Ciro Donalek (Caltech)

Classification

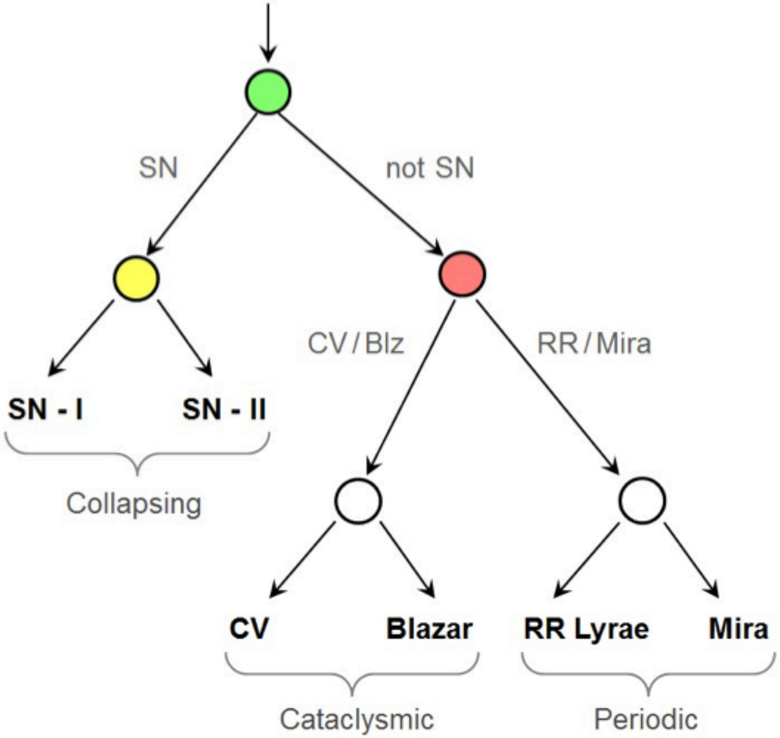
JPL Jet Propulsion Laboratory
Keck Institute for Space Studies
Sponsored by the Jet Propulsion Laboratory and by the Keck Institute for Space Studies, California Institute of Technology
Caltech

Outline

- Supervised Learning: recap
- Classification
- Accuracy and Error Measures
- Evaluation
- Challenges

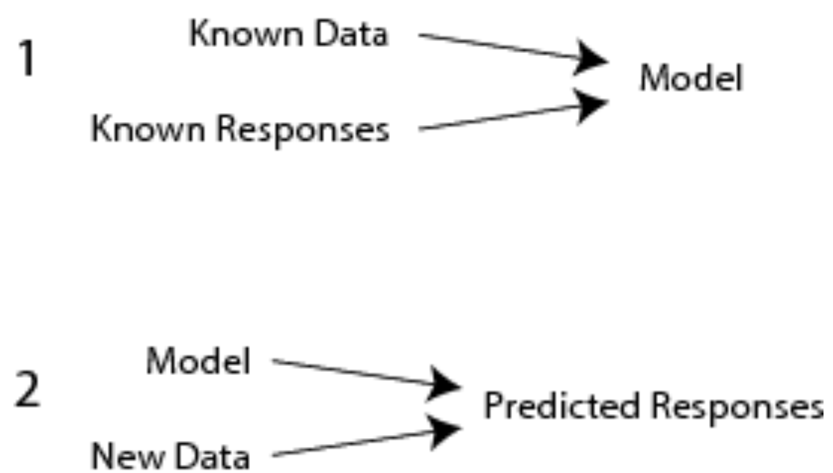
Classification

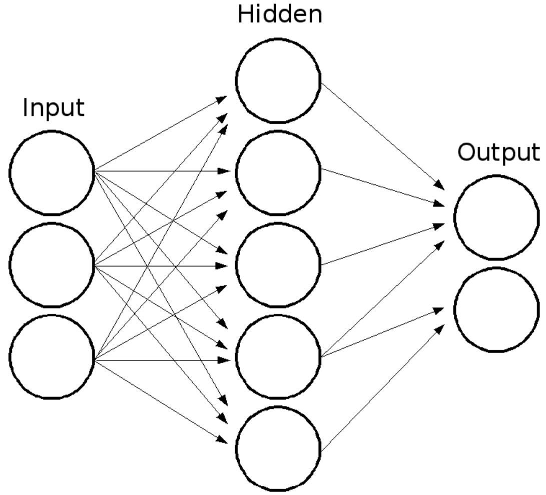
- Assign samples into categories (classes) based on a predictable attribute.
- The goal of classification is to accurately predict the target class for each case in the data set.
- Supervised Learning.



