Model Evaluation

MSc. Bui Quoc Khanh

khanhbq@hanu.edu.vn

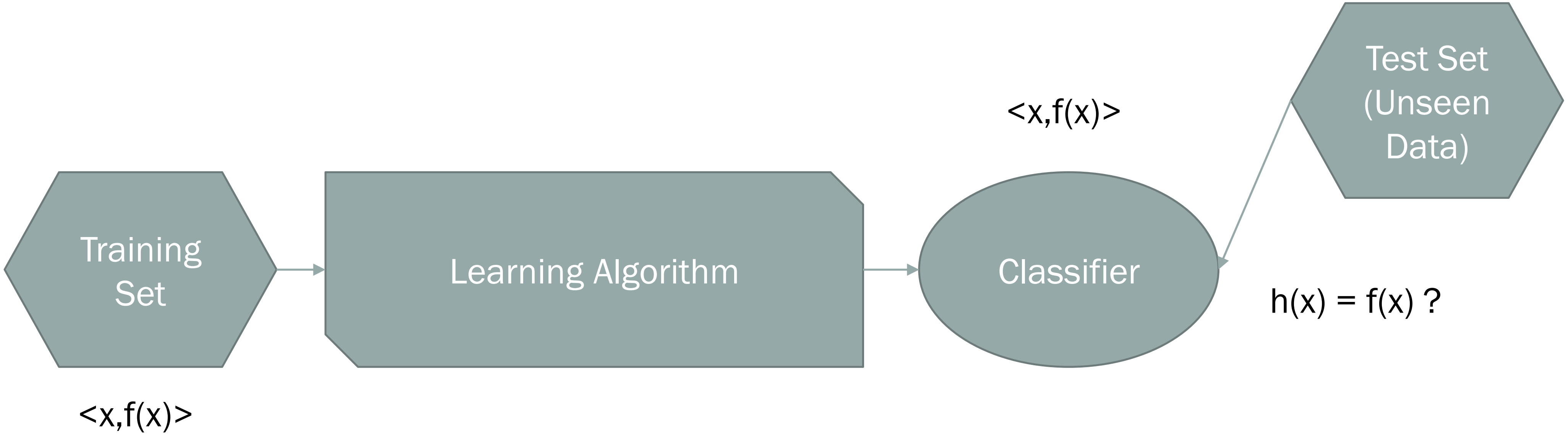
MODEL EVALUATION

- How do we assess the generalization capabilities of a learned hypothesis?
- Metrics for Performance Evaluation
  - How to evaluate the performance (prediction capability) of a model?
- Methods for Performance Evaluation
- How to obtain reliable estimates?

Testing the classifier

- We are given a learning algorithm A and a data set S of labeled instances <x, f(x)>
- The general idea to assess the generalization capabilities of A is that of splitting S into a subset used for training and a subset used for testing

Testing the classifier



Testing the classifier

Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt
Sunny	Warm	Normal	Strong	Warm	Change	Y
Rainy	Cold	Normal	Light	Warm	Same	N
Sunny	Warm	High	Light	Warm	Same	Y
Sunny	Cold	Normal	Strong	Warm	Same	Y
Sunny	Cold	high	Strong	Cool	Change	Y
Rainy	Warm	Normal	Light	Warm	Change	N
Rainy	Warm	Low	Light	Warm	Same	N
Sunny	Cold	Normal	Strong	Cool	Change	Y
Rainy	Warm	Normal	Light	Warm	Change	N
Sunny	Cold	Normal	Strong	Warm	Change	Y
Sunny	Warm	Normal	Strong	Cool	Change	N
Rainy	Cold	Normal	Light	Warm	Same	N
Sunny	Warm	Normal	Light	Warm	Same	Y
Sunny	Cold	Normal	Strong	Warm	Same	Y
Sunny	Warm	high	Strong	Cool	Change	N
Rainy	Warm	Normal	Light	Warm	Change	N
Rainy	Warm	Normal	Light	Warm	Same	N
Sunny	Cold	Normal	Strong	Warm	Change	Y

Target function

Testing the classifier

Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt
Sunny	Warm	Normal	Strong	Warm	Change	Y
Rainy	Cold	Normal	Light	Warm	Same	N
Sunny	Warm	High	Light	Warm	Same	Y
Sunny	Cold	Normal	Strong	Warm	Same	Y
Sunny	Cold	high	Strong	Cool	Change	Y
Rainy	Warm	Normal	Light	Warm	Change	N
Rainy	Warm	Low	Light	Warm	Same	N
Sunny	Cold	Normal	Strong	Cool	Change	Y
Rainy	Warm	Normal	Light	Warm	Change	N
Sunny	Cold	Normal	Strong	Warm	Change	Y
Sunny	Warm	Normal	Strong	Cool	Change	N
Rainy	Cold	Normal	Light	Warm	Same	N
Sunny	Warm	Normal	Light	Warm	Same	Y
Sunny	Cold	Normal	Strong	Warm	Same	Y
Sunny	Warm	high	Strong	Cool	Change	N
Rainy	Warm	Normal	Light	Warm	Change	N
Rainy	Warm	Normal	Light	Warm	Same	N
Sunny	Cold	Normal	Strong	Warm	Change	Y

