Introduction to Machine Learning Applications

Spring 2021

Lecture-9

Lydia Manikonda

manikl@rpi.edu

Rensselaer

Today’s agenda

- Overview of Modeling
- Accuracy metrics

Announcements

- Homework-3 due tonight 11:59 pm ET via LMS

Modeling

What is a model?

- Mathematical representation of a real-world process.
- In other words, description of a system using mathematical concepts.

- Three different types of models can be built:
    - Supervised learning
    - Unsupervised learning
    - Semi-supervised learning

Definition of Classification

Given a collection of records (training set)

- Each record is by characterized by a tuple ($\boldsymbol{x}$,y), where $\boldsymbol{x}$ is the attribute set and y is the class label

  # $\boldsymbol{x}$: attribute, predictor, independent variable, input

  # y: class, response, dependent variable, output

Task:

- Learn a model that maps each attribute set $\boldsymbol{x}$ into one of the predefined class labels y

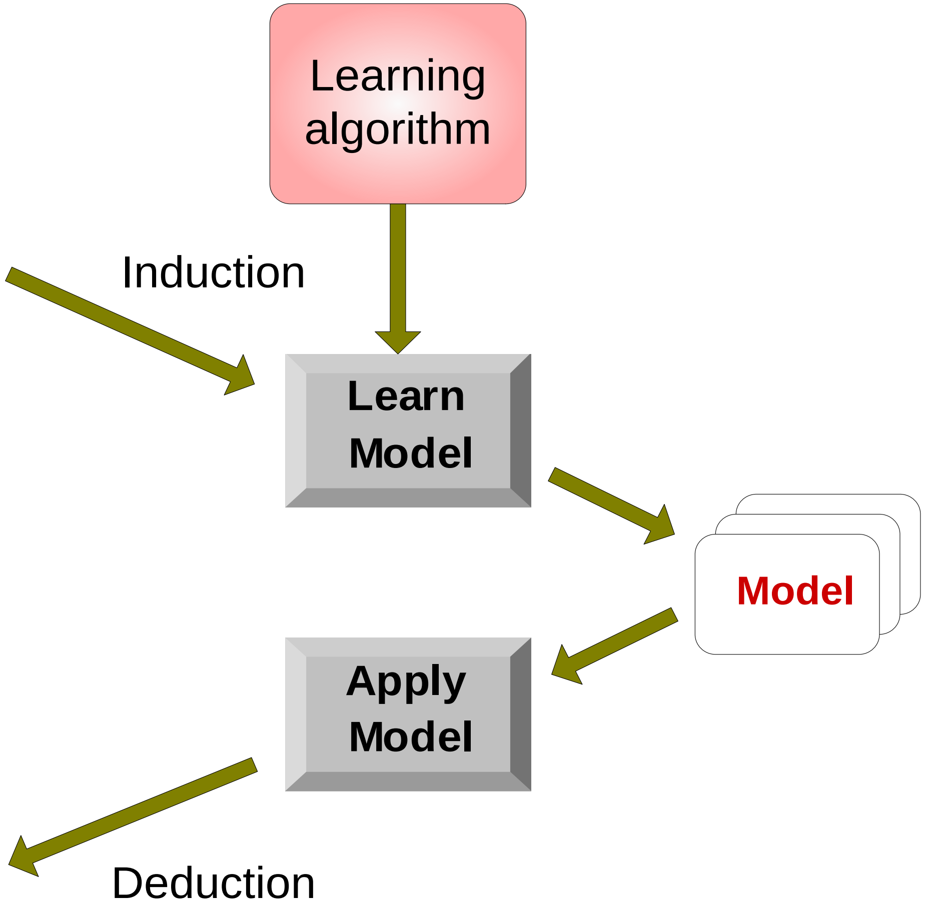
Example -- Classification tasks

Task	Attribute set, ***x***	Class label, *y*
Categorizing email messages	Features extracted from email message header and content	spam or non-spam
Identifying tumor cells	Features extracted from MRI scans	malignant or benign cells
Cataloging galaxies	Features extracted from telescope images	Elliptical, spiral, or irregular-shaped galaxies

Classification model

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	**No**
2	No	Medium	100K	**No**
3	No	Small	70K	**No**
4	Yes	Medium	120K	**No**
5	No	Large	95K	**Yes**
6	No	Medium	60K	**No**
7	Yes	Large	220K	**No**
8	No	Small	85K	**Yes**
9	No	Medium	75K	**No**
10	No	Small	90K	**Yes**