Recap: Supervised Learning

- For some examples the correct results (targets) are known and are given in input to the model during the learning process.
- Generalization: ability of a learning machine to perform accurately on new, unseen examples.





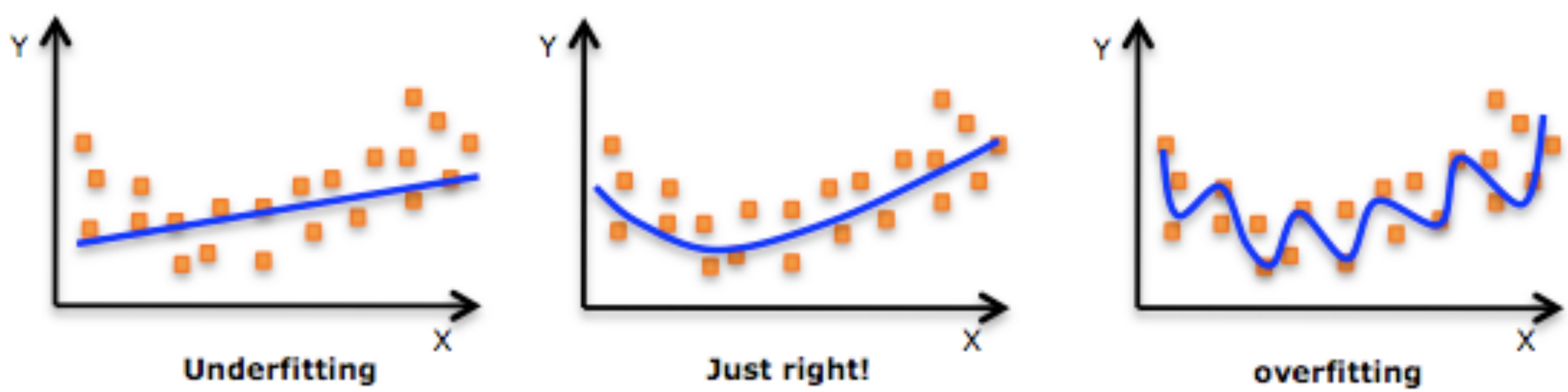
Images credits: Mathworks

Recap: datasets for learning

- Representative of the underlying model.
- Split the data in three independent data sets:
  - training set;
  - validation set;
  - test set.

Recap: Cross-Validation

- C-V techniques are used for assessing how the results of a statistical analysis will generalize to an independent data set.
- Exhaustive Cross-Validation
  - leave one out cross validation (LOOCV)
  - leave p-out cross validation
- Non-exhaustive Cross-Validation
  - *k*-fold cross validation
  - repeated random sub-sampling validation
- Choose also according to your model/task.

