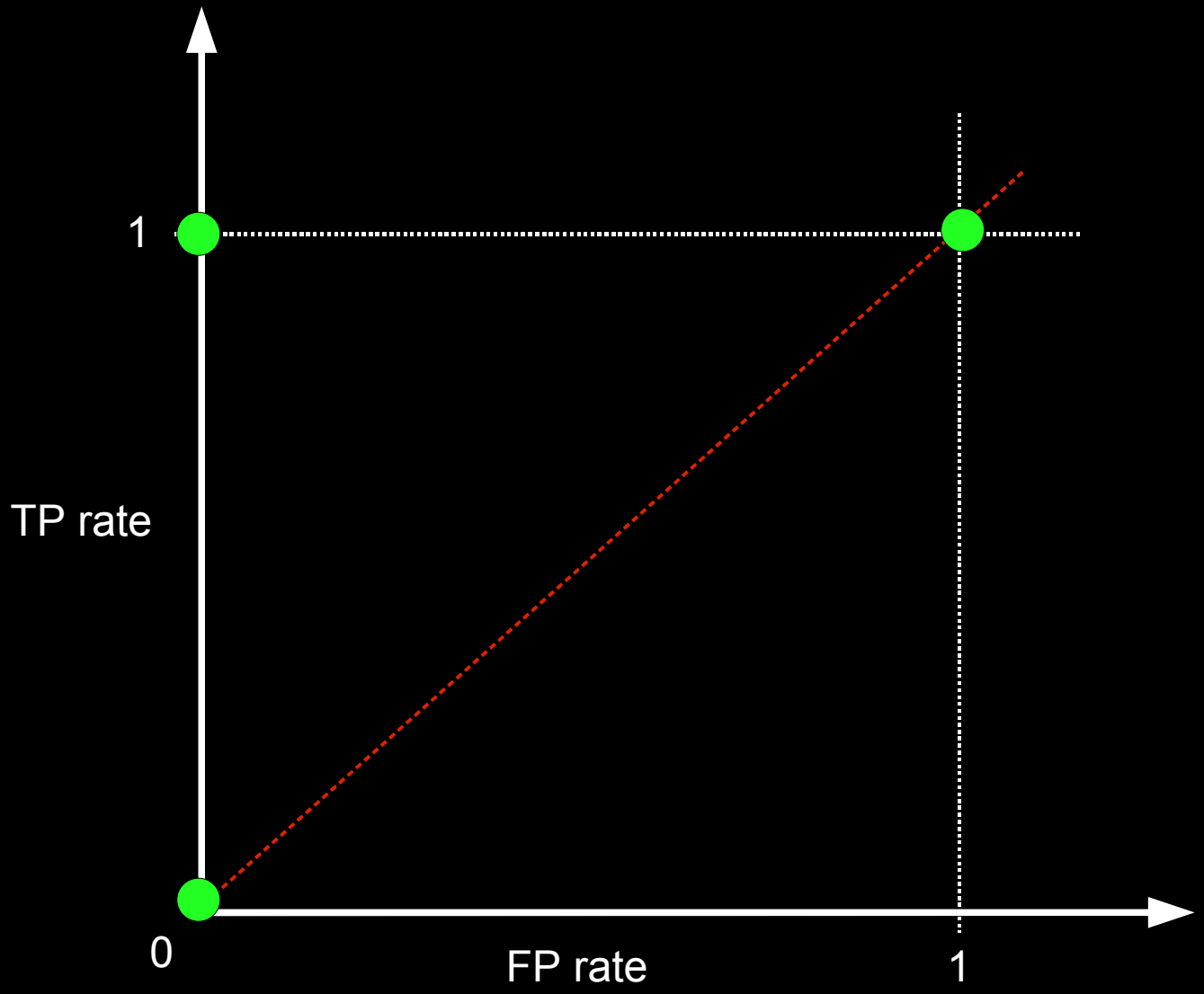
ROC Curve



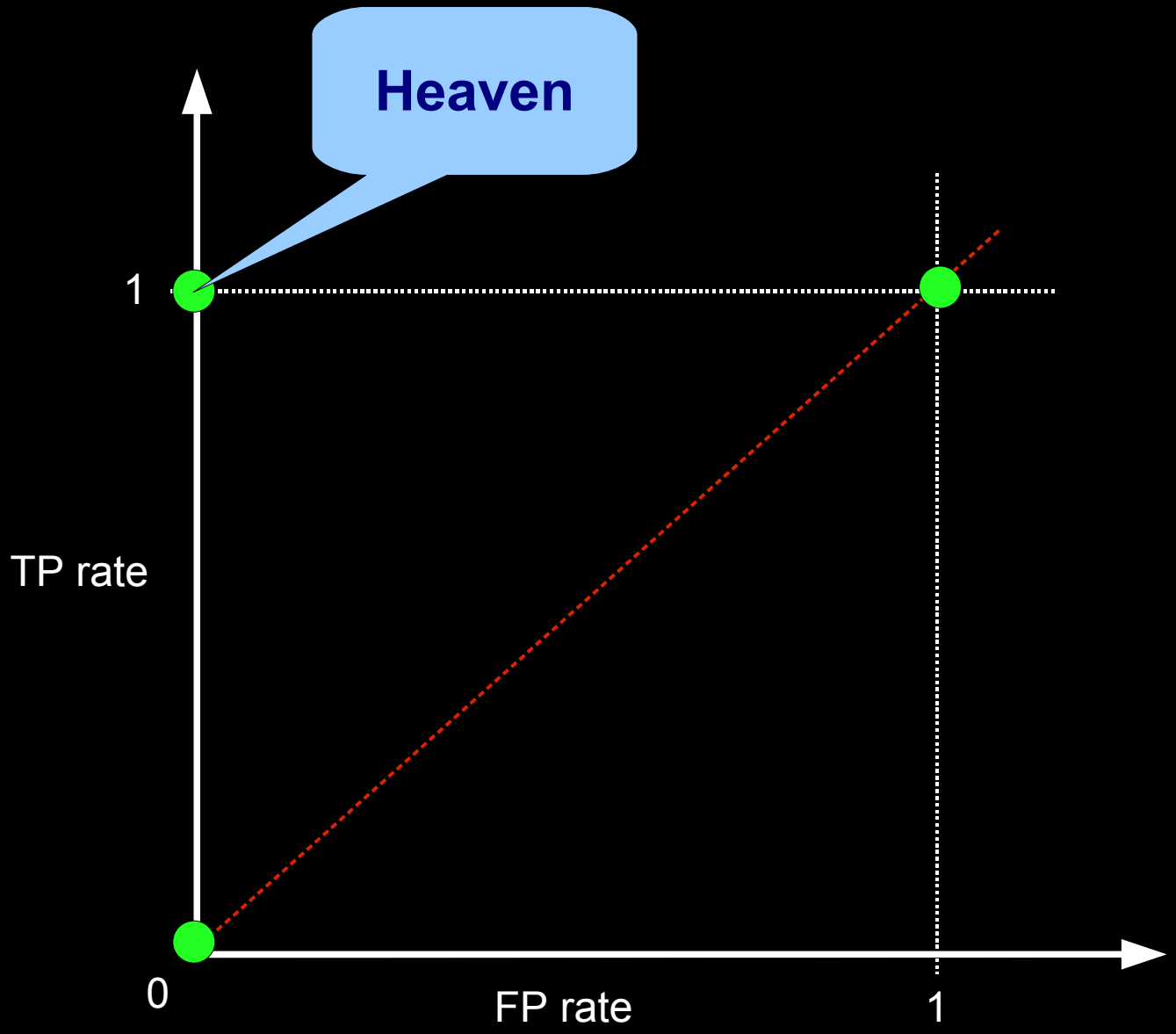
ROC Curve



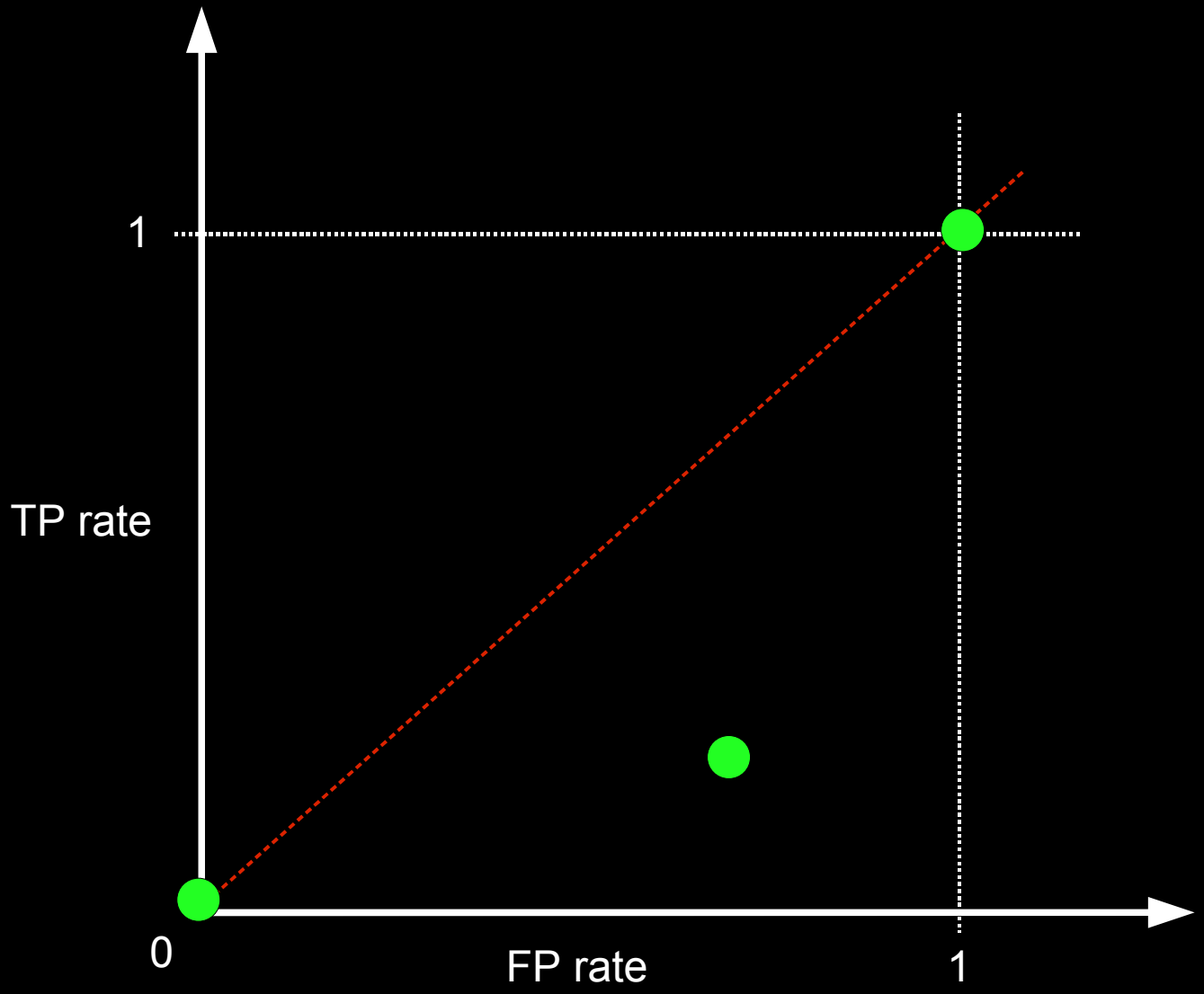
ROC Curve



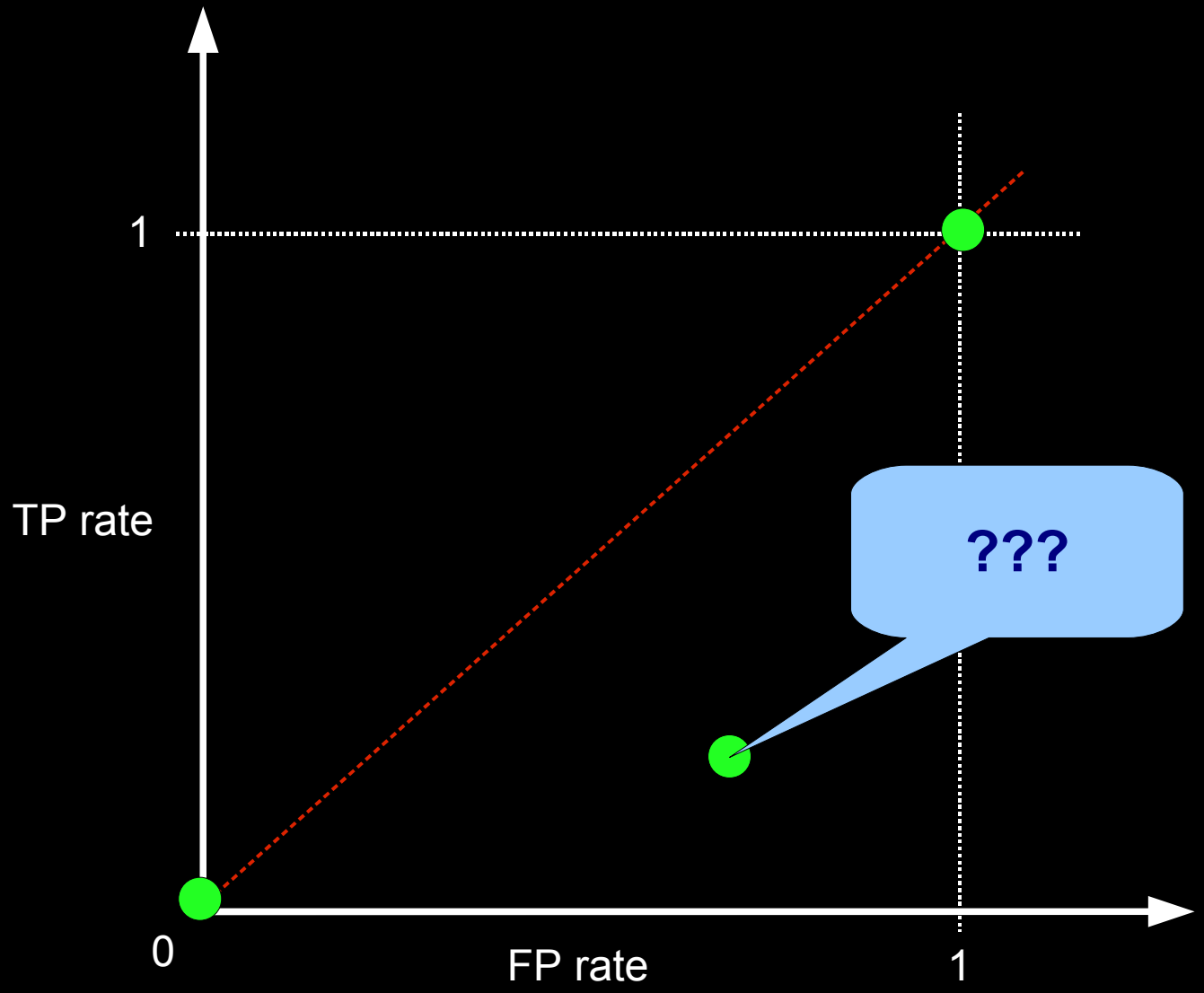