Training Set



Learning algorithm

Induction

Learn Model

Model

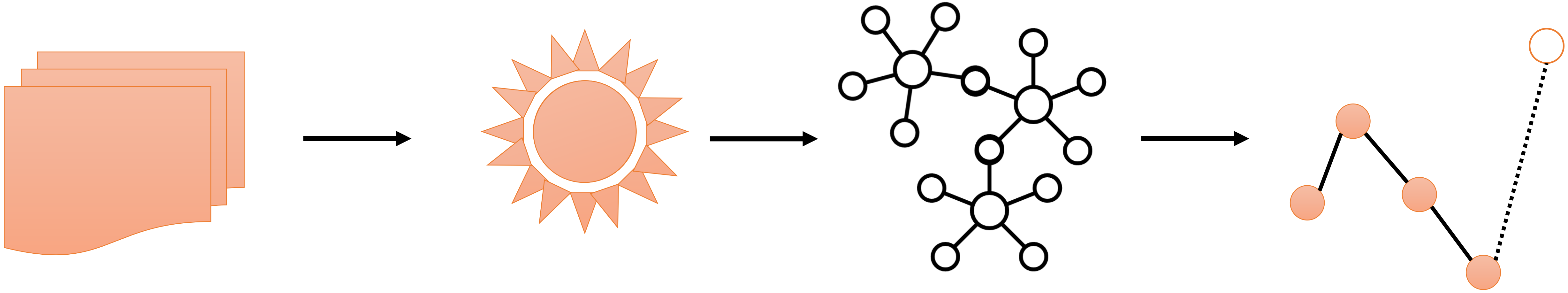
Apply Model

Deduction

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	**?**
12	Yes	Medium	80K	**?**
13	Yes	Large	110K	**?**
14	No	Small	95K	**?**
15	No	Large	67K	**?**

Test Set

A standard learning pipeline



DATA **Training** **Model** **Prediction**

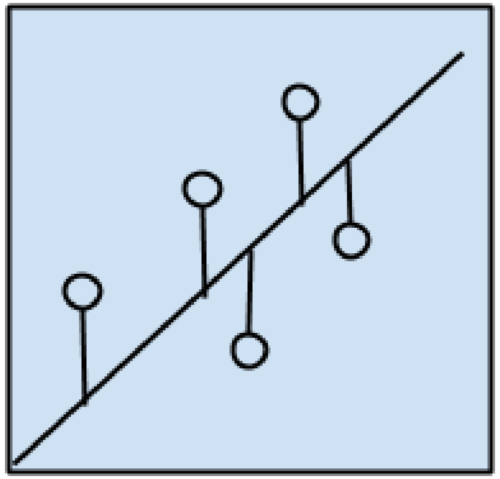
Classification Techniques

- Base Classifiers
  - Decision Tree based Methods
  - Rule-based Methods
  - Nearest-neighbor
  - Neural Networks
  - Deep Learning
  - Naïve Bayes and Bayesian Belief Networks
  - Support Vector Machines

- Ensemble Classifiers
  - Boosting, Bagging, Random Forests

Regression Algorithms

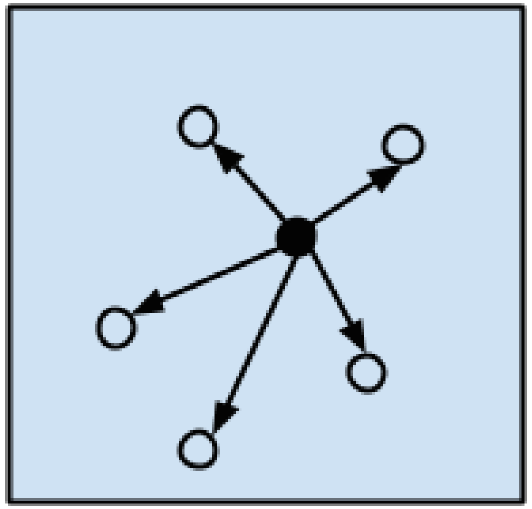
- Modeling the relationship between variables that are iteratively refined using a measure of error.
- Most popular regression algorithms are:
  - Ordinary least squares regression
  - Linear regression
  - Logistic regression
  - Multivariate adaptive regression splines
  - …



Regression Algorithms

Instance-based algorithms

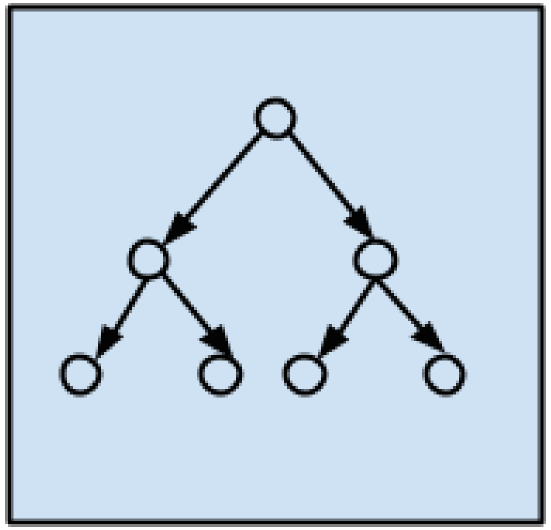
- This model is a decision problem with instances of training data that are deemed important or required to the model.
- Focus is put on the representation of the stored instances and similarity measures used between instances.
- Most popular instance-based algorithms are:
  - K-Nearest Neighbor (KNN)
  - Support Vector Machines (SVM)
  - Learning Vector Quantization
  - Self-Organizing Maps
  - …



