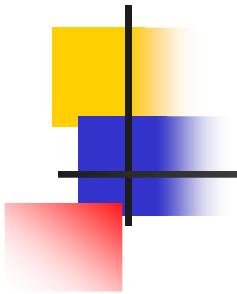
Machine Learning Approaches in Bioinformatics and Computational Biology

Byron Olson

Center for Computational Intelligence, Learning, and Discovery



Machine Learning

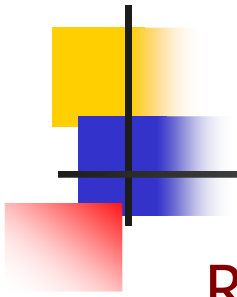
- Background and Motivation
- What is learning?
- What is machine learning?
- How can we specify a learning problem?
- Taxonomy of learning algorithms
- Representative applications in bioinformatics and computational biology

Motivation

- Simply by being here, you've shown you're an example of the amazing computational ability of humans.
- There are many tasks easily accomplished by animals that are difficult to reproduce using computers
- Motivated by understanding and modeling this ability to learn, machine learning researchers develop algorithms to tackle difficult problems

Learning Defined

Learning is a process by which the learner improves his performance on a task or a set of tasks as a result of experience within some environment



Types of learning

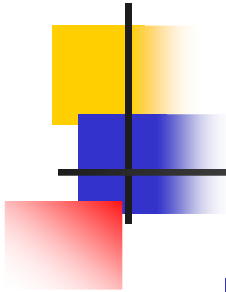
Rote Learning – useful when it is less expensive to store and retrieve some information than to compute it

Learning from Instruction – transform instructions into operationally useful knowledge

Learning form Examples (and counter-examples) – extract predictive or descriptive regularities from data

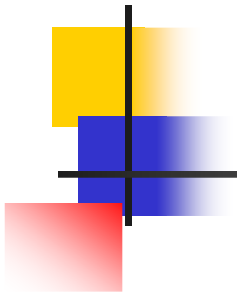
Learning from Deduction (and explanation) – generalize instances of deductive problem-solving

Learning from Exploration – learn to choose actions that maximize reward



Machine Learning Defined

- **Machine learning** is an area of artificial intelligence concerned with the development of techniques which allow computers to "learn". More specifically, machine learning is a method for creating computer programs by the analysis of data sets. (empirical approach)
  - From wikipedia



Machine Learning: Contributing Disciplines

Computer Science – Artificial Intelligence, Algorithms and Complexity, Databases, Data Mining

