Machine Learning

Juan Facorro - Marcos Almonacid

Introduction

What is ML?

Introduction

Introduction

Machine learning is a subfield of artificial intelligence (AI) that deals with the construction and study of systems that can **learn** from data.

Introduction

Learn

“To generalize is to learn.” - Unknown

Introduction - Bias

8 = T	4 = T	
2 = T	13 = F	12 = ?
5 = F		
9 = F		

- To generalize we introduce a bias.
- **Occam's Razor**.
- Each method introduces a different bias.

Types of Problems - Classification

Assign a class (or many classes) to an object formed by various attributes.

Examples

- Spam mail
- Identify a digit from a group of pixels
- Assign a subject(s) to a news article

Types of Problems - Regression

Assign a real number to an object.

<u>Examples</u>

- The price of a house
- Predict the stock prices

Types of Problems - Search & Rank

Find and order answers to a problem according to their likelihood.

Examples

- Search engines
- Recommendation systems

Types of Problems - Novelty Detection

Detect outliers which are objects different from the average.

Examples

- Credit card alarms
- Critical systems failure

Types of Problems - Clustering

Form groups of similar objects.

<u>Examples</u>

- Consumer segmentation

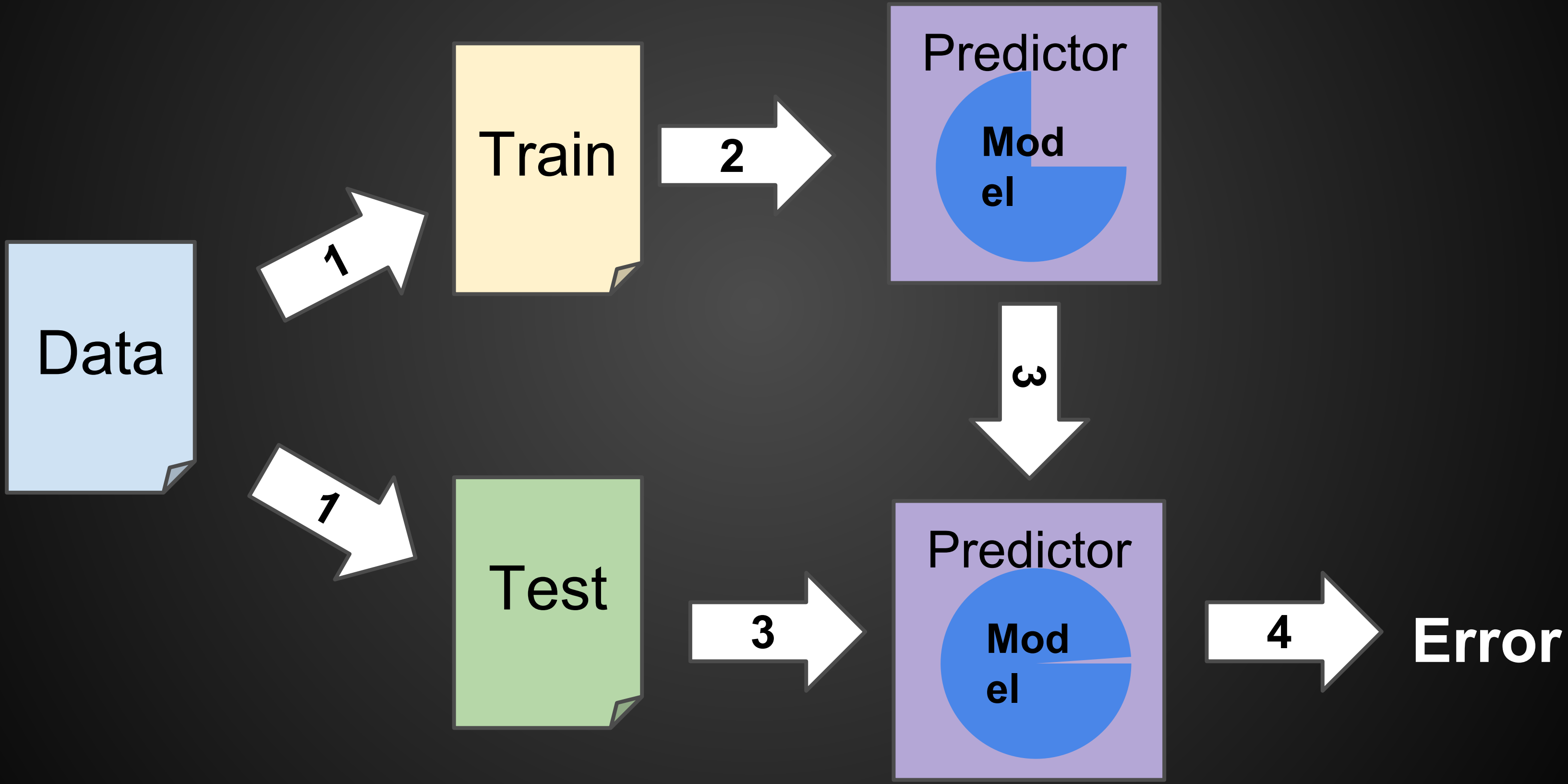
Types of Problems - Relevant Input

Figure out what attributes are responsible for the solution while solving the problems presented before or others.

<u>Examples</u>

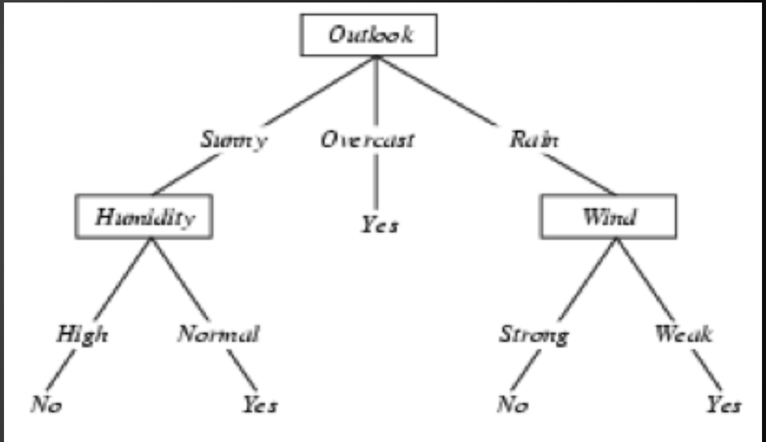
- Identify relevant symptoms in medical research.

Basic ML Mechanism



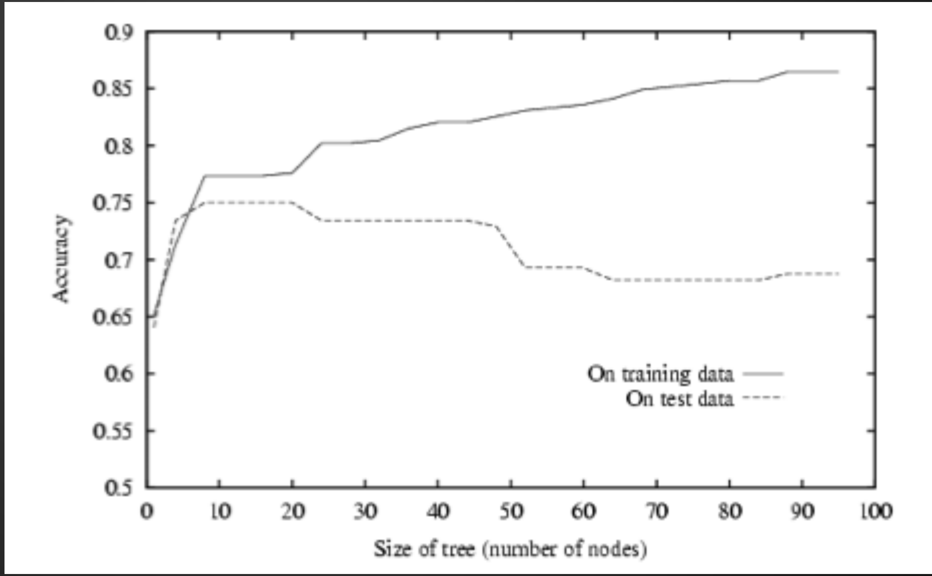
Basic Methods - Decision Trees

- Used in classification.
- Most relevant class on top.
- Done building when leaves are pure classes.