Instance-based Algorithms

Decision Tree-based algorithms

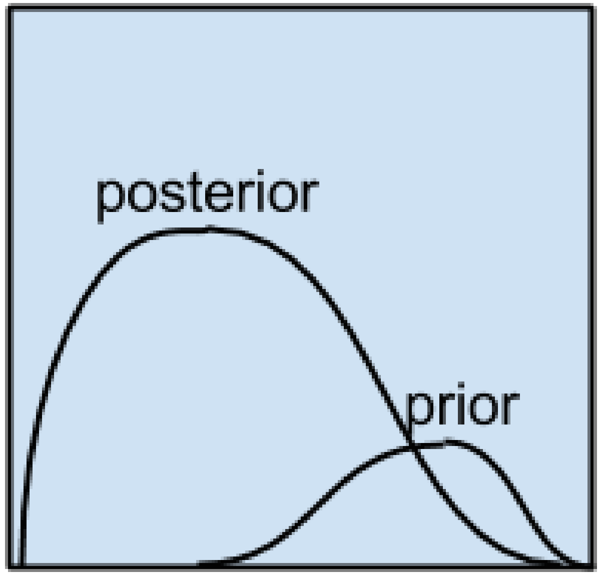
- These methods construct a model of decisions based on the actual values of attributes in the data.
- These decisions built are in the form of a tree.
- Most popular algorithms are:
  - Classification and Regression Tree
  - Conditional Decision Trees
  - ID3
  - C4.5 and C5.0
  - …



Decision Tree Algorithms

Bayesian Algorithms

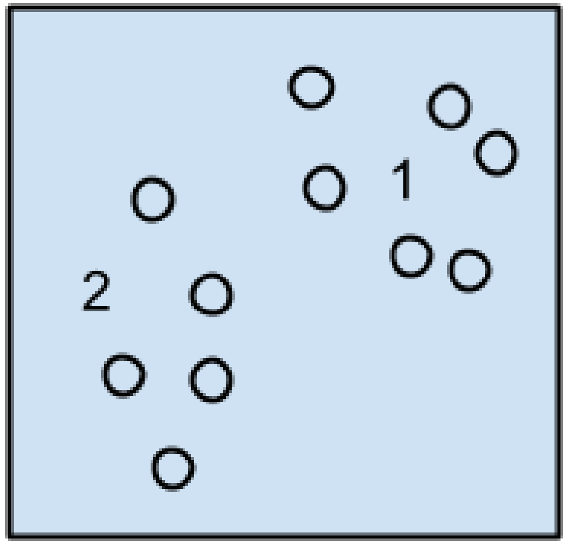
- Bayesian methods explicitly apply the Bayes Theorem for problems such as classification and regression.
- Bayes Theorem
- Most popular algorithms are:
  - Naïve Bayes
  - Gaussian Naïve bayes
  - Bayesian network
  - Bayesian belief network
  - …



Bayesian Algorithms

Clustering Algorithms

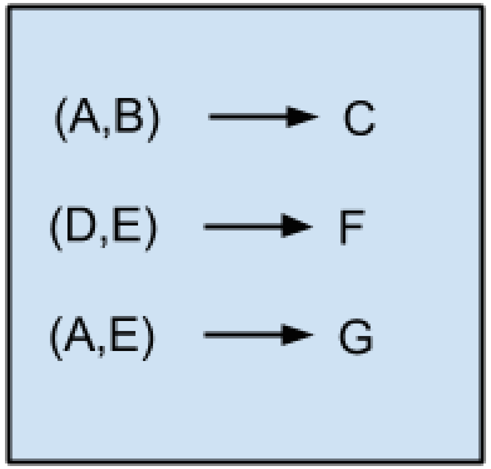
- These algorithms utilize the inherent structures in the data to organize them into various groups.
- Main goal is to find clusters that have high intra similarity and high inter similarity distances.
- Most popular clustering algorithms are:
  - K-Means
  - K-Medoids
  - Expectation Maximization
  - Hierarchical Clustering
  - …



Clustering Algorithms

Association Rule Learning Algorithms

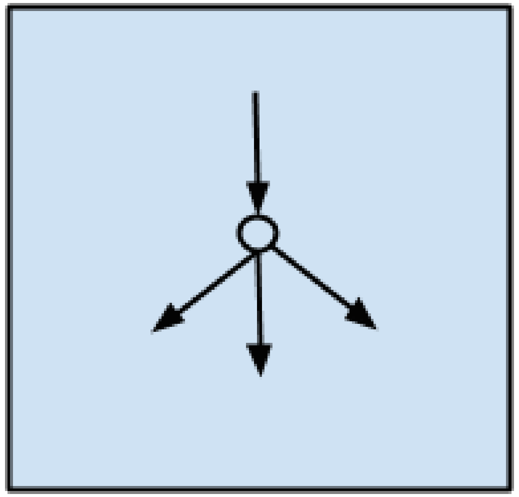
- These methods extract rules that best explain the observed relationships between variables in the data
- Most popular algorithms are:
  - Apriori
  - Eclat
  - FP-growth
  - …