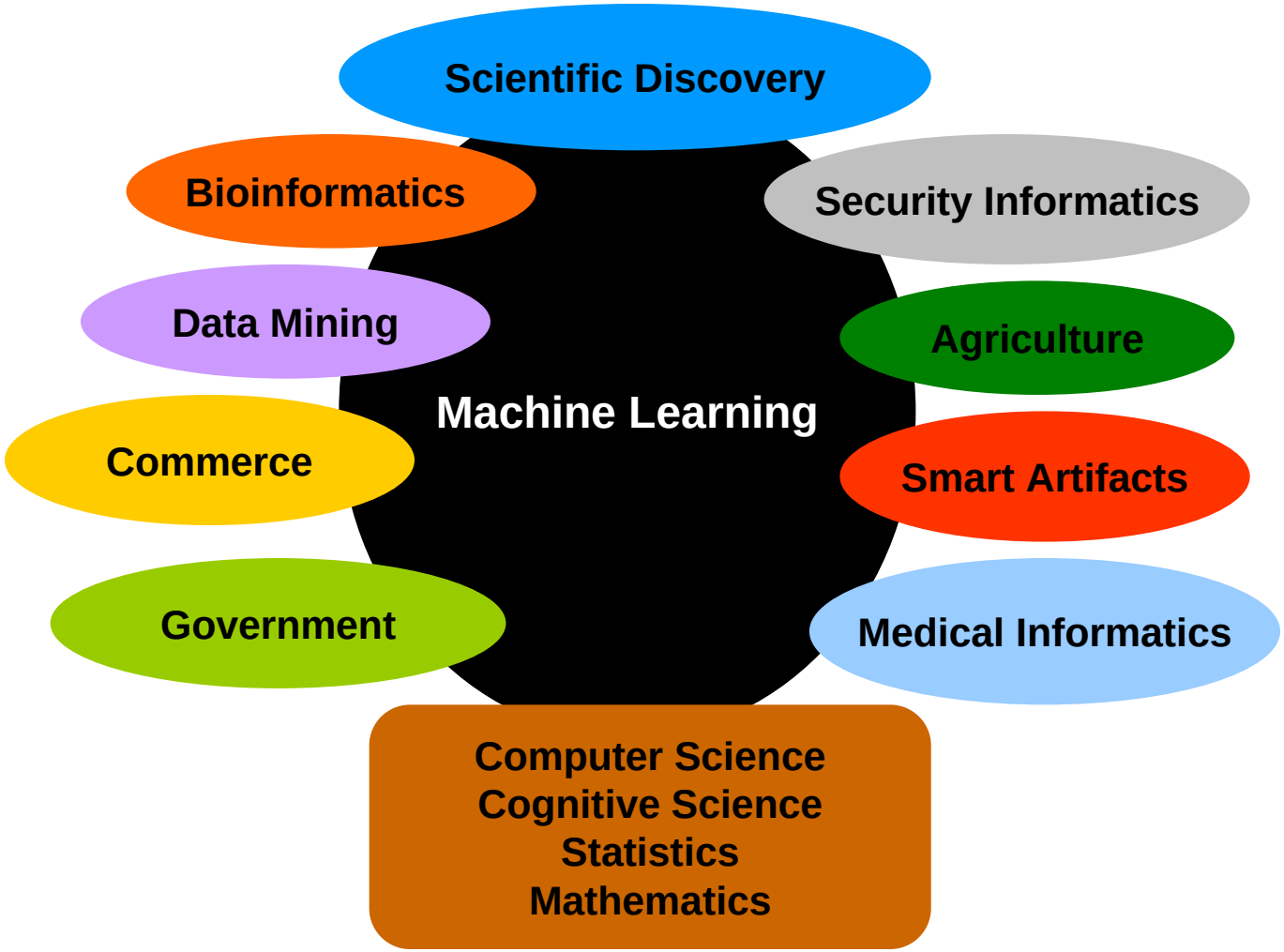
Statistics – Statistical Inference, Experiment Design, Exploratory Data Analysis

Mathematics – Abstract Algebra, Logic, Information Theory, Probability Theory

Psychology and Neuroscience – Behavior, Perception, Learning, Memory, Problem solving

Philosophy – Ontology, Epistemology, Philosophy of Mind, Philosophy of Science

Machine Learning in Context



Machine Learning: Applications

Bioinformatics and Computational Biology

Cognitive Science

e-Commerce, e-Enterprises, e-Government

e-Science

Environmental Informatics

Human Computer Interaction

Intelligent Information Infrastructure

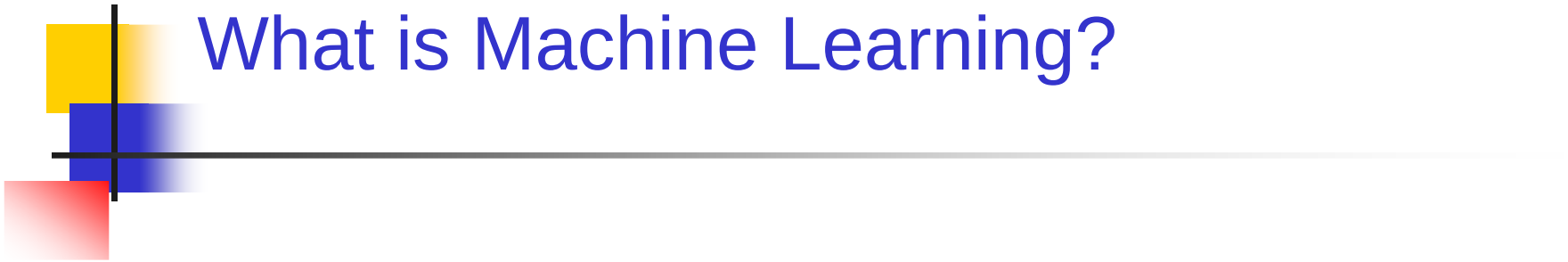
Medical Informatics

Security Informatics

Smart Artifacts

Robotics

Engineering



What is Machine Learning?