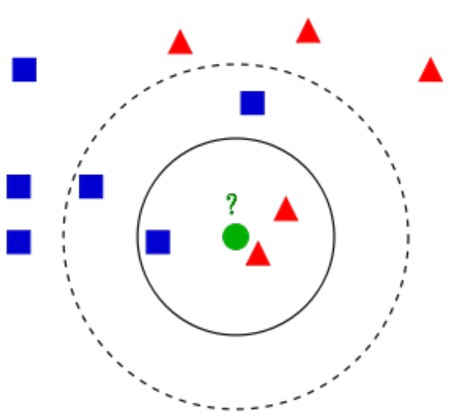


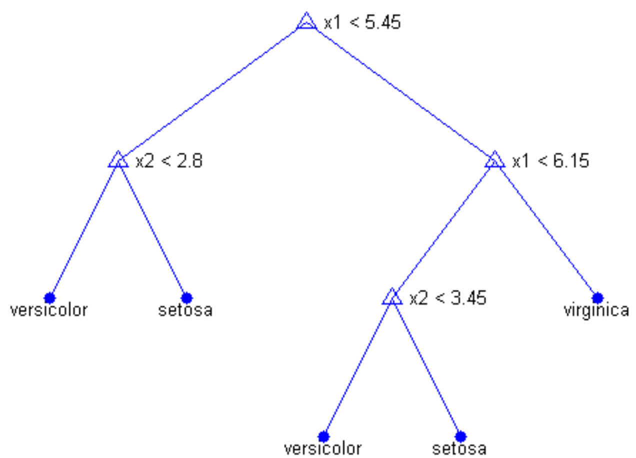
A two step process: model construction

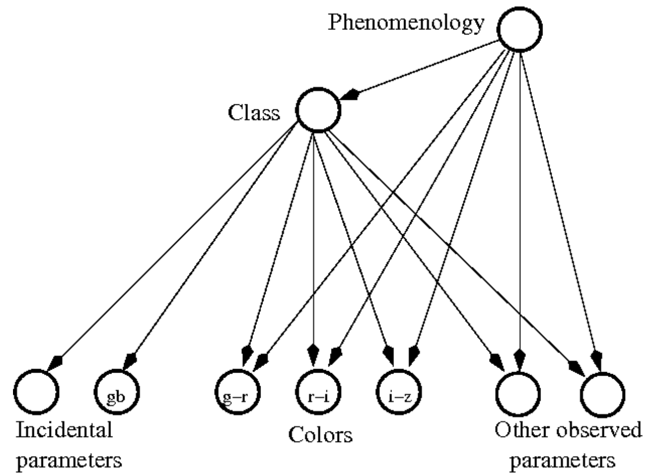
- Model construction:
  - each sample is assumed to belong to a predefined class, according to its label;
  - use training and validation set for learning;
  - model represented as classification rules, decision trees, or mathematical formulae.

A two step process: model usage

- Model usage: classifying future or unknown objects
  - estimate accuracy;
    - use an independent test set;
    - eg, accuracy rate: percentage of test samples that are correctly classified;
    - if the accuracy is acceptable, use the model to classify data whose labels are not known.
  - Output: crispy or probabilistic.







Crispy vs Probabilistic

- **Crispy classification**
  - given an input, the classifier returns its label

- **Probabilistic classification**
  - given an input, the classifier returns its probabilities to belong to each class;
  - useful when some mistakes can be more costly than others;
  - allow thresholds (e.g., give me only data >90%)
  - winner take all and other rules
    - assign the object to the class with the highest probability (WTA)
    - …but only if its probability is greater than 40% (WTA with thresholds)