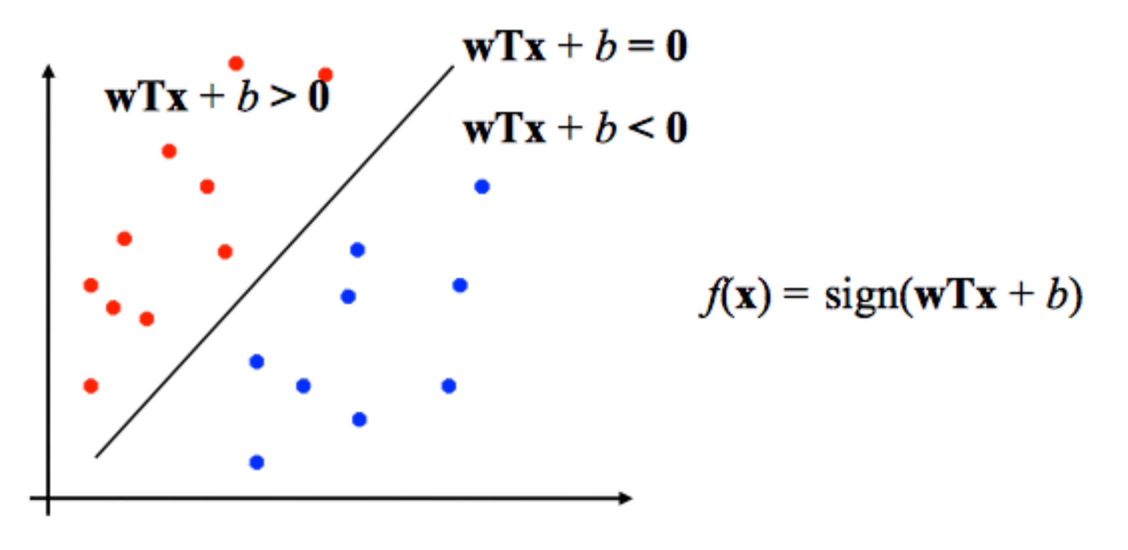


Basic Methods - Decision Trees

Overfitting!

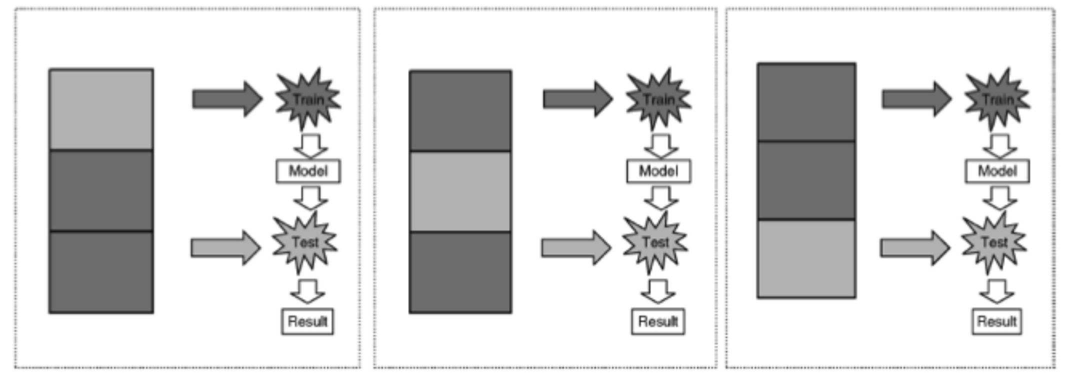


Basic Methods - Perceptron



Results Evaluation - Error Estimation

- Measure with a test data set.
- No test set available, cross-validation.



Cross-Validation. Figure 1. Procedure of three-fold cross-validation.