A program *M* is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P* if its performance as measured by *P* on tasks in *T* in an environment *Z* improves with experience *E*.

Example 1

T – cancer diagnosis

E – a set of diagnosed cases

P – accuracy of diagnosis on new cases

Z – noisy measurements, occasionally misdiagnosed training cases

M – a program that runs on a general purpose computer

What is Machine Learning?

A program *M* is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P* if its performance as measured by *P* on tasks in *T* in an environment *Z* improves with experience *E*.

Example 2

T – solving calculus problems

E – practice problems + rules of calculus

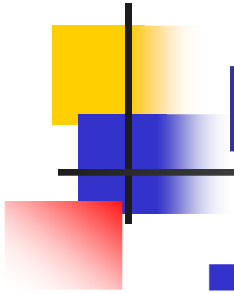
P – score on a test

Example 3

T – driving on the interstate

E – a sequence of sensor measurements and driving actions recorded while observing an expert driver

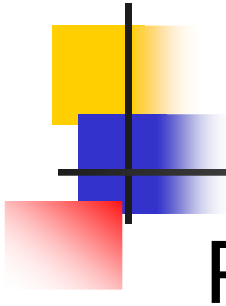
P – mean distance traveled before an error as judged by a human expert



Machine Learning Algorithms

- Many types of Algorithms differing in the structure of the learning problem as well as the approach to learning used

- Regression vs Classification
- Supervised vs Unsupervised vs Semi-Supervised
- Generative vs Discriminative
- Linear vs Non-Linear



Machine Learning Algorithms

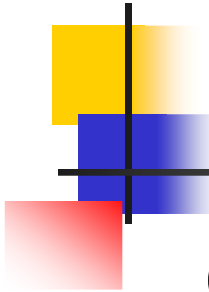
Regression vs Classification

- Structural Difference
  - Regression Algorithms attempt to map inputs into continuous output (Integers, Real Numbers, Vectors, etc.)
  - Classification Algorithms attempt to map inputs into one of a set of classes (Colors, Cellular Locations, Good and Bad Credit Risks, Blogs vs Product Webpages)

Machine Learning Algorithms

Supervised vs Unsupervised vs Semi-Supervised

- Data Difference
  - Supervised Learning involves using pairs of input/output relationships to learn an input output mapping
    - Called 'labeled pairs' often denoted $\{X_i , Y_i\}$
  - Unsupervised Learning involves examining 'input' data to find patterns (clustering)
  - Semi-Supervised Learning uses both labeled data to find input/output mappings and unlabeled data to understand the distribution of the input space
    - Why? There are often many more unlabeled data points than properly labeled data points (ex. webpages)