Association Rule
Learning Algorithms

Artificial Neural Network Algorithms

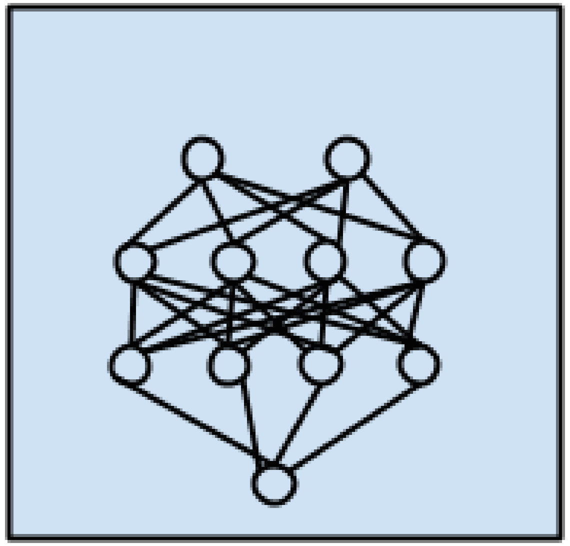
- Models that are inspired by the structure and function of biological neural networks.
- Most popular algorithms are:
  - Perceptron
  - Multilayer perceptron
  - Backpropagation
  - …



Artificial Neural Network Algorithms

Deep Learning Algorithms

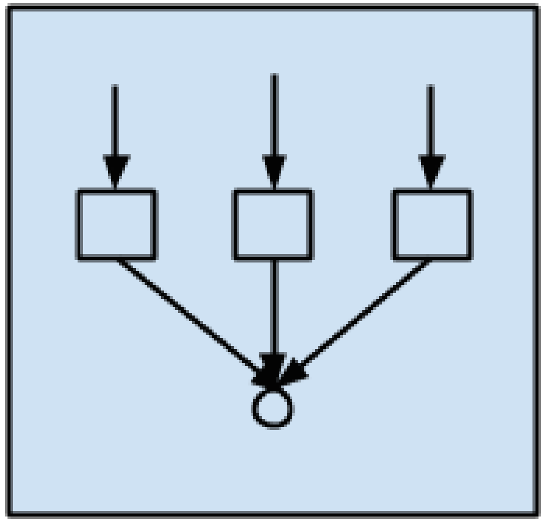
- Update to Artificial Neural Networks
- Main goal is to build a much larger and more complex neural networks.
- Most popular algorithms are:
  - Convolutional Neural Network (CNN)
  - Recurrent Neural Networks (RNNs)
  - Long Short-Term Memory Networks (LSTMs)
  - Deep Belief Networks (DBN)



Deep Learning
Algorithms

Ensemble Algorithms

- These are the models composed of multiple weaker models that are independently trained and the predictions are combined to make the overall prediction.
- Some of the popular algorithms are:
  - Boosting
  - Boostrapped Aggregation
  - AdaBoost
  - Gradient Boosting Machines
  - Random Forest
  - …



Ensemble Algorithms

How can we measure the quality of an ML model?

Confusion Matrix

- A table that is often used to describe the performance of a classification model on a set of test data.
- This allows the visualization of the algorithm’s performance.

		Actual Class	
		Class = 1	**Class = 0**
Predicted Class	**Class = 1**	f_{11}	f_{10}
	Class = 0	f_{01}	f_{00}

		Actual Class	
		Class = 1	**Class = 0**
Predicted Class	**Class = 1**	f_{11}	f_{10}
	Class = 0	f_{01}	f_{00}

- f_{11} – True Positive
- f_{10} – False Positive – Type I error
- f_{01} – False Negative – Type II error
- f_{00} – True Negative

		Actual Class	
		Class = 1	**Class = 0**
Predicted Class	**Class = 1**	f_{11}	f_{10}
	Class = 0	f_{01}	f_{00}

- f_{11} – True Positive
- f_{10} – False Positive – Type I error
- f_{01} – False Negative – Type II error
- f_{00} – True Negative

$$Accuracy = \frac{(f_{11} + f_{00})}{(f_{11} + f_{10} + f_{01} + f_{00})}$$

		Actual Class	
		Class = 1	**Class = 0**
Predicted Class	**Class = 1**	*10*	*5*
	Class = 0	*5*	*10*

Compute the value of Accuracy?