(Un)Supervised Learning

- Supervised
  - Trained on labelled examples.
  - Attempts to generalise a function or mapping from inputs to outputs.
- Unsupervised
  - Operate on *unlabelled* examples, i.e., input where the desired output is unknown.
  - The objective is to discover structure in the data.

Steps for Solving ML Problems

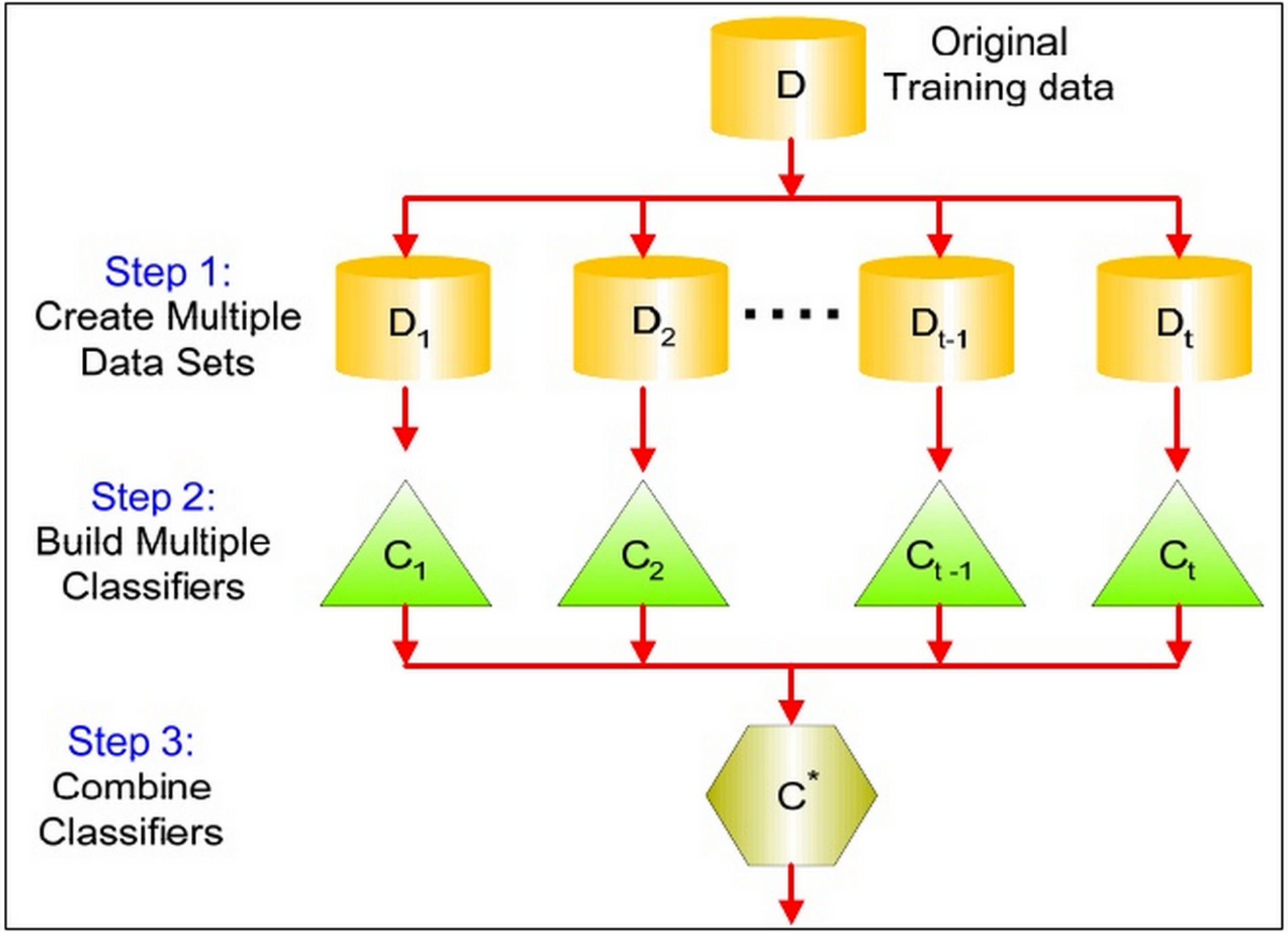
- Identify the problem and get expert knowledge.
- Get lots and lots of data!
- Choose one or more adequate methods.
- Train various models with the **training set**.
- Evaluate them with the **validation set**.
- Estimate error with the **test set**.