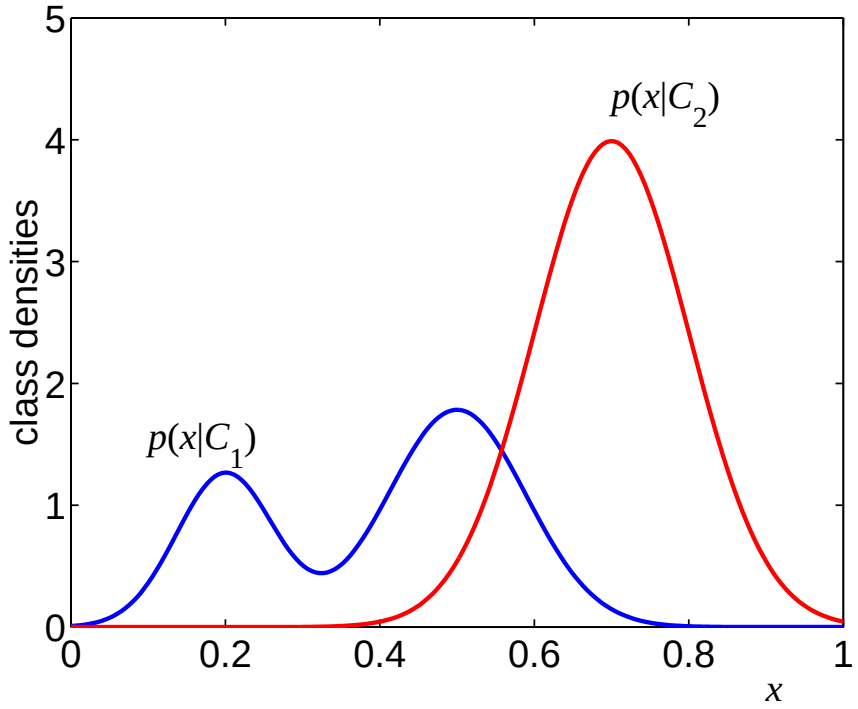


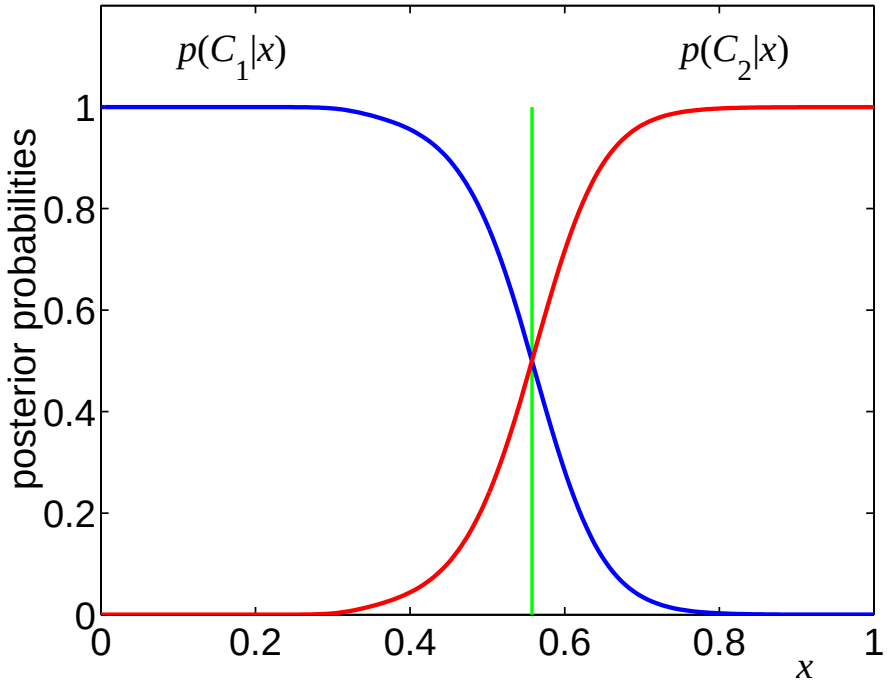
Machine Learning Algorithms

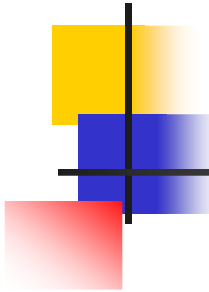
Generative vs Discriminative

- 'philosophical' difference
  - Generative models attempt to recreate or understand the process that generated the data
  - Discriminative models attempt to simply separate or determine the class of input data without regard to the process