Target function

Training set

Test set

Testing the classifier

Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt
Sunny	Warm	Normal	Strong	Warm	Change	Y
Rainy	Cold	Normal	Light	Warm	Same	N
Sunny	Warm	High	Light	Warm	Same	Y
Sunny	Cold	Normal	Strong	Warm	Same	Y
Sunny	Cold	high	Strong	Cool	Change	Y
Rainy	Warm	Normal	Light	Warm	Change	N
Rainy	Warm	Low	Light	Warm	Same	N
Sunny	Cold	Normal	Strong	Cool	Change	Y
Rainy	Warm	Normal	Light	Warm	Change	N
Sunny	Cold	Normal	Strong	Warm	Change	Y
Sunny	Warm	Normal	Strong	Cool	Change	N
Rainy	Cold	Normal	Light	Warm	Same	N
Sunny	Warm	Normal	Light	Warm	Same	Y
Sunny	Cold	Normal	Strong	Warm	Same	Y
Sunny	Warm	high	Strong	Cool	Change	N
Rainy	Warm	Normal	Light	Warm	Change	N
Rainy	Warm	Normal	Light	Warm	Same	N
Sunny	Cold	Normal	Strong	Warm	Change	Y

Target function

Learn hypothesis h

Apply h

Testing the classifier

Target function

TEST SET						
Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt
Sunny	Warm	Normal	Strong	Cool	Change	N
Rainy	Cold	Normal	Light	Warm	Same	N
Sunny	Warm	Normal	Light	Warm	Same	Y
Sunny	Cold	Normal	Strong	Warm	Same	Y
Sunny	Warm	high	Strong	Cool	Change	N
Rainy	Warm	Normal	Light	Warm	Change	N
Rainy	Warm	Normal	Light	Warm	Same	N
Sunny	Cold	Normal	Strong	Warm	Change	Y

Testing the classifier

Is h a good predictive capability?

TEST SET							
Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt	h(x)
Sunny	Warm	Normal	Strong	Cool	Change	No	No
Rainy	Cold	Normal	Light	Warm	Same	No	No
Sunny	Warm	Normal	Light	Warm	Same	Yes	Yes
Sunny	Cold	Normal	Strong	Warm	Same	Yes	No
Sunny	Warm	high	Strong	Cool	Change	No	No
Rainy	Warm	Normal	Light	Warm	Change	No	No
Rainy	Warm	Normal	Light	Warm	Same	No	Yes
Sunny	Cold	Normal	Strong	Warm	Change	Yes	No