How many are truly labeled = 10+10

Total data points that you have: 10+10+5+5 = 30

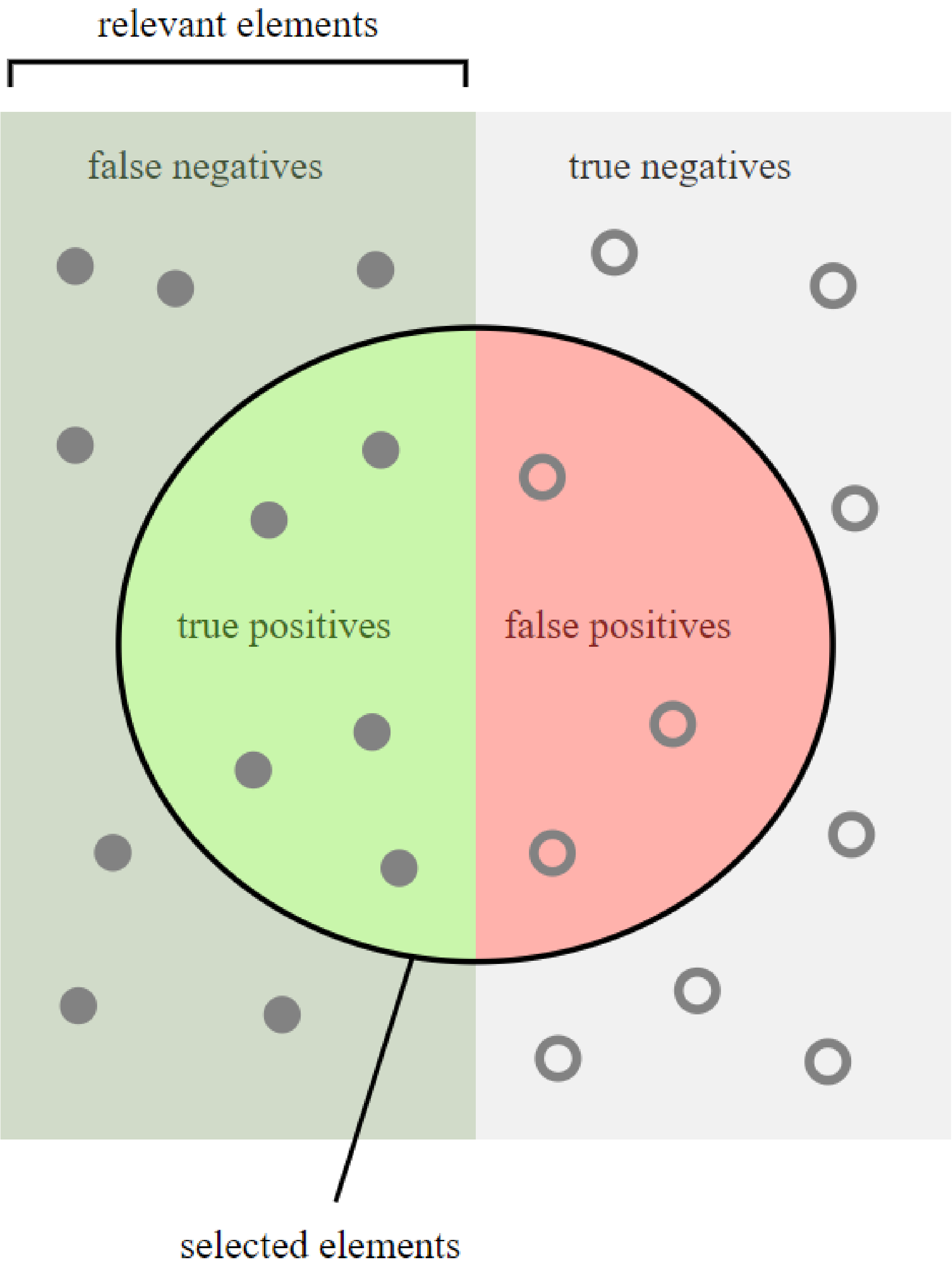
Accuracy = 20/30 = 2/3 = 66.67

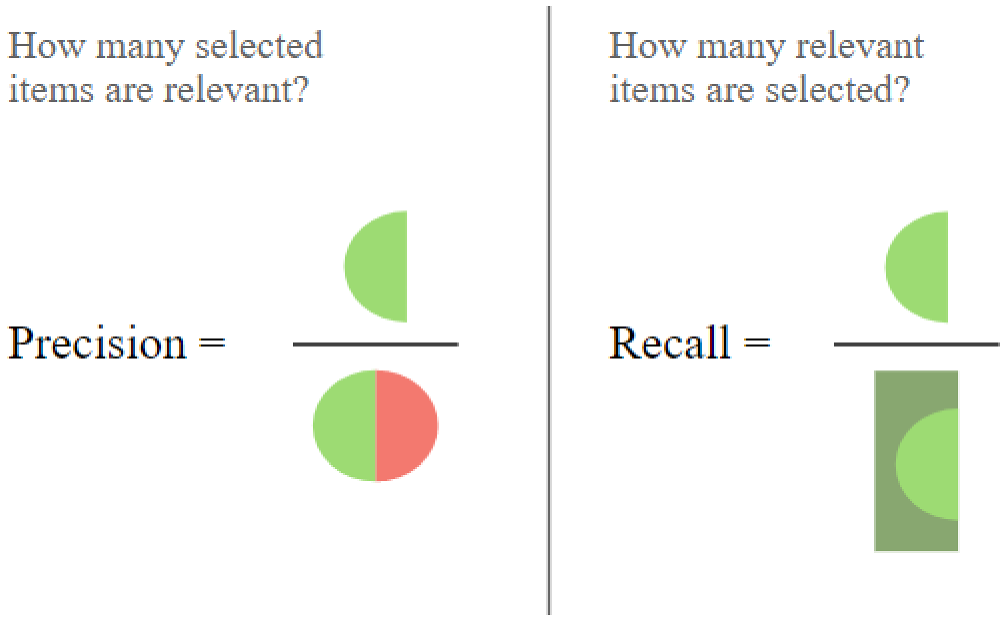
		Actual Class	
		Class = 1	**Class = 0**
Predicted Class	**Class = 1**	f_{11}	f_{10}
	Class = 0	f_{01}	f_{00}

- f_{11} – True Positive
- f_{10} – False Positive – Type I error
- f_{01} – False Negative – Type II error
- f_{00} – True Negative

Precision: How many selected items are relevant?