Generative vs. Discriminative Models





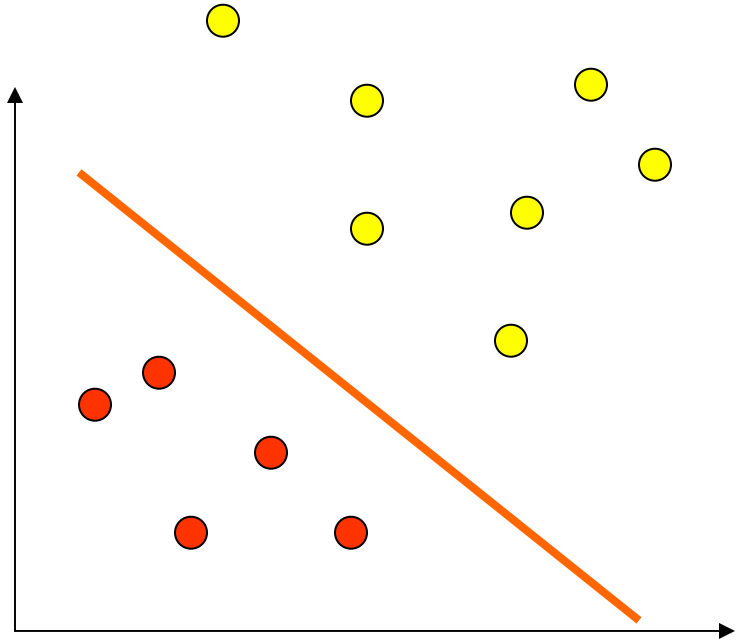


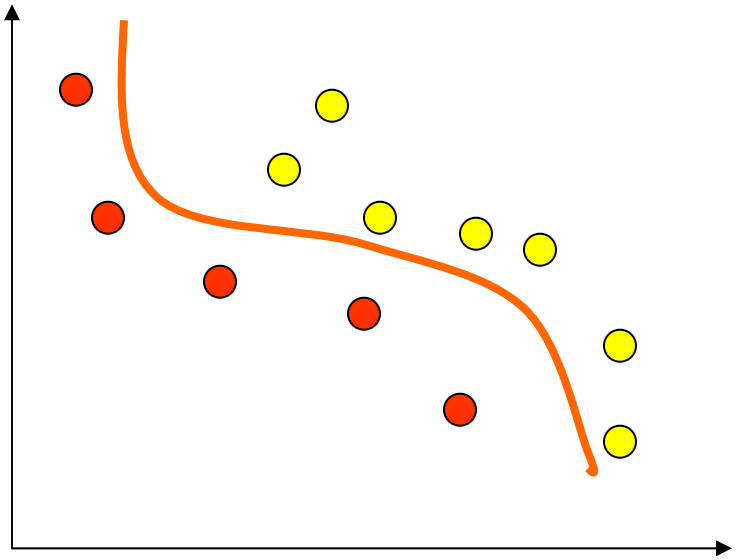
Machine Learning Algorithms

Linear vs Non-Linear

- Modeling Difference
  - Linear models involve only linear combinations of input variables
    - Ex – $a_1x_1 + a_2x_2 + a_3x_3 + b$
  - Non-Linear models are not restricted in their form
    - Common examples exponentials, quadratic terms