Target function

Predicted class

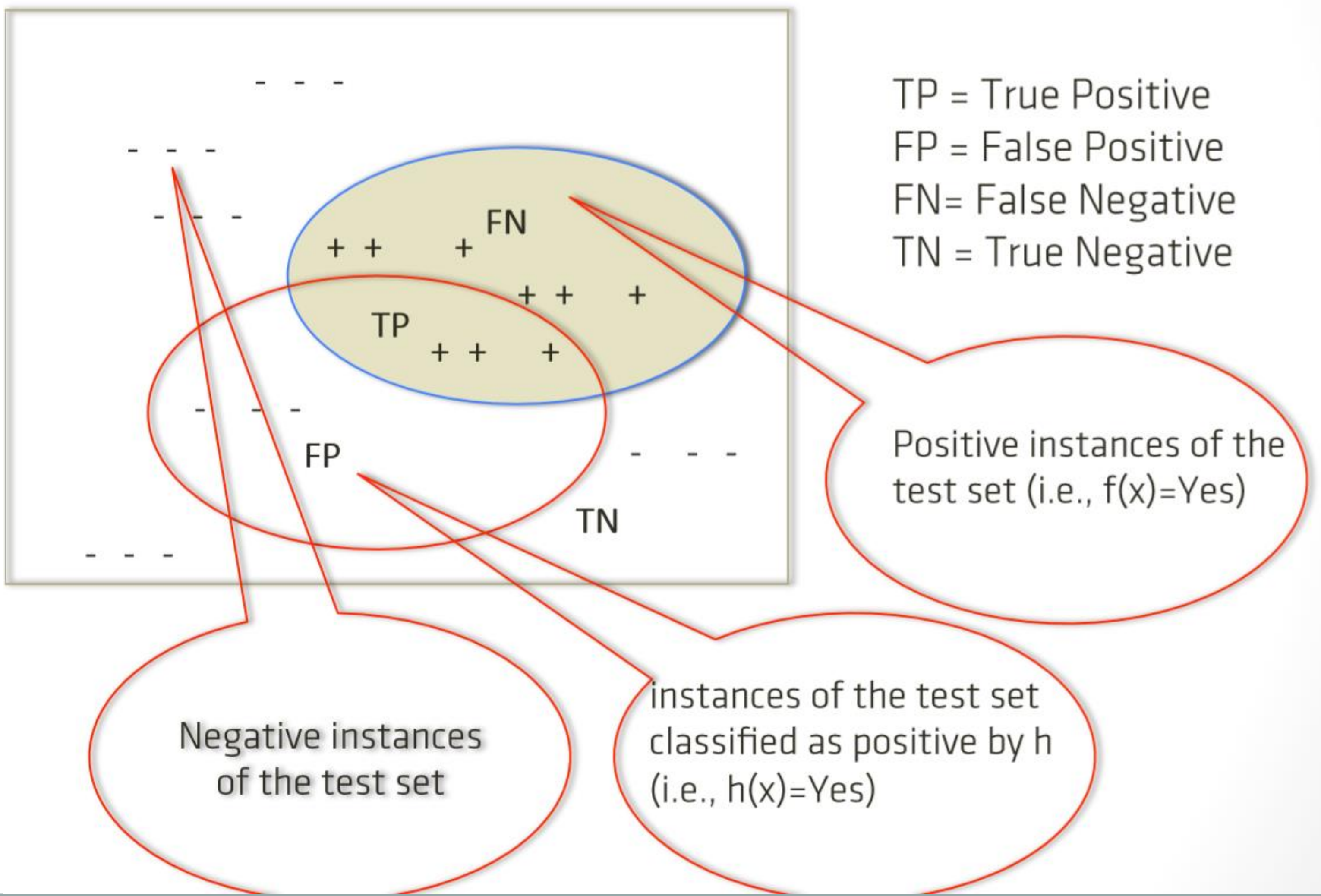
Testing the classifier

TEST SET								
Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt	h(x)	
Sunny	Warm	Normal	Strong	Cool	Change	No	No	TN
Rainy	Cold	Normal	Light	Warm	Same	No	No	TN
Sunny	Warm	Normal	Light	Warm	Same	Yes	Yes	TP
Sunny	Cold	Normal	Strong	Warm	Same	Yes	No	FN
Sunny	Warm	high	Strong	Cool	Change	No	No	TN
Rainy	Warm	Normal	Light	Warm	Change	No	No	TN
Rainy	Warm	Normal	Light	Warm	Same	No	Yes	FP
Sunny	Cold	Normal	Strong	Warm	Change	Yes	No	FN

Has h a good predictive capability?

To answer this question we compare h(x) with the target function EnjoySpt

Metrics for Performance Evaluation



Metrics for Performance Evaluation Error and Accuracy

- Accuracy: number of instances correctly classified over the total number of predictions
  - $Accuracy = \frac{(TP+TN)}{N}$ where N = TP+TN+FN+FP
- Error: number of instances misclassified over the total number of predictions
  - $Error = \frac{(FP+FN)}{N} = 1 - Accuracy$

Metrics for Performance Evaluation
An Example

TEST SET								
Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt	h(x)	
Sunny	Warm	Normal	Strong	Cool	Change	No	No	TN
Rainy	Cold	Normal	Light	Warm	Same	No	No	TN
Sunny	Warm	Normal	Light	Warm	Same	Yes	Yes	TP
Sunny	Cold	Normal	Strong	Warm	Same	Yes	No	FN
Sunny	Warm	high	Strong	Cool	Change	No	No	TN
Rainy	Warm	Normal	Light	Warm	Change	No	No	TN
Rainy	Warm	Normal	Light	Warm	Same	No	Yes	FP
Sunny	Cold	Normal	Strong	Warm	Change	Yes	No	FN

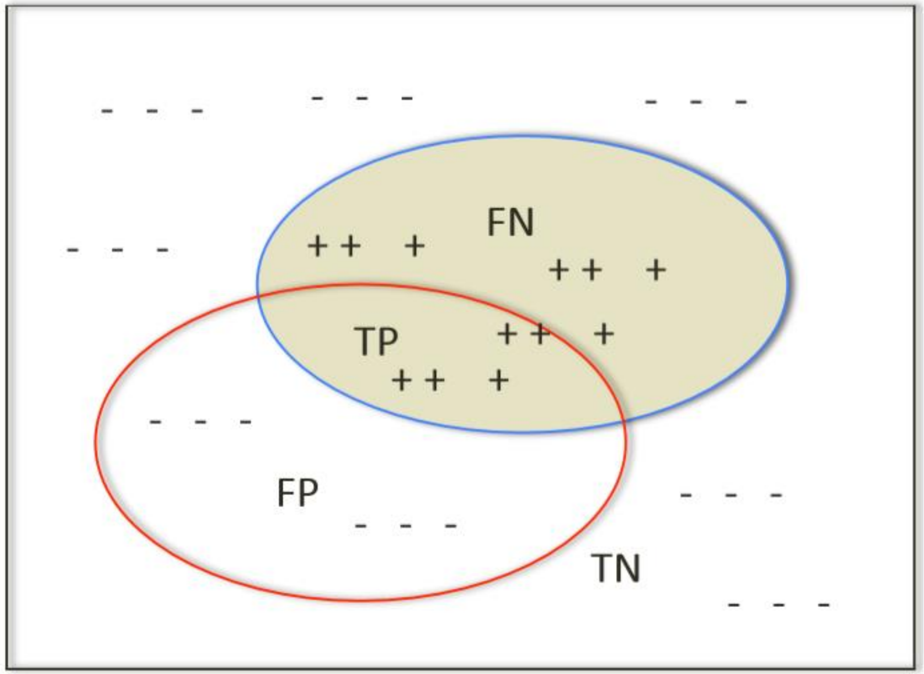
TP = 1 FP= 1 FN = 2 TN = 4

- Acc = 5/8 = 62,5%
- Err = 3/8 = 37,5%

Metrics for Performance Evaluation
An Example (cont'ed)