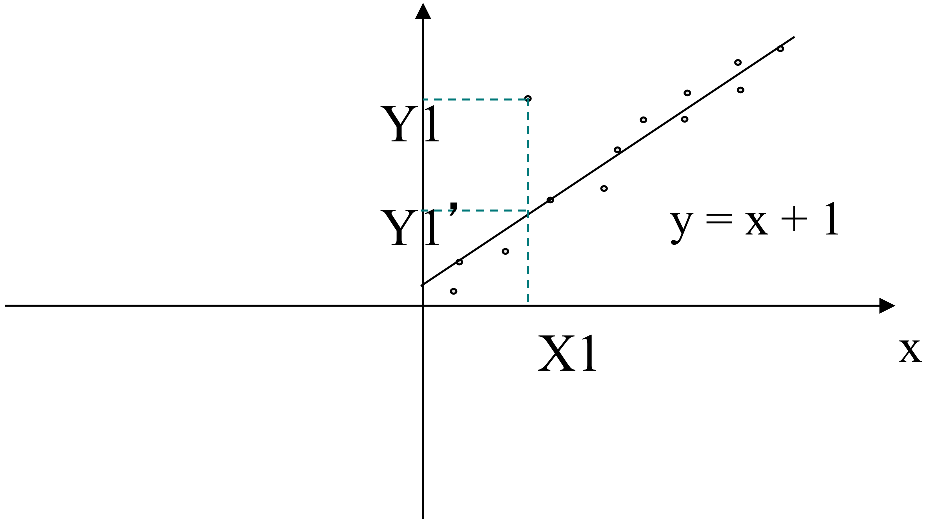
Classifiers evaluation

- Accuracy
  - ability of correctly predicti class labels.
- Speed
  - training time, classification time.
- Scalability
  - classifying data sets with millions of examples and hundreds of attributes with reasonable speed.
- Robustness
  - ability of handling missing data, noise, etc.
- Interpretability

Preparing the data

- Pre-processing steps.
- Data cleaning
  - – remove or reduce noise;
  - – treatment of missing values.

❑ Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
* Partition into equal-frequency (equi-depth) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
* Smoothing by bin means:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
* Smoothing by bin boundaries:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

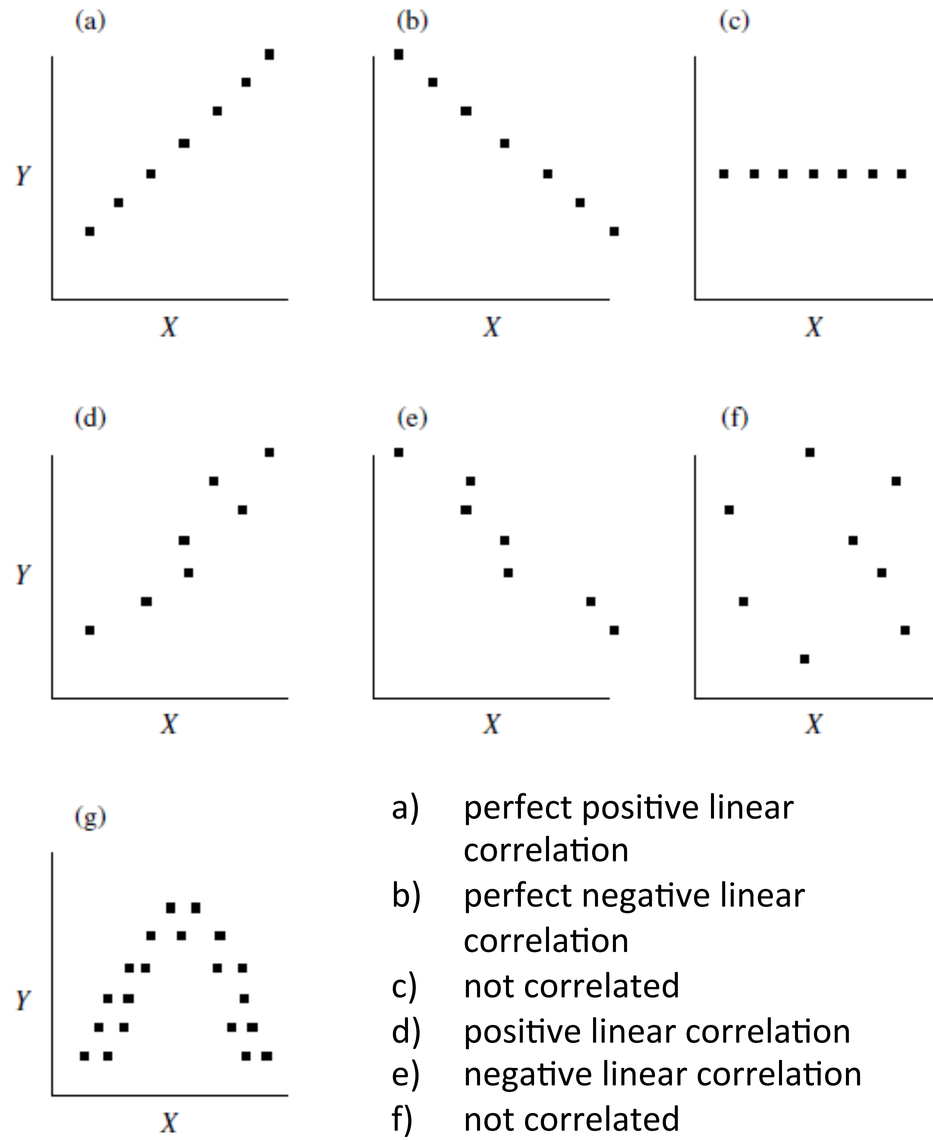


Preparing the data

- Relevance analysis
  - remove redundant and irrelevant attributes;
  - correlation analysis can be used to examine whether two variables changes together in a consistent manner.