Coffee Break & Questions

On to the next one…

Ensembles

What is an ensemble?


D
Original
Training data
Step 1:
Create Multiple
Data Sets
D_1
D_2
D_{t-1}
D_t
Step 2:
Build Multiple
Classifiers
C_1
C_2
C_{t-1}
C_t
Step 3:
Combine
Classifiers
C^*

Ensembles - Categories

Plain

- all experts
- all good individually
- different opinions in some cases

Divisive

- divide and conquer
- useful for big problems
- need a selective function

Ensembles - Success Case

NETFLIX

Netflix Prize

COMPLETED

Home | Rules | Leaderboard | Update

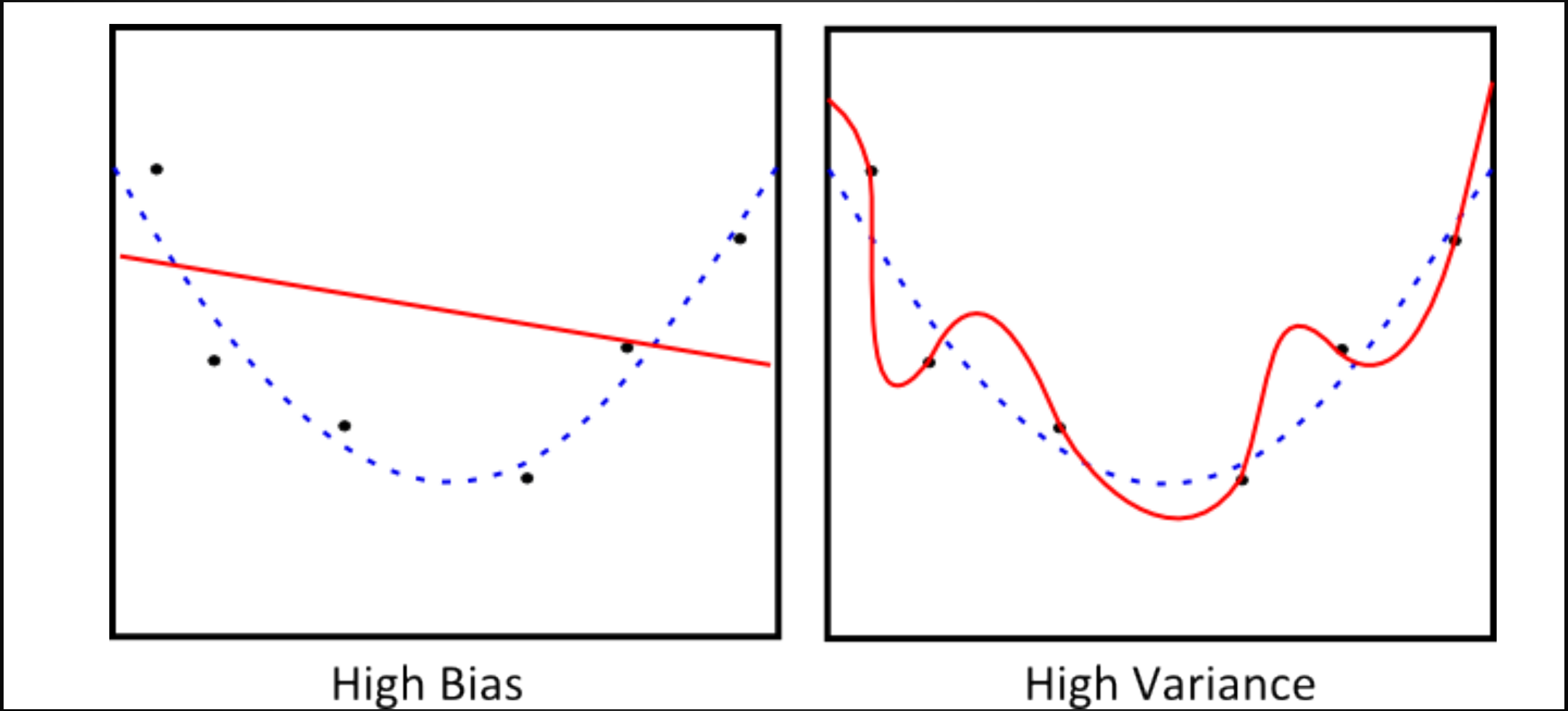
Leaderboard

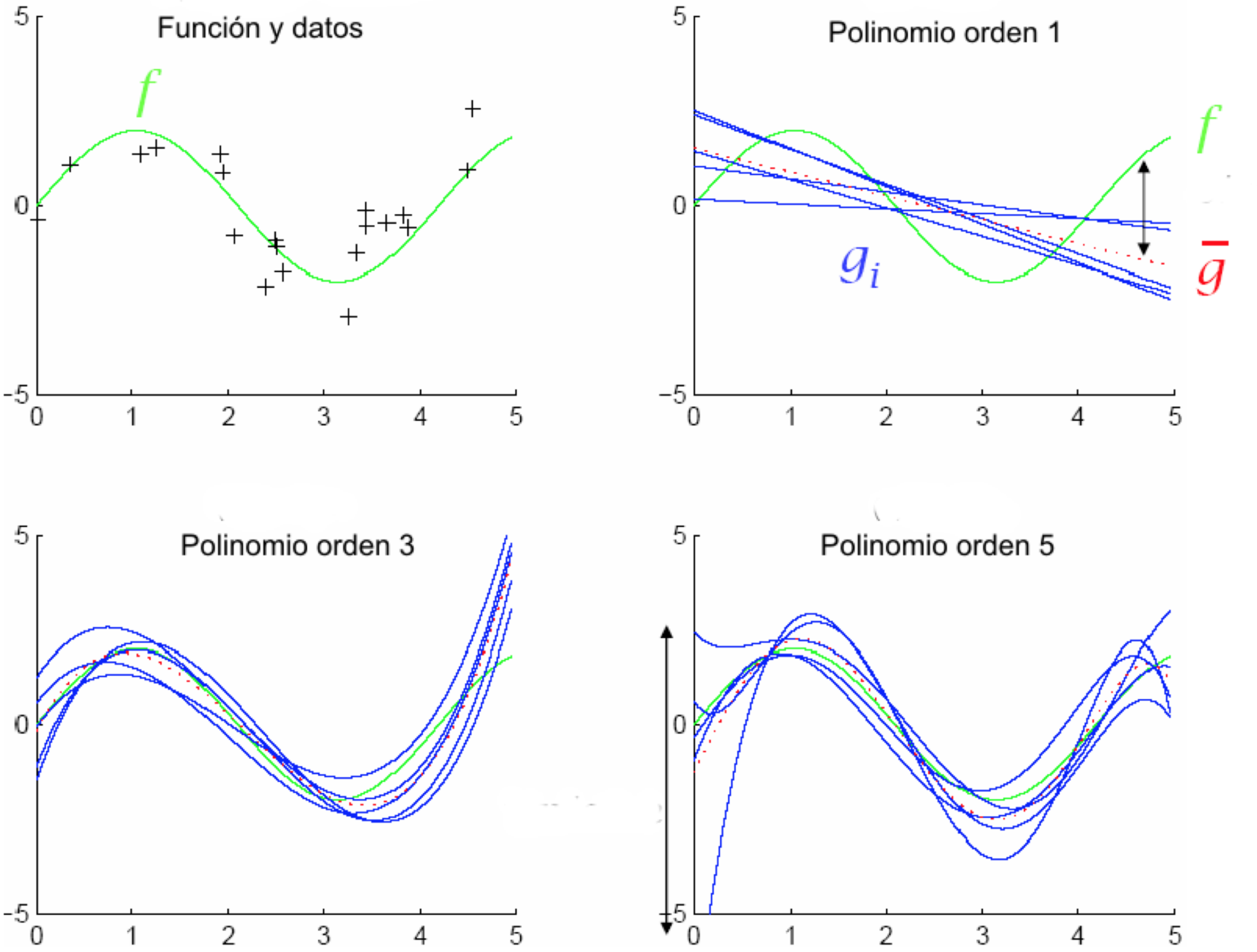
Showing Test Score. Click here to show quiz score