- Assume that no instance is classified positively by h (rejector)
  - TP=0, FP=0, FN=3, TN=5 (Acc= 62,5% (!!))
- The accuracy may not be an adequate performance measure when the number of negative cases is much greater than the number of positive ones
- Suppose there are 1000 examples, 995 of which are negative cases and 5 are positive cases. If the system classifies them all as negative (rejector)
  - TP=0, TN=995, FP=0, FN=5
- the accuracy would be 99.5%, even though the classifier missed all positive cases

Metrics for Performance Evaluation
Precision



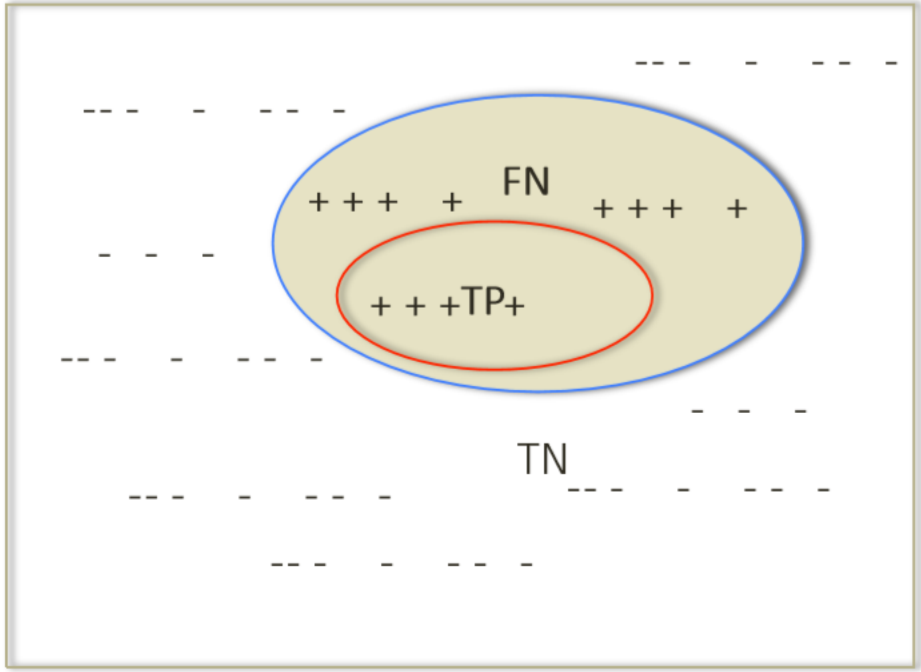
- $Precision = \frac{TP}{TP+FP}$
- fraction of instances correctly classified

Metrics for Performance Evaluation An Example

TEST SET								
Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt	h(x)	
Sunny	Warm	Normal	Strong	Cool	Change	No	No	TN
Rainy	Cold	Normal	Light	Warm	Same	No	No	TN
Sunny	Warm	Normal	Light	Warm	Same	Yes	Yes	TP
Sunny	Cold	Normal	Strong	Warm	Same	Yes	No	FN
Sunny	Warm	high	Strong	Cool	Change	No	No	TN
Rainy	Warm	Normal	Light	Warm	Change	No	No	TN
Rainy	Warm	Normal	Light	Warm	Same	No	Yes	FP
Sunny	Cold	Normal	Strong	Warm	Change	Yes	No	FN