Linear vs Non-Linear





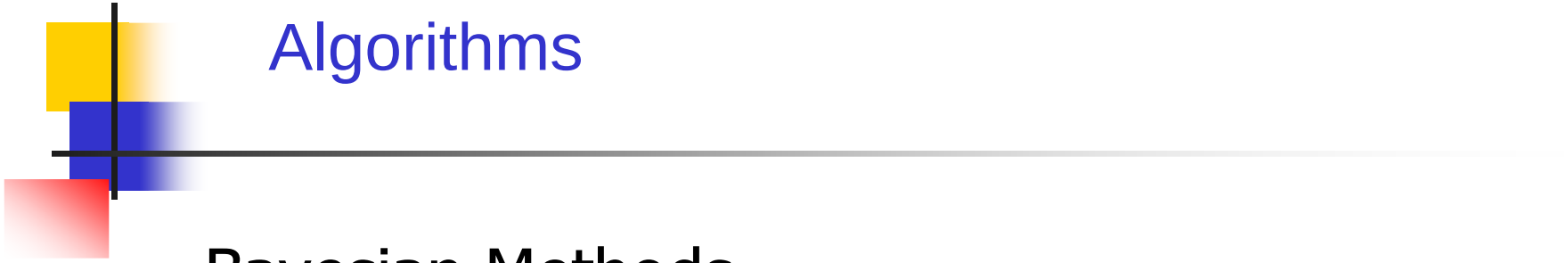
Summary

These aspects are only the tip of the iceberg

No single algorithm works best for every application

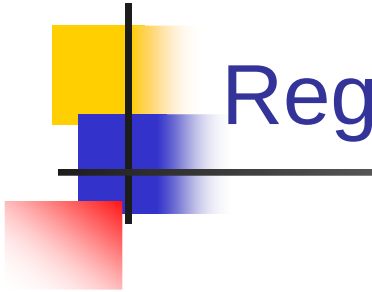
Some simple algorithms are effective on many data sets

Better results can be obtained by preprocessing the data to suit the algorithm or adapting the algorithm to suit the characteristics of the data



Algorithms

- Bayesian Methods
  - Naïve Bayes
- Artificial Neural Networks
  - Perceptrons
  - Hidden Layer Networks
  - Winner Takes All Networks
  - ART – SOM
- SVM



Regression - Measuring Performance

Mostly 'mean squared error'

$$MSE = \sum_i (f(x_i) - y_i)^2$$

Classifier Learning - Measuring Performance

N: Total number of instances in the data set

TP_j: True positives for class j

FP_j : False positives for class j

TN_j: True Negatives for class j

FN_j: False Negatives for class j

$$Accuracy_j = \frac{TP_j + TN_j}{N}$$

$$Accuracy = \frac{\sum_j TP_j}{N}$$

$$Recall_j = \frac{TP_j}{TP_J + FN_j}$$

$$Precision_j = \frac{TP_j}{TP_J + FP_j}$$

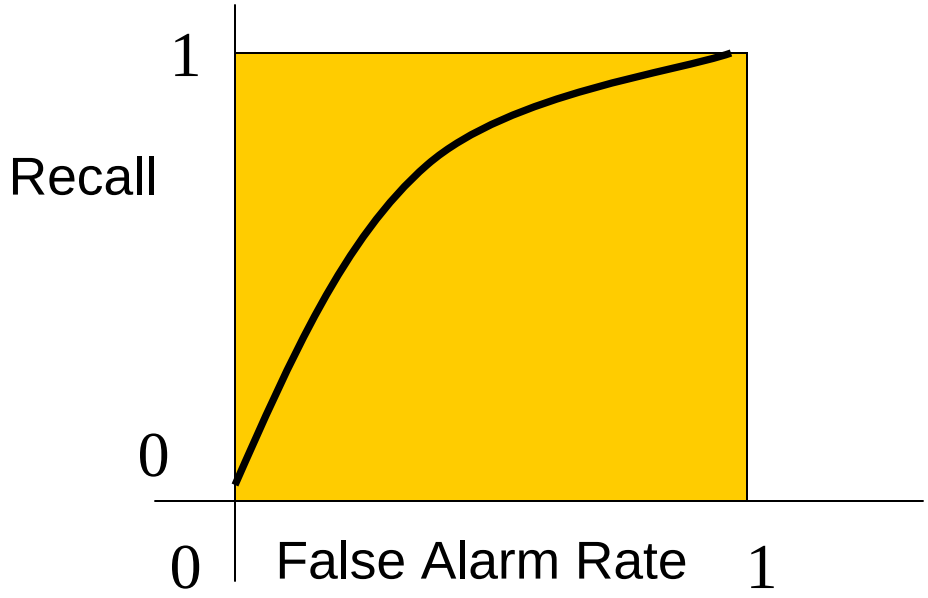
$$FalseAlarm_j = \frac{FP_j}{TP_J + FP_j} = 1 - Precision_j$$

$$CorrelationCoeff_j = \frac{(TP_j \times TN_j) - (FP_j \times FN_j)}{\sqrt{(TP_J + FN_j)(TP_j + FP_j)(TN_j + FP_j)(TN_j + FN_j)}}$$