ROC Curve



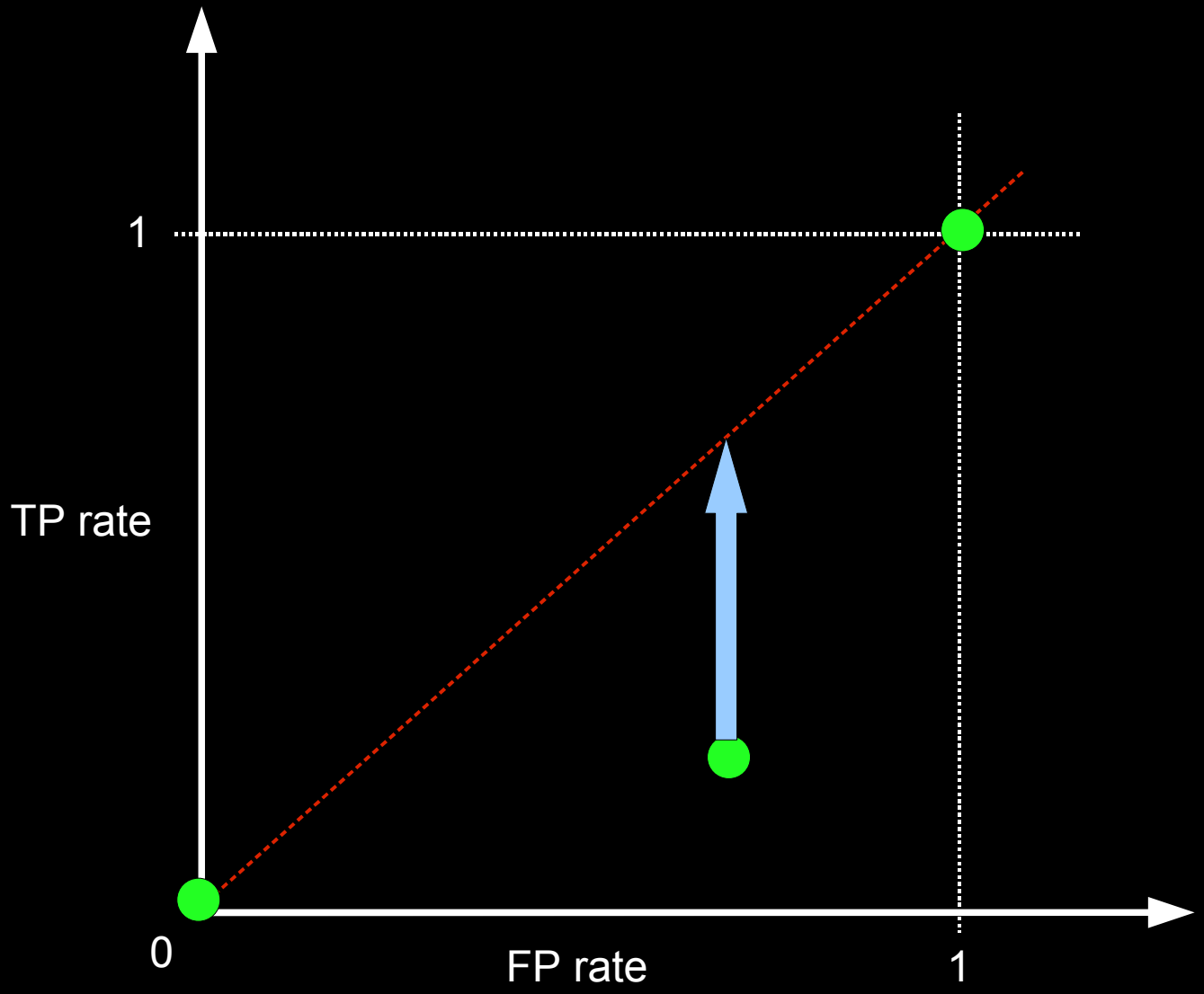
ROC Curve



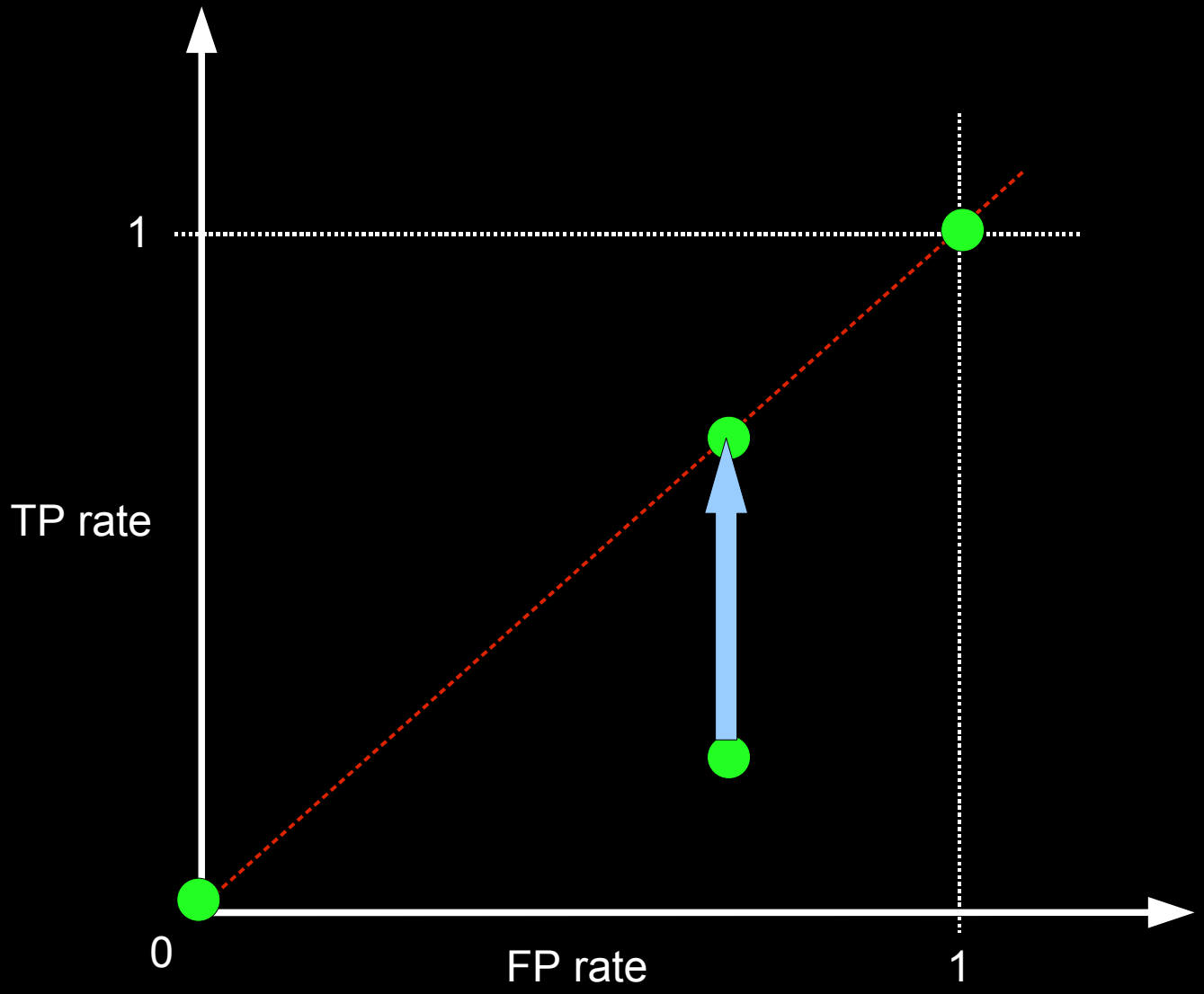
ROC Curve



ROC Curve

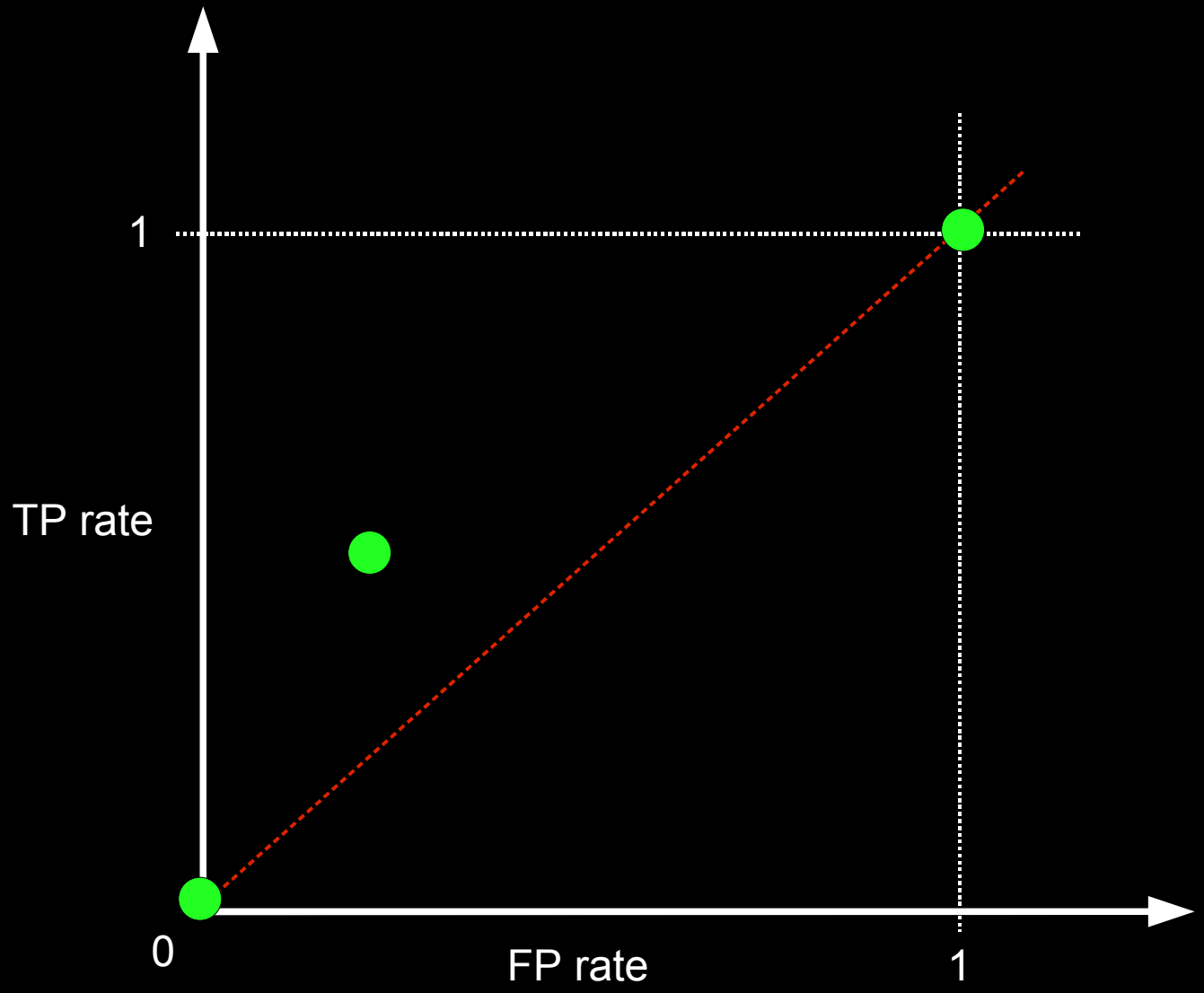


ROC Curve