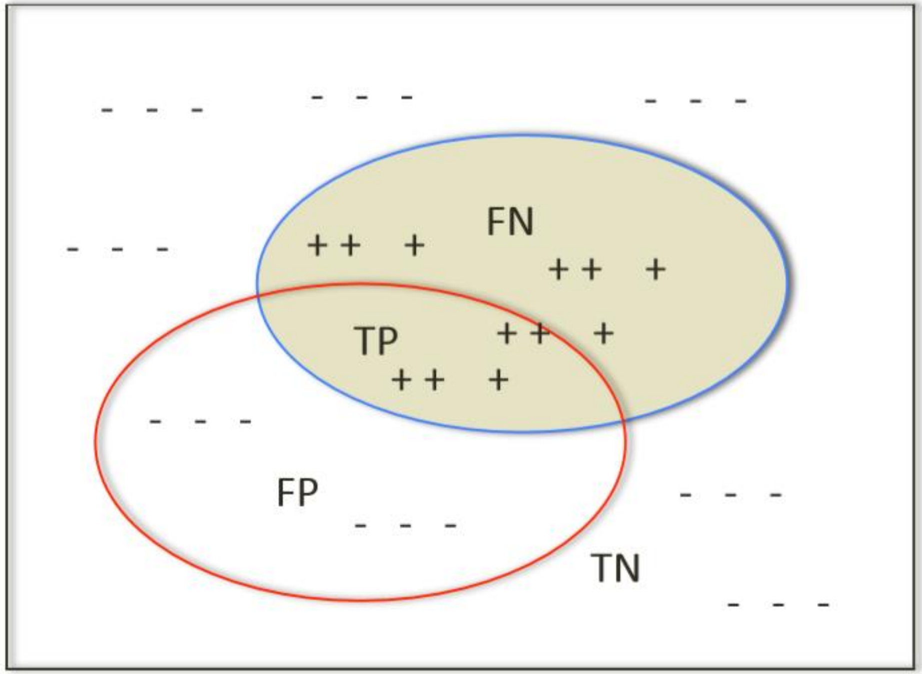
- TP = 1 FP= 1 FN = 2 TN = 4
- Pr= TP/(TP+FP) = 1/2 = 0.5

Metrics for Performance Evaluation Precision



- FP=0 => Pr=1 – i.e., no negative examples classified as positive
- A classifier may have high precision but low coverage
- Precision alone not sufficient

Metrics for Performance Evaluation
Recall



- $Recall = \frac{TP}{TP+FN}$
- Fraction of positive examples in the test set that have been correctly classified
- Also called coverage, sensitivity or True Positive Rate

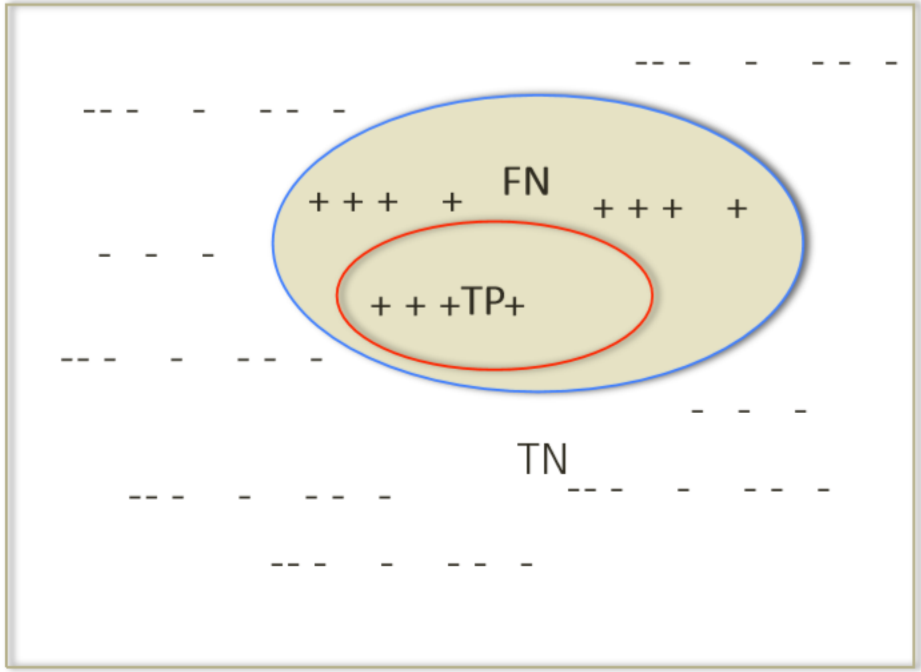
Metrics for Performance Evaluation An Example

TEST SET								
Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt	h(x)	
Sunny	Warm	Normal	Strong	Cool	Change	No	No	TN
Rainy	Cold	Normal	Light	Warm	Same	No	No	TN
Sunny	Warm	Normal	Light	Warm	Same	Yes	Yes	TP
Sunny	Cold	Normal	Strong	Warm	Same	Yes	No	FN
Sunny	Warm	high	Strong	Cool	Change	No	No	TN
Rainy	Warm	Normal	Light	Warm	Change	No	No	TN
Rainy	Warm	Normal	Light	Warm	Same	No	Yes	FP
Sunny	Cold	Normal	Strong	Warm	Change	Yes	No	FN

TP = 1 FP= 1 FN = 2 TN = 4

- Re= TP/(TP+FN) = 1/3 = 0.33

Metrics for Performance Evaluation
Recall



- FN=0 => Re=1 – i.e., all positive examples have been correctly classified
- A classifier may have high coverage but low precision
- Recall alone not sufficient

Metrics for Performance Evaluation F-measure

- $F - Measure\ F = \frac{2PrRe}{Pr+Re}$
- F ≈ min(Pr, Re)
- F is high when both Pr and Re are high (good classifier)