Receiver Operating Characteristic (ROC) Curve

We can trade off recall versus precision – e.g., by adjusting classification threshold

ROC curve is a plot of Recall against False Positive Rate (1-Precision)



Classifier Learning -- Measuring Performance

The contingency table consisting of *FN TP, FP, TN* contains all the information needed to assess the performance of binary classifiers

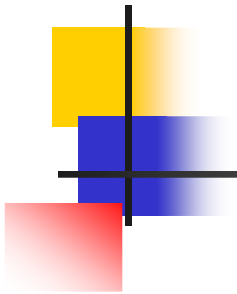
Measures like *Precision, Recall, Accuracy* summarize this information in the form of a single scalar. Any such summary necessarily loses information

Each measure is useful in its own way, but must be used with care – For example, accuracy is misleading when data set has an uneven proportion of examples of different classes

If a single measure of performance is to be reported, perhaps one of the least biased and the most useful measures is the *Correlation Coefficient* – Value of 1 corresponds to the perfect classifier; 0 corresponds to random predictions

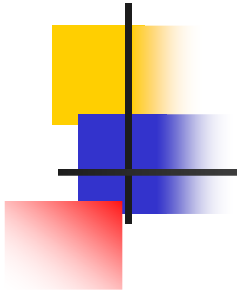
Correlation coefficient can be defined for the case *of M-ary classifiers*

It is often possible to trade off precision against recall



Representative Machine Learning Applications in Bioinformatics and Computational Biology

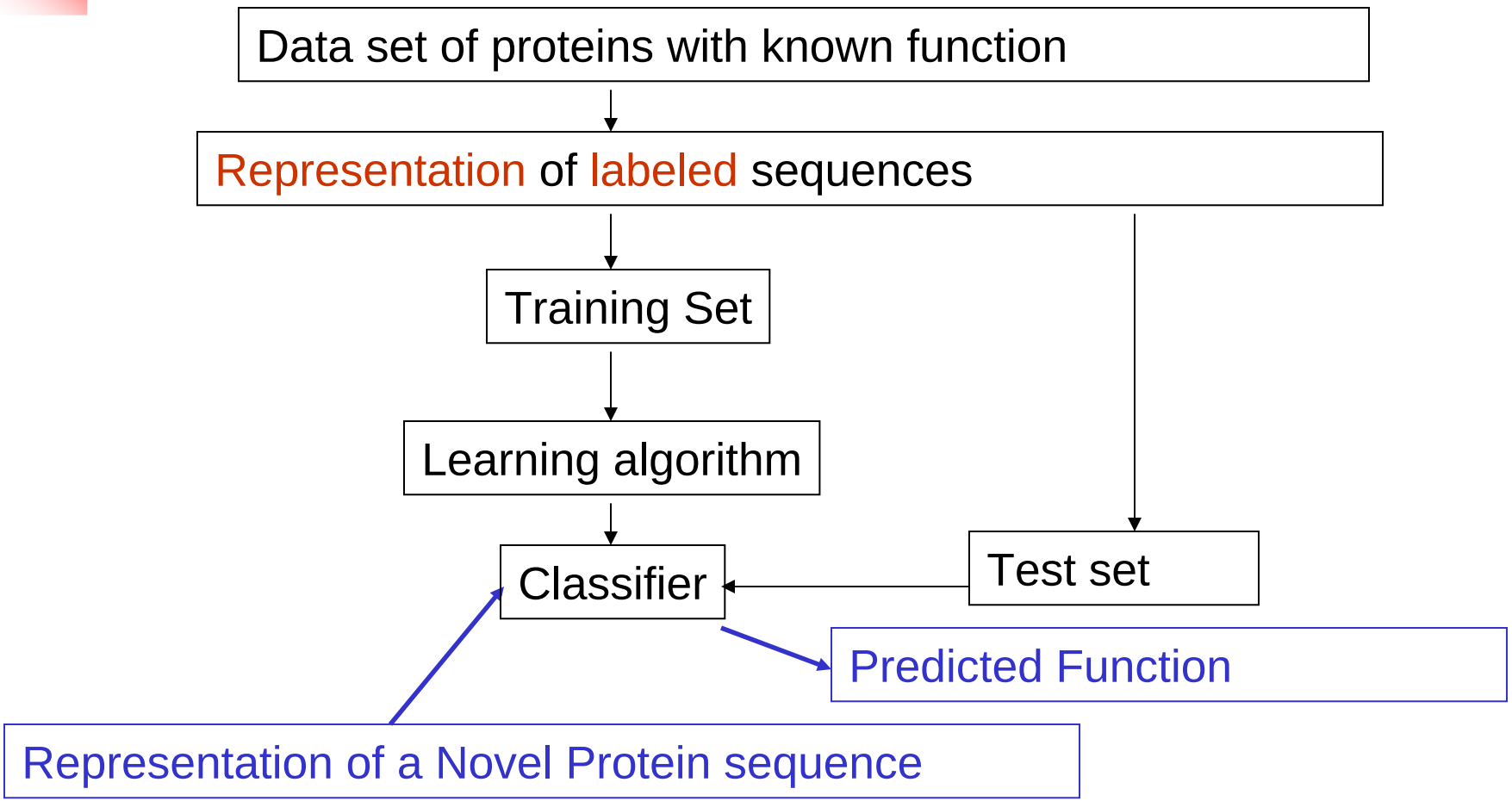
- Gene finding
- Ribosome binding site identification
- Promoter identification
- Prediction of protein structural features
- Protein binding site identification
- Prediction of protein function
- Genetic network inference
- Cancer diagnosis
- Gene annotation