Preparing the data

- Relevance analysis
  - remove redundant and irrelevant attributes;
  - correlation analysis can be used to examine whether two variables changes together in a consistent manner.
    - Pearson coefficient for linear correlation;
  - feature selection.



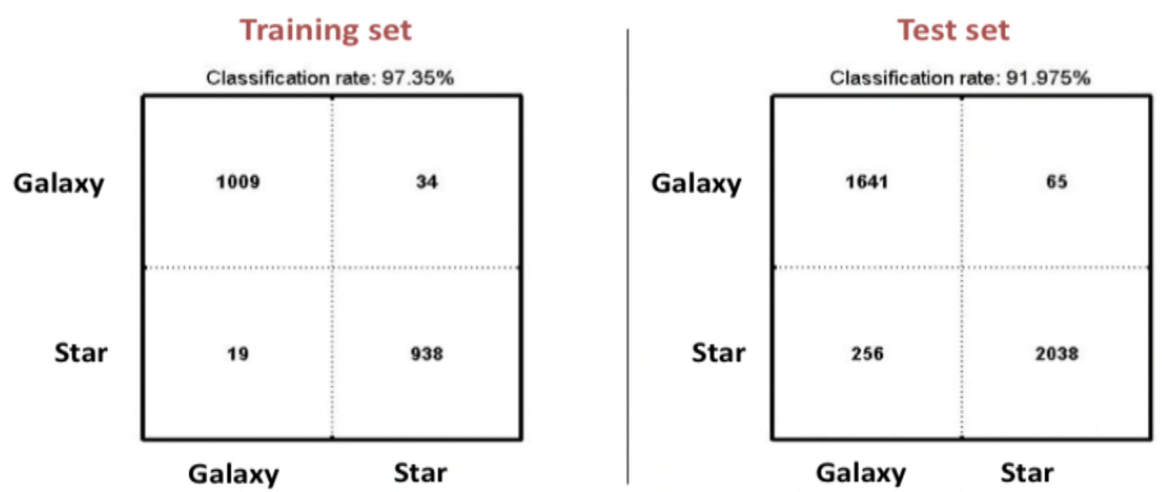
a) perfect positive linear correlation
b) perfect negative linear correlation
c) not correlated
d) positive linear correlation
e) negative linear correlation
f) not correlated
g) non-linear correlation

Image credits: cnfolio

Accuracy Measures - 1

- **Classification Rate M**: the overall percentage of objects correctly classified.
- **Error rate** (misclassification rate, loss): 1-M

In the confusion matrix the network prediction Y are compared with the target T: the rows represent the true classes and the columns the predicted classes.



Accuracy Measures - 1

- **Classification Rate M**: the overall percentage of objects correctly classified.
- **Error rate** (misclassification rate, loss): 1-M

In the confusion matrix the network prediction Y are compared with the target T: the rows represent the true classes and the columns the predicted classes.

Training set

Classification rate: 97.35%

	Galaxy	Star
Galaxy	1009	34
Star	19	938

Test set

Classification rate: 91.975%

	Galaxy	Star
Galaxy	1641	65
Star	256	2038

Confusion Matrix: example

- Confusion matrices for training, testing, and validation, and the three kinds of data combined.
- Model outputs: accurate
  - high numbers of correct responses in the green squares
  - low numbers of incorrect responses in the red squares.
  - the lower right blue squares illustrate the overall accuracies.