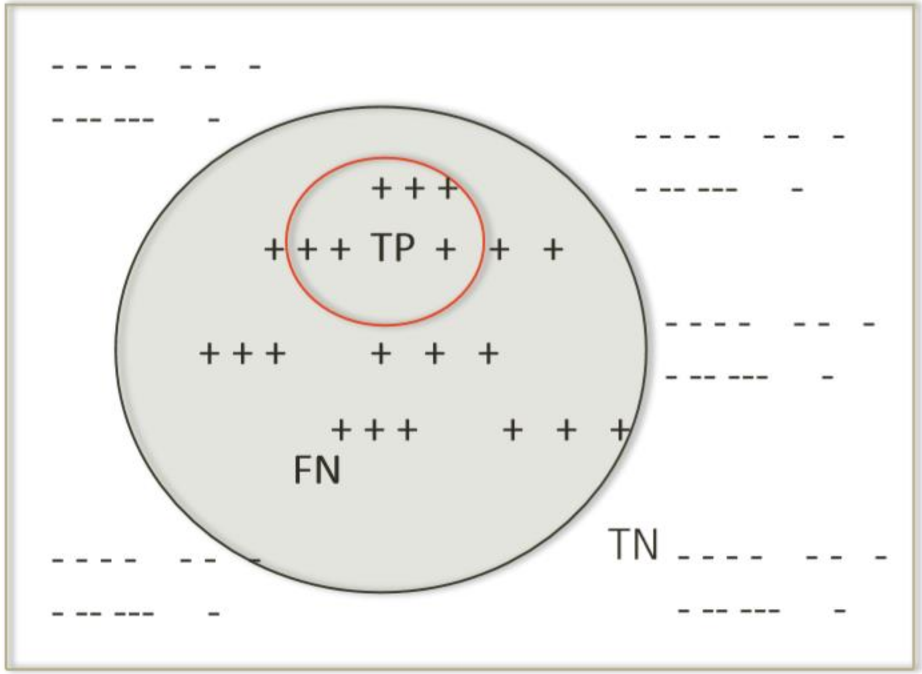
Metrics for Performance Evaluation
An Example

TEST SET								
Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt	h(x)	
Sunny	Warm	Normal	Strong	Cool	Change	No	No	TN
Rainy	Cold	Normal	Light	Warm	Same	No	No	TN
Sunny	Warm	Normal	Light	Warm	Same	Yes	Yes	TP
Sunny	Cold	Normal	Strong	Warm	Same	Yes	No	FN
Sunny	Warm	high	Strong	Cool	Change	No	No	TN
Rainy	Warm	Normal	Light	Warm	Change	No	No	TN
Rainy	Warm	Normal	Light	Warm	Same	No	Yes	FP
Sunny	Cold	Normal	Strong	Warm	Change	Yes	No	FN

TP = 1 FP= 1 FN = 2 TN = 4

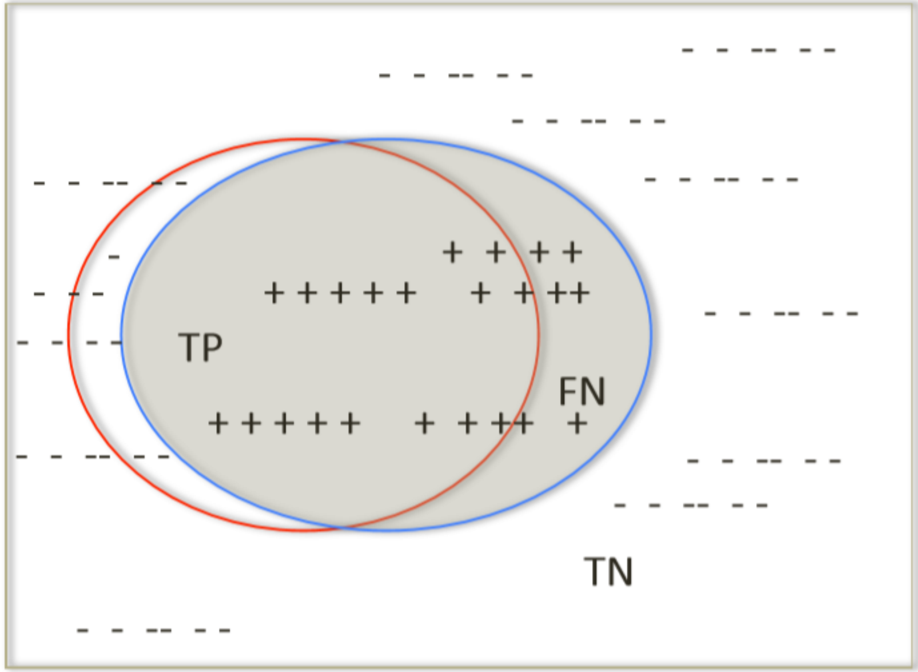
- Pr = 0.5
- Re= 0.33
- F = 2PrRe/(Pr+Re) = 0,37

Metrics for Performance Evaluation F- Measure



- Pr=1, Re=0.2
- F = 0.33
- F ≈ min(Pr, Re)

Metrics for Performance Evaluation
F- Measure



- Pr=0.9, Re=0.8
- F = 0.85
- F ≈ min(Pr, Re)
- High F-measure is indicative of both high precision and high recall