Simple Implementation Example

- Predicting Protein Function
  - Inputs –
  - Outputs –

Sample Learning Scenario: Synthesis of protein function classifiers

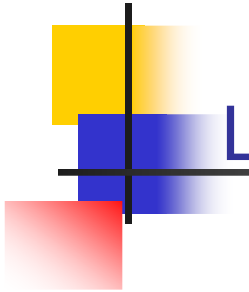


K-fold cross-validation

Recommended procedure for evaluating classifiers when data are limited

Use *K*-fold cross-validation (*K*=5 or 10)

Better still, repeat *K*-fold cross-validation *R* times and average the results



Leave-one-out cross-validation

K-fold cross validation with *K = n* where *n* is the total number of samples available

n experiments – using *n*-1 samples for training and the remaining sample for testing

Leave-one-out cross-validation does not guarantee the same class distribution in training and test data!

Extreme case: 50% class 1, 50% class 2

Predict majority class label in the training data

True error – 50%; Leave-one-out error estimate – 100%!!!!!

Evaluating the performance of classifiers

- Rigorous statistical evaluation of learned classifiers is important
- How good is a learned hypothesis?
- Is one hypothesis better than another?
- Is one learning algorithm better than another on a particular learning task? (No learning algorithm outperforms all others on all tasks – No Free Lunch Theorem)
- Different procedures for evaluation are appropriate under different conditions (large versus limited versus small sample) – Important to know when to use which evaluation method and be aware of pathological behavior (tendency to grossly overestimate or underestimate the target value under specific conditions)

Classifier Learning -- Measuring Performance

The contingency table consisting of *FN TP, FP, TN* contains all the information needed to assess the performance of binary classifiers

Measures like *Precision, Recall, Accuracy* summarize this information in the form of a single scalar. Any such summary necessarily loses information

Each measure is useful in its own way, but must be used with care – For example, accuracy is misleading when data set has an uneven proportion of examples of different classes

If a single measure of performance is to be reported, perhaps one of the least biased and the most useful measures is the *Correlation Coefficient* – Value of 1 corresponds to the perfect classifier; 0 corresponds to random predictions

Correlation coefficient can be defined for the case *of M-ary classifiers*

It is often possible to trade off precision against recall

END OF BYRON'S SLIDES