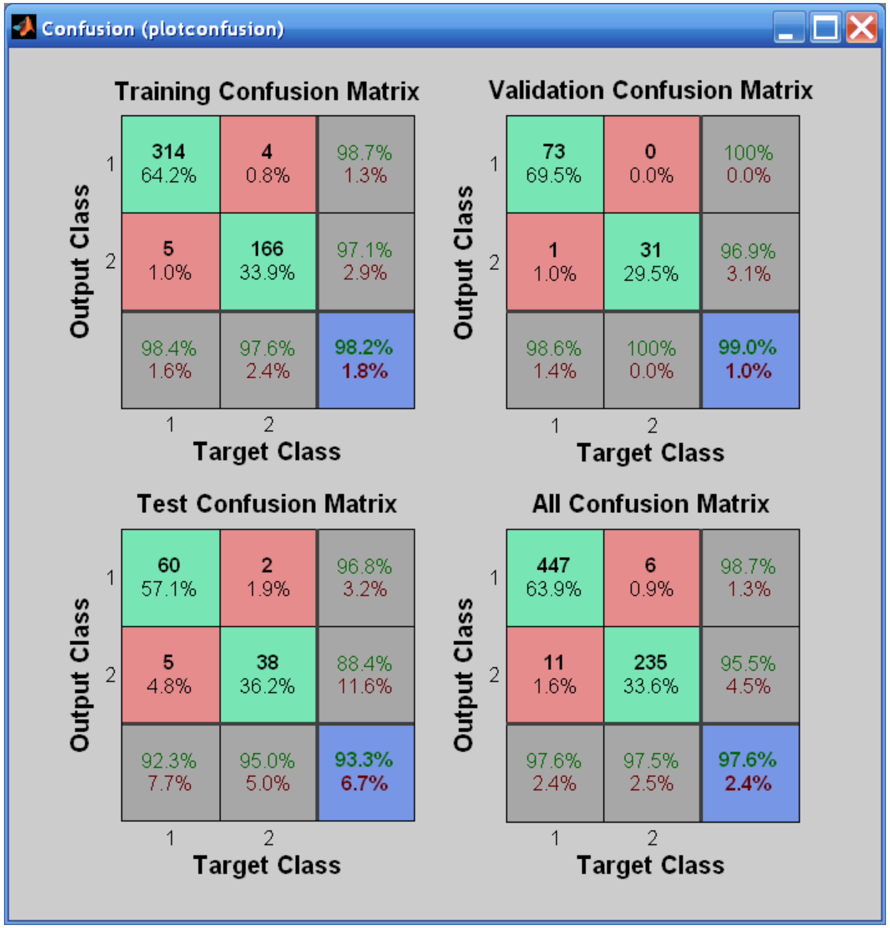
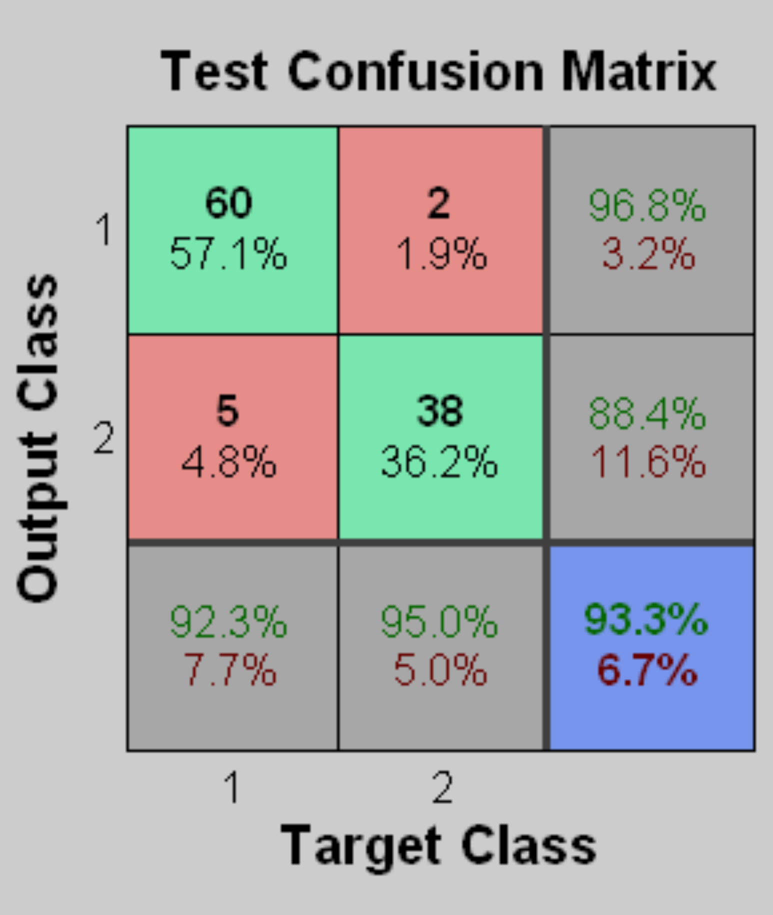