Recall: How many relevant items are selected?




How many selected items are relevant?
How many relevant items are selected?
Precision =
Recall =

Credits: https://en.wikipedia.org/wiki/Precision_and_recall

Precision

How many selected items are relevant?

$$Precision = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{retrieved\ documents\}|}$$

$$Precision = \frac{f_{11}}{(f_{10} + f_{11})}$$

Recall

How many relevant items are selected?

$$Recall = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{relevant\ documents\}|}$$

$$Recall = \frac{f_{11}}{(f_{01} + f_{11})}$$

F-measure

Better measure that considers the harmonic mean of *precision* and *recall*

$$f - measure = \frac{2*(precision*recall)}{(precision+recall)}$$

$$f1score = \frac{2 * precision * recall}{(precision + recall)}$$

Compute precision, recall and f-measure

tp = 8

fp = 4

fn = 2

tn = 6

		Actual Class	
		True	**False**
Predicted class	**True**	8	*4*
	False	2	6

Precision = tp/(tp+fp) = 8/(8+4) = 8/12

Recall = 8/(8+2) = 8/10

F-score = (2*precision*recall)/(precision+recall) = (2*8*8/120)/(8/10 + 8/12) =

Sensitivity

Also considered as the True positive rate or equivalent to recall

$$Sensitivity = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{relevant\ documents\}|}$$

Specificity

Also known as True Negative Rate (actual negatives that are correctly identified)

$$Specificity = \frac{True\ Negative}{(True\ Negative + False\ Positive)}$$

$$Specificity = \frac{f_{00}}{(f_{00} + f_{10})}$$

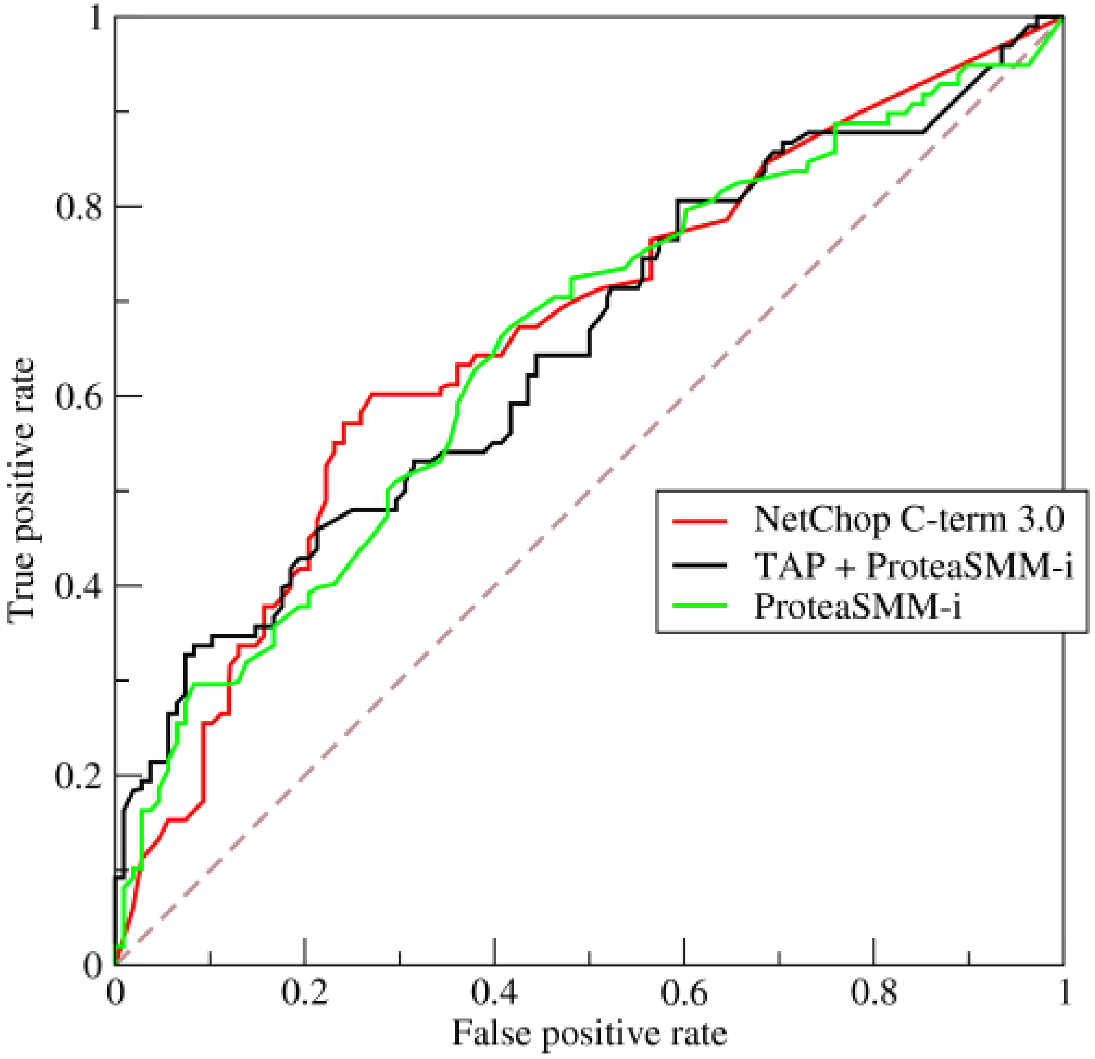
ROC curve

- Receiver operating characteristic curve
- Created by plotting True-positive rate vs False positive rate