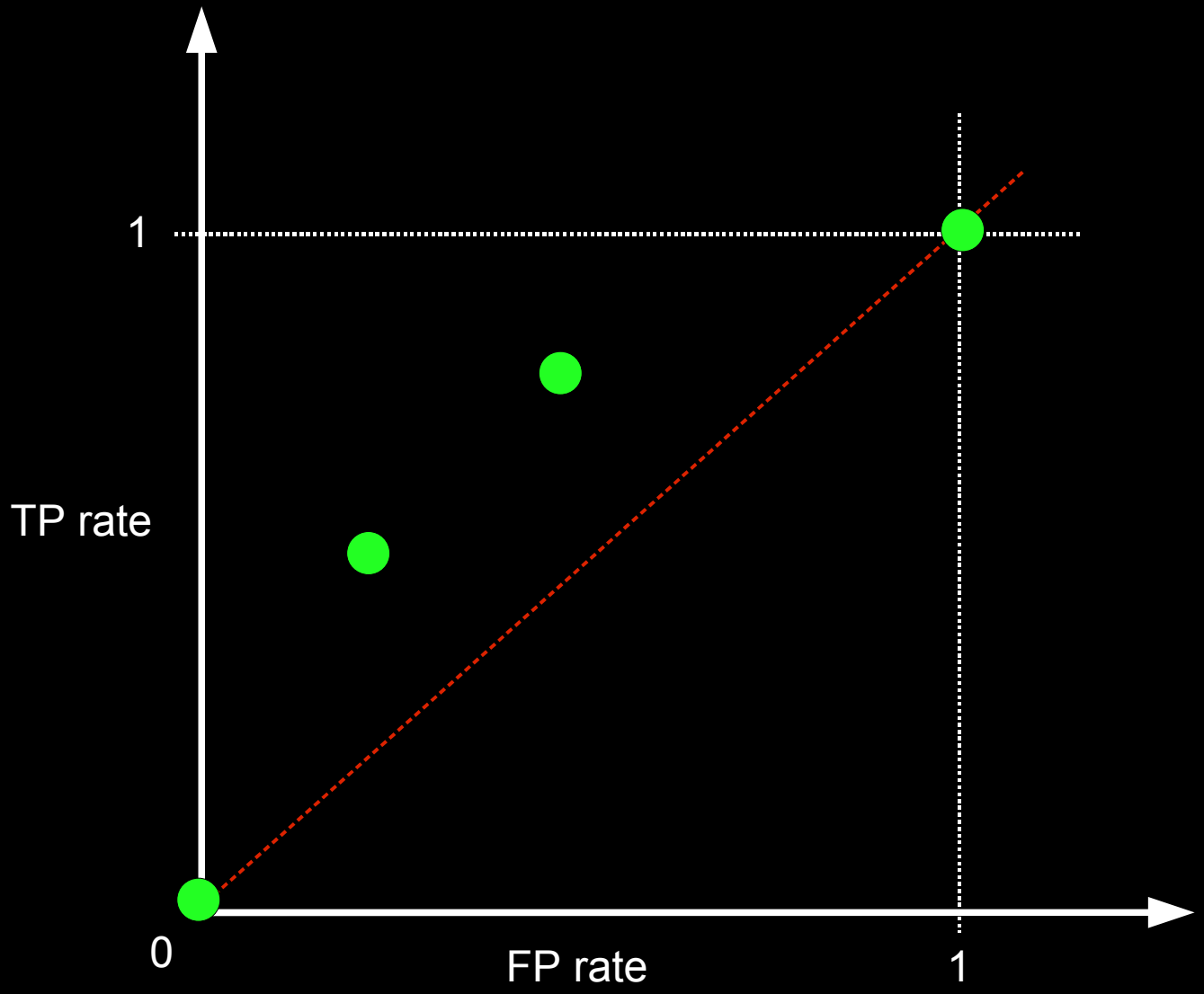


ROC Curve

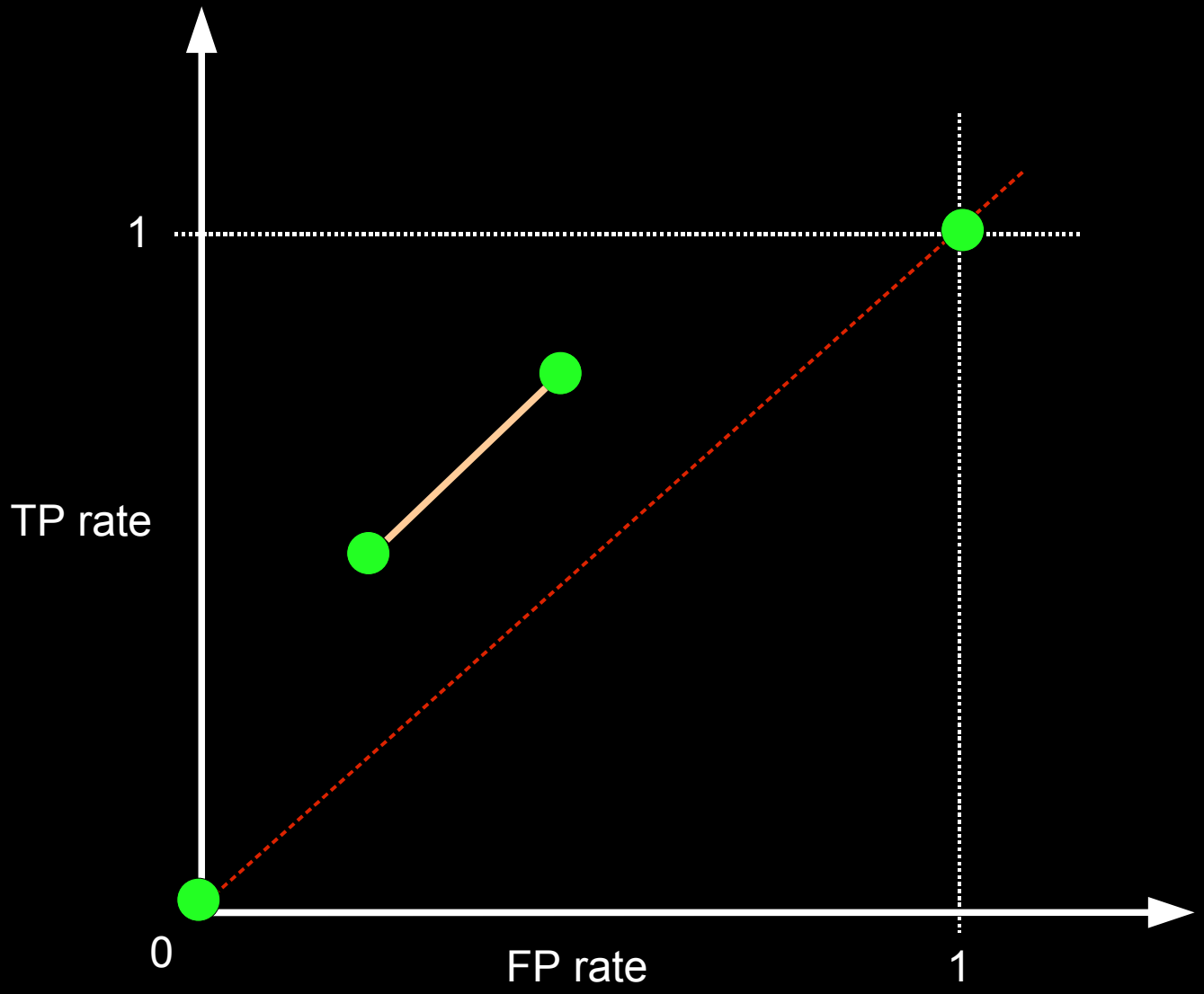
ROC Curve



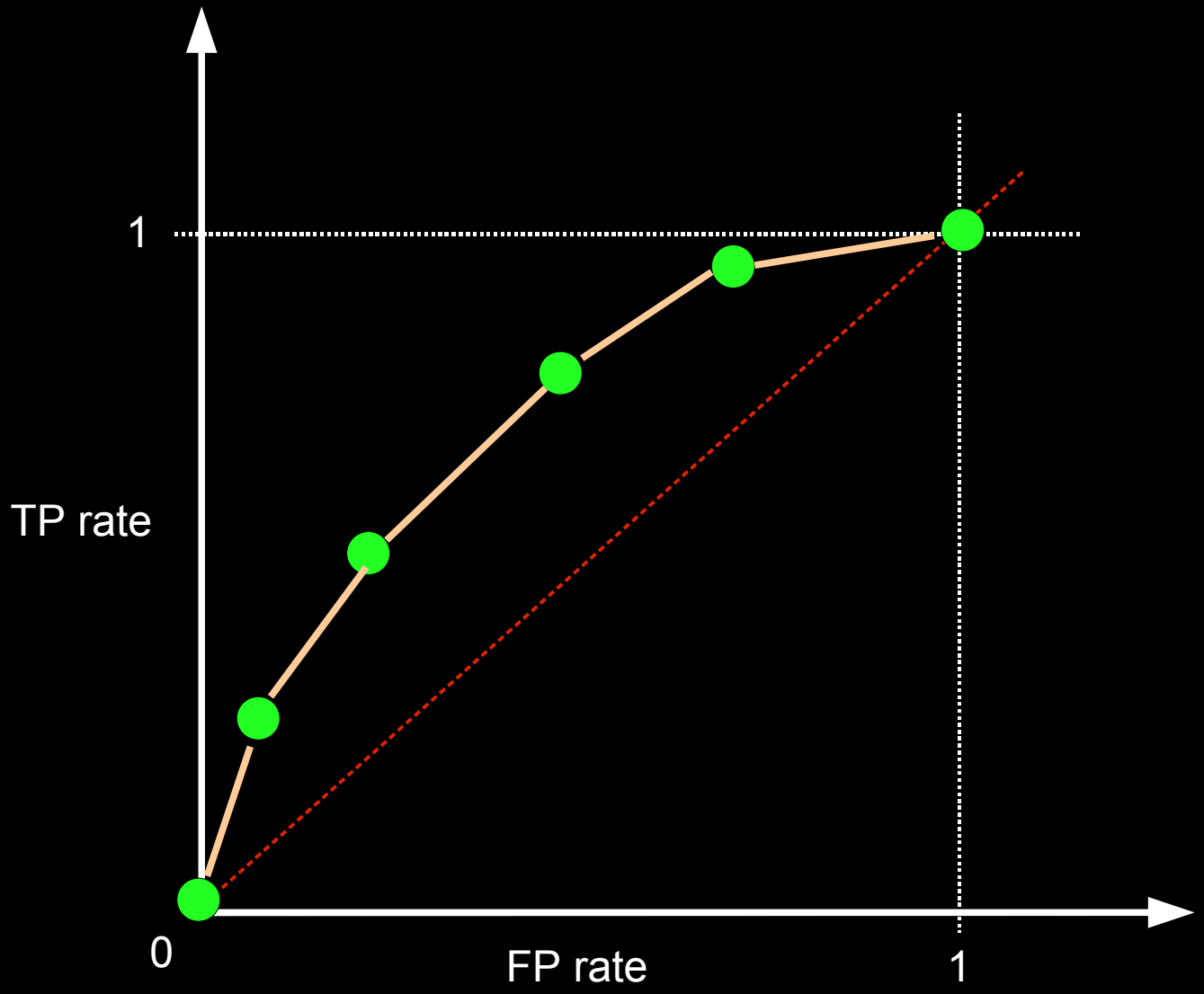
ROC Curve



ROC Curve



ROC Curve



Alternative Terminology

- Precision = TP / (TP+FP)
  (= positive predictive value)
- Recall = TP / (TP+FN)
  (= sensitivity = true positive rate)
- F-measure - the harmonic mean of precision and recall:
  F-score = 2 Precision × Recall / (Precision+Recall)

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

Alternative Terminology (cont.)

- Specificity = TN / (FP+TN)
  (= 1 - false positive rate)
- Negative predictive value = TN / (FN+TN)

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

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.

Choosing a Point on the Curve

- Depends on the application:
  - Medical screening tests (e.g., mammography) - high TP.
  - Spam filtering - low FP.

Summary

- Methods for:
  - Estimation properties of an estimator.
  - Model selection.

References

- Bootstrap Methods and their Applications. A, Davison and D. Hinkley.
- The Elements of Statistical Learning. T. Hastie, R. Tibshirani and J. H. Friedman.
- ROC Graphs: Notes and Practical Considerations for Researchers. T. Fawcett.