Metrics for Performance Evaluation Summary

- Let N = TP+TN+FN+FP
  - Precision P = TP/(TP+FP)
  - Recall R = TP/(TP+FN)
  - F-measure F1 = 2PxR/(P+R)
  - Accuracy A = (TP+TN)/N
  - Error E = (FP+FN)/N= 1- A

Metrics for Performance Evaluation Confusion Matrix

- A classification system has been trained to distinguish between dogs, mouses and chickens
- The classification results can be summarized by the following Confusion Matrix

	predicted class h(x)			
actual class f(x)	dog	mouse	chicken	
dog	10	4	0	14
mouse	3	8	1	12
chicken	0	2	10	12

Metrics for Performance Evaluation Confusion Matrix

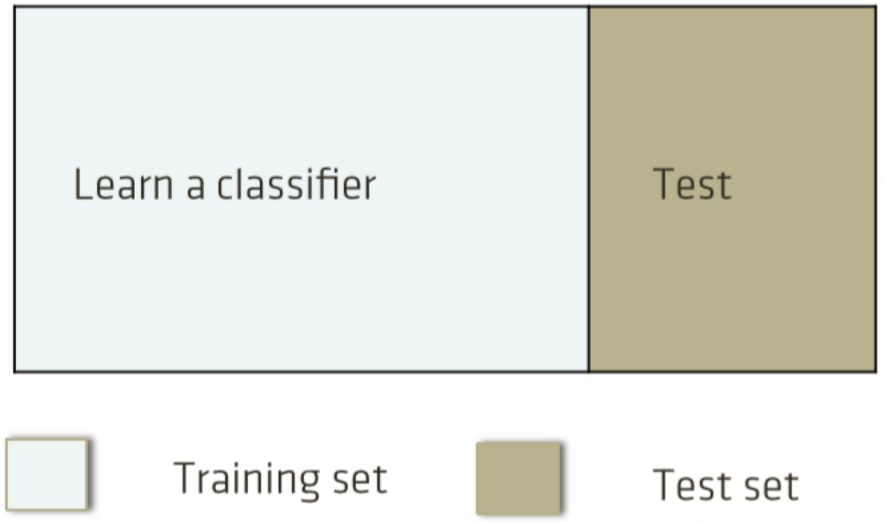
- A binary confusion matrix for a class, is a table with 2 rows and 2 columns that reports the number of false positives, false negatives, true positives, and true negatives.

CM for class cane	Predicted classes	
Actual classes	dog	Other animals
dog	TP=10	FN=4
Other animals	FP=3	TN=21

Testing the classifier
Holdout

- We split the data set into a training set and a test set – as in the previous examples
- We use the former to learn a hypothesis, and the latter to test its generalization capability

Compute performance measures over the test set



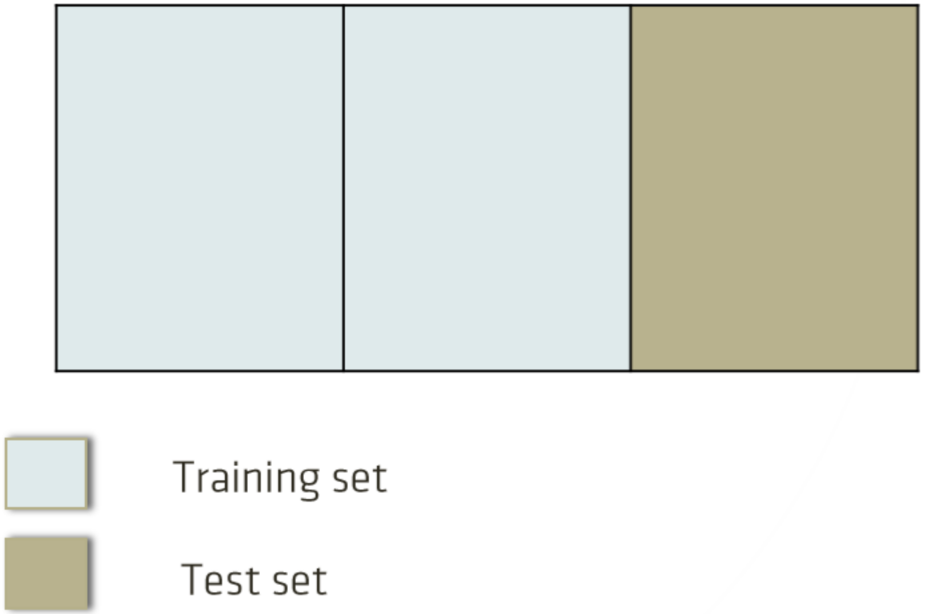
Methods for Performance Cross validation

- K-fold-cross-validation :
  - the data set is segmented into k equal-sized partitions.
  - During each run, one of the partitions is chosen for testing, and the remaining k-1 for training.
  - The procedure is repeated k times
  - The average (over the k folds) performance measures are finally given