$$FalsePositiveRate = \frac{False\ Positive}{(True\ Negative+False\ Positive)}$$

$$TruePositiveRate = \frac{True\ Positive}{(False\ Negative+True\ Positive)}$$

ROC curve – Example



https://upload.wikimedia.org/wikipedia/commons/6/6b/Roccurves.png

Class Exercises

		Actual Class	
		Cat (true)	**~Cat** (false)
Predicted Class	**Cat** (true)	5	*2*
	~Cat (false)	3	3

- Compute
  - Accuracy
  - Precision,
  - Recall,
  - f-measure,
  - Specificity
  - False positive rate

Accuracy = (5+3)/13

Tp=5

Fp = 2

Fn = 3

tn = 3

Precision = (5/5+2)

Recall = 5/(5+3)

F-measure = 2*p*r/(p+r)

$$\text{Accuracy} = \frac{TP + TN}{\#\text{Total}} = \frac{5 + 3}{5 + 2 + 3 + 3} = \frac{8}{13}$$

$$= 61.53\%$$

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{5}{5 + 2} = \frac{5}{7} = 71.42\%$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{5}{5 + 3} = \frac{5}{8} = 62.5\%$$

$$\text{fscore} = \frac{2 \times \text{Pre} \times \text{Rec}}{(\text{Pre} + \text{Rec})} = \frac{2 \times \frac{5}{7} \times \frac{5}{8}}{\left(\frac{5}{7} + \frac{5}{8}\right)} =$$

$$= \left(\frac{25}{28}\right) \Big/ \left(5 \times \left[\frac{15}{56}\right]\right)$$

$$= \frac{\cancel{25}}{\cancel{28}} * \frac{\cancel{56}\ 2}{\cancel{5} \times \cancel{15}\ 3} = \frac{2}{3} = 66.7\%$$

$$\text{Specificity} = \frac{TN}{TN + FP} = \frac{3}{3 + 2} = \frac{3}{5} = 60\%$$

$$\text{FPR} = \frac{FP}{FP + TN} = \frac{2}{2 + 3} = \frac{2}{5} = 40\%$$

k-fold Cross-validation

- Resampling procedure to evaluate machine learning models on a given data sample.
- The parameter *k* refers to the number of groups that a given data sample is to be split into.
- If k=10, it is 10-fold cross-validation where the sample data is divided into 10 groups.

k-fold Cross-validation

> Shuffle the dataset (better)

> Split the dataset into *k* disjoint groups

> For each unique group:

  > Take the group as a hold out or test (validation) data set

  > Take the remaining groups as a training data set

  > Fit a model on the training set and evaluate it on the test set

  > Record the evaluation score

> Find the mean of all the sample of model evaluation scores

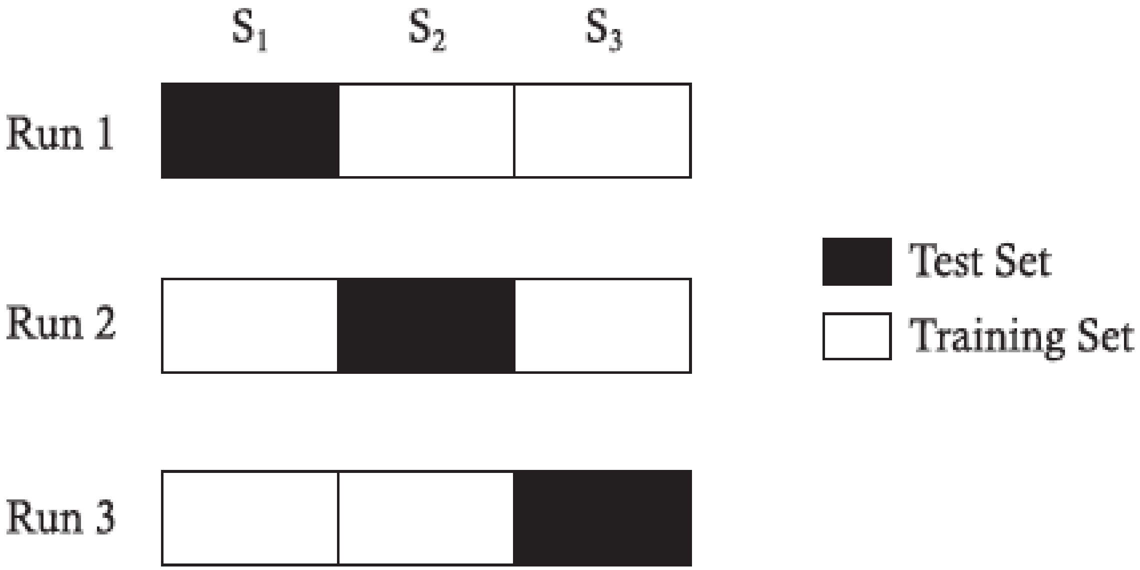
k-fold Cross-validation

[1, 2, 3, 4, 5, 6]