Image credits: Mathworks

Completeness and Contamination

- **Completeness**: the percentage of objects of a given class correctly classified as such. (ex, class 1: 96.8% compl.);
- **Contamination**: for each class, the percentage of objects of other classes incorrectly classified as objects belonging to that class (ex, class 1: 7.7% cont.)
- **Precision**: 1-Contamination



Confusion Matrix

	SNIa	SNIb	SNIc	SNIIn	SNIIp
SNIa	**845**	2	1	8	23
SNIb	31	**18**	2	1	3
SNIc	16	1	**15**	5	8
SNIIn	12	0	1	**64**	9
SNIIp	34	3	3	7	**235**

Binary Classifiers

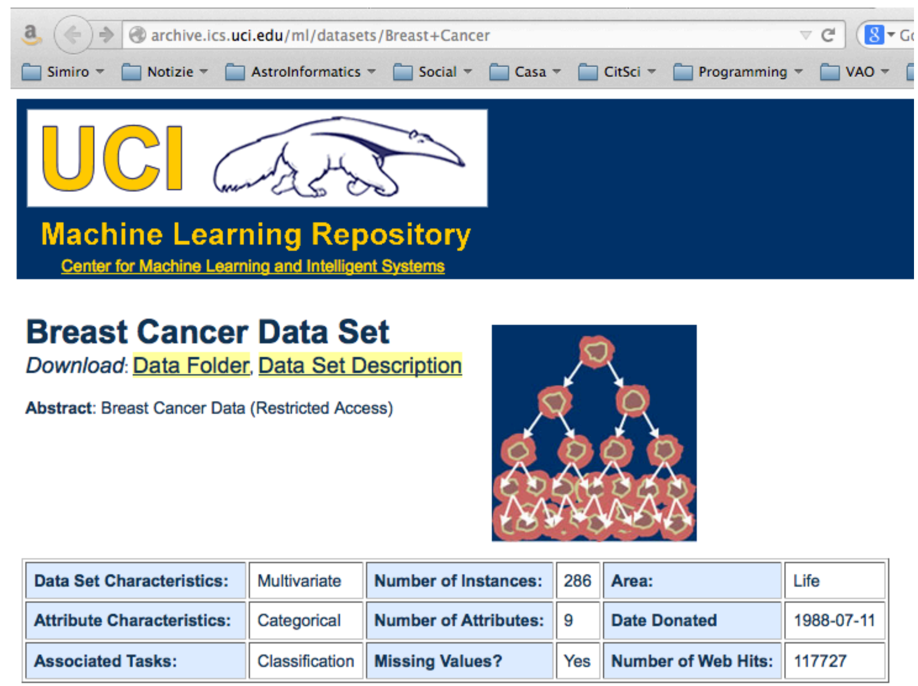
- Two classes problem: negative vs. positive (e.g., cancer)

$\downarrow$ **actual** \ **predicted** $\rightarrow$	**negative**	**positive**
negative	*TN*	*FP*
positive	*FN*	*TP*

- Is an accuracy rate of 90% acceptable?
  - accuracy rate defined as the percentage of objects correctly classified.

Example: cancer diagnosis

- Is an accuracy rate of 90% acceptable?
- **NOT NECESSARILY!** Supposing only 3-4% of training set is labeled as cancer, a classifier that classify all the elements as “not cancer” would have 96% classification rate!
- The cost associated with a false negative may be far greater than that of a false positives.
  – eg, incorrectly classifying a cancerous patient as not cancerous.