Methods for Performance Cross validation

- 3-fold Cross Validation:
  - 2 folds for learning
  - 1 fold for testing

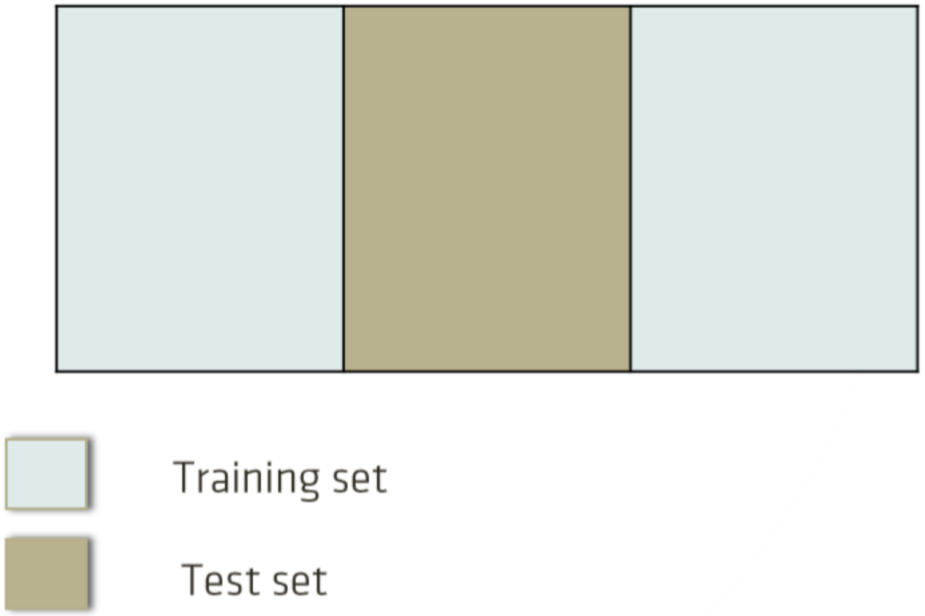
Compute performance measures over the test set, e.g., Pr1, Re1, F1



Methods for Performance Cross validation

- 3-fold Cross Validation:
  - 2 folds for learning
  - 1 fold for testing

Compute performance measures over the test set, e.g., Pr2, Re2, F2

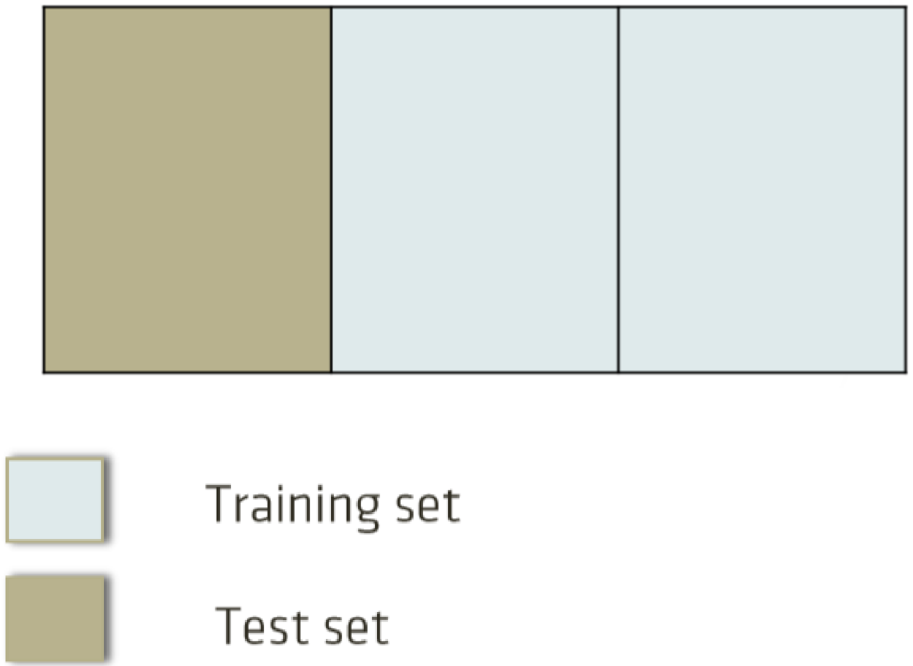


Methods for Performance Cross validation

- 3-fold Cross Validation:
  - 2 folds for learning
  - 1 fold for testing

Compute performance measures over the test set, e.g., Pr3, Re3, F3

Finally compute average performance measures



Holdout vs Cross Validation

Performance assessment based on holdout may be affected by the choice of the training and test data

Cross validation exploits all the available data, so more reliable performance measures are obtained

Quality of a Classifier

- Predictive capability: accuracy, precision, recall, …
- Descriptive capability:
  - Interpretability of the model
  - Decreases with the size of the classifier (e.g., number of rules)

The Simplicity Bias of Occam's Razor

- Entities should not be multiplied without necessity
- Given two models with the same generalization capability, the simpler one should be preferred as simplicity is desirable by itself
- Its easier to work with simple hypotheses than with complex ones
- Simple hypotheses deals better with the overfitting problem