Fold1: [5, 3]

Fold2: [1, 6]

Fold3: [2,4]



Model1: Trained on Fold2 + Fold3, Tested on Fold1

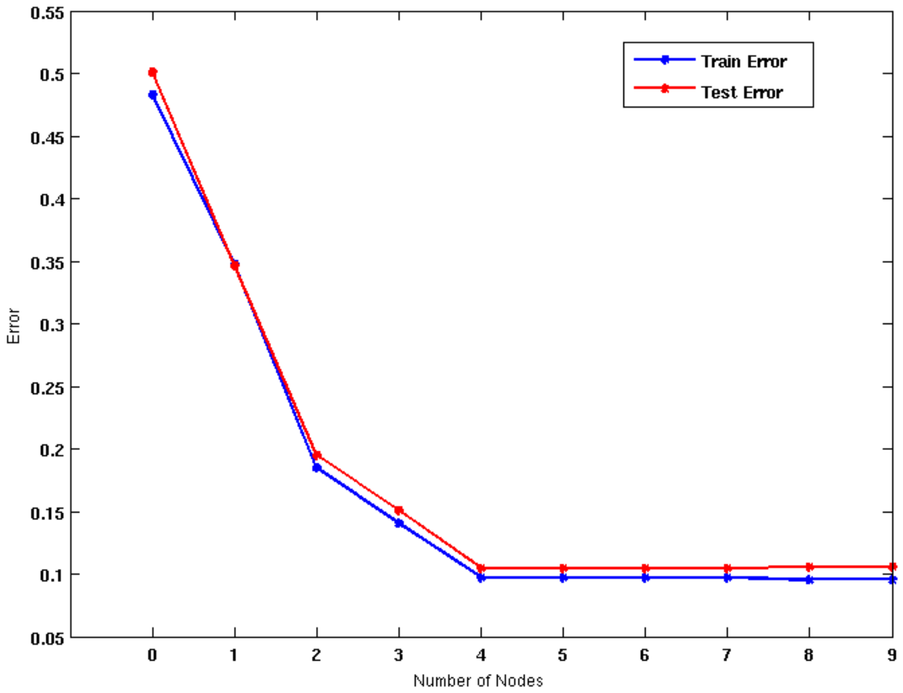
Model2: Trained on Fold1 + Fold3, Tested on Fold2

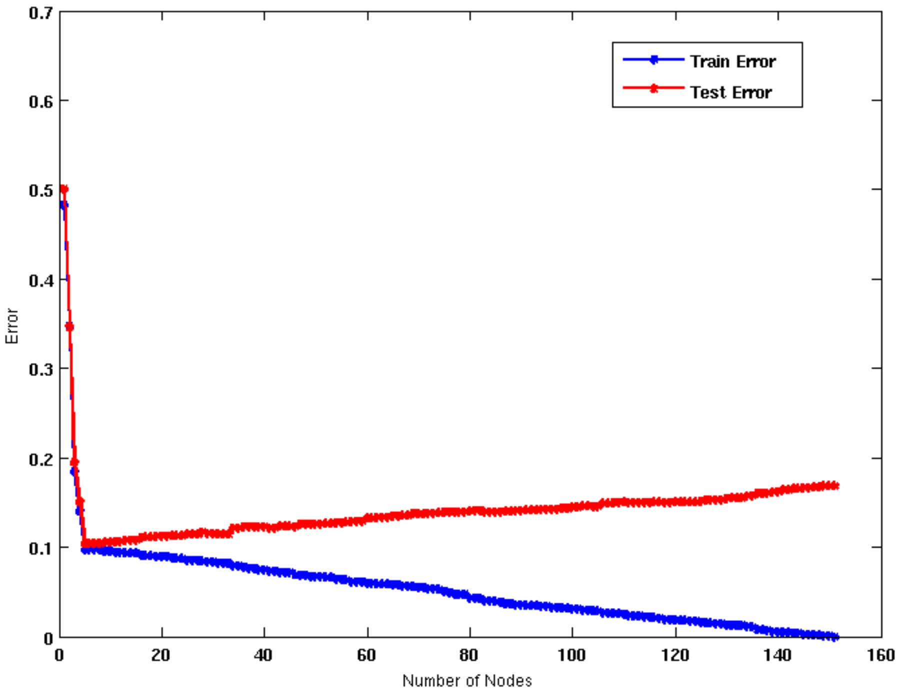
Model3: Trained on Fold1 + Fold2, Tested on Fold3

Example

- Given a set of data points – {a, b, c, d, e, f, g, h}
  - Perform 4-fold cross validation
  - Explain it in your own terms – what are the folds and how do you use them?

Model Overfitting & Underfitting





Underfitting: when model is too simple, both training and test errors are large

Overfitting: when model is too complex, training error is small but test error is large

How do we convert ground truth data and predictions to a confusion matrix?

- Notebook example

- Exercises in the notebook to follow..