Example: cancer diagnosis

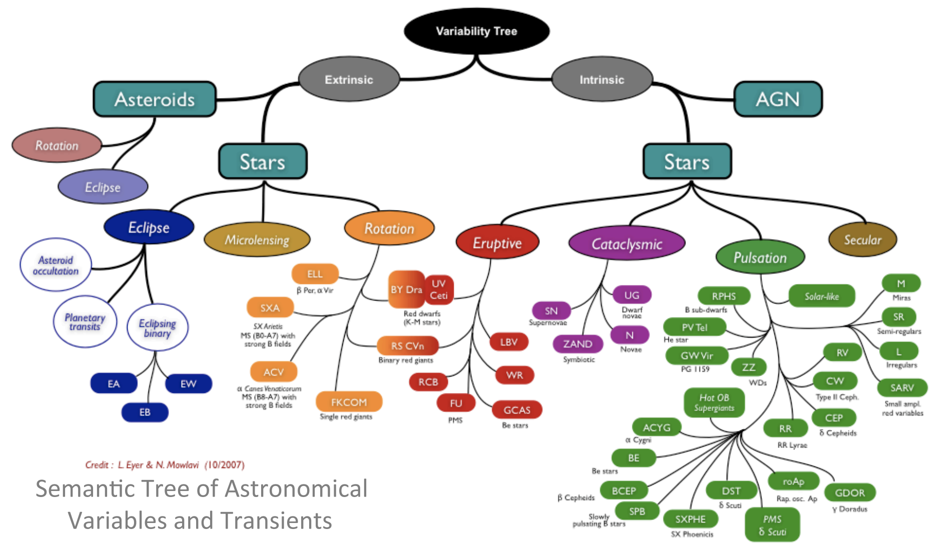
- Measures useful to assess costs and benefits associated with a classification model.
- Sensitivity=*true_positives/total_#_of_positives*
- Specificity=*true_negatives/total_#_of_negatives*
- Precision= *t_pos / (t_pos + f_pos)*
  - e.g., percentage of samples labeled as cancer that are actually cancer.

↓ actual \ predicted →	negative	positive
negative	*TN*	*FP*
positive	*FN*	*TP*

Sensitivity: TP / (FN+TP) // Type II error
Specificity: TN / (TN+FP) // Type I error
Precision: TP / (TP+FP)
Accuracy: (TN+TP) / (TN+FP+FN+TP)

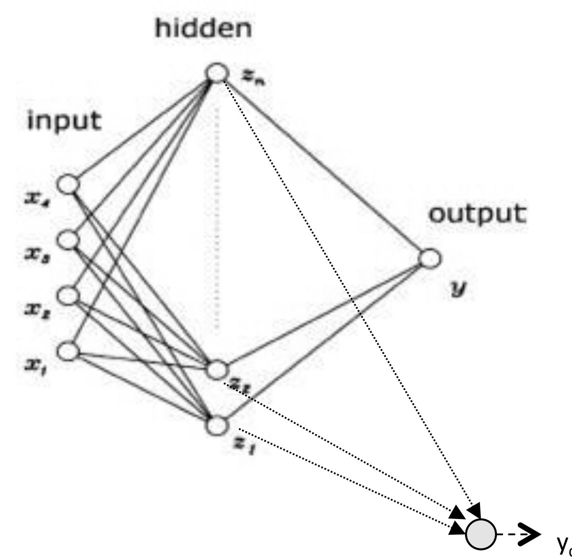
Classification Challenges

- Massive multiparametric dataset
  - Petascale ready
  - Sparse Data
  - Heterogeneous Data
- Classification
  - Real Time
  - Reliable
  - High completeness
  - Low contamination
  - Use minimum amount of points
  - *Learn* from the past experience
  - As automated as possible
- Include External Knowledge



Credit : L.Eyer & N. Mowlavi (10/2007)

Semantic Tree of Astronomical Variables and Transients



[redacted]