Display top 20 **leaders.**

Rank	Team Name	Best Test Score	% Improvement	Best Submit Time
Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos				
1	BellKor's Pragmatic Chaos	0.8567	10.06	2009-07-26 18:18:28
2	The Ensemble	0.8567	10.06	2009-07-26 18:38:22
3	Grand Prize Team	0.8582	9.90	2009-07-10 21:24:40
4	Opera Solutions and Vandelay United	0.8588	9.84	2009-07-10 01:12:31
5	Vandelay Industries !	0.8591	9.81	2009-07-10 00:32:20
6	PragmaticTheory	0.8594	9.77	2009-06-24 12:06:56
7	BellKor in BigChaos	0.8601	9.70	2009-05-13 08:14:09
8	Dace_	0.8612	9.59	2009-07-24 17:18:43
9	Feeds2	0.8622	9.48	2009-07-12 13:11:51
10	BigChaos	0.8623	9.47	2009-04-07 12:33:59
11	Opera Solutions	0.8623	9.47	2009-07-24 00:34:07
12	BellKor	0.8624	9.46	2009-07-26 17:19:11

Ensembles - Bias/Variance Dilemma





Función y datos
f
Polinomio orden 1
g_i
$\bar{g}$
Polinomio orden 3
Polinomio orden 5

Ensembles - Bias/Variance Dilemma

SOLUTIONS

Reduce Variance when no Bias
(Bagging and Random Forest)

Reduce Bias when Stable Predictors
(Boosting)

Ensembles - Bagging

- Simple and effective
- Uses bootstraps and unstable predictors
- Reduces variance averaging predictors

Ensembles - Bagging

Analogy:

Diagnosis based on multiple doctors’ majority vote.

Ensembles - Bagging

Bootstrap

Data Set

2

4

6

7

10

34

Samples

7, 7, 34, 2, 6, 2

10, 2, 4, 6, 6, 7

7, 34, 4, 6, 4, 7

Ensembles - Bagging

Combination Rules

For Classification:

Each model votes for a class. The class with more votes is chosen

Ensembles - Bagging

Combination Rules

For Regression/Ranking:

Simple average of predictions

Ensembles - Random Forest

- Improves Bagging only for decision trees
- Very used. Almost automatic
- Results comparable with current methods

Ensembles - Random Forest

Adds little changes to tree growth

No Bias

Adds Variance (but is fine)

Ensembles - Boosting

Multiple Weak Predictors

Bootstrap

Weight

Ensembles - Boosting

Combination Rules

Takes a vote from each model

Weighted average of votes

Ensembles - Boosting

- constructing D_t:
  - $D_1(i) = 1/m$
  - given D_t and h_t:

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} e^{-\alpha_t} & \text{if } y_i = h_t(x_i) \\ e^{\alpha_t} & \text{if } y_i \neq h_t(x_i) \end{cases}$$

$$= \frac{D_t(i)}{Z_t} \exp(-\alpha_t \, y_i \, h_t(x_i))$$

where $Z_t = \text{normalization constant}$

$$\alpha_t = \tfrac{1}{2} \ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right) > 0$$

- final classifier:
  - $H_{\text{final}}(x) = \text{sign}\left(\sum_t \alpha_t h_t(x)\right)$

AND THAT'S ALL I HAVE

TO SAY ABOUT THAT.

Support Vector Machines

SVM

Huh... What is it good for?

Classification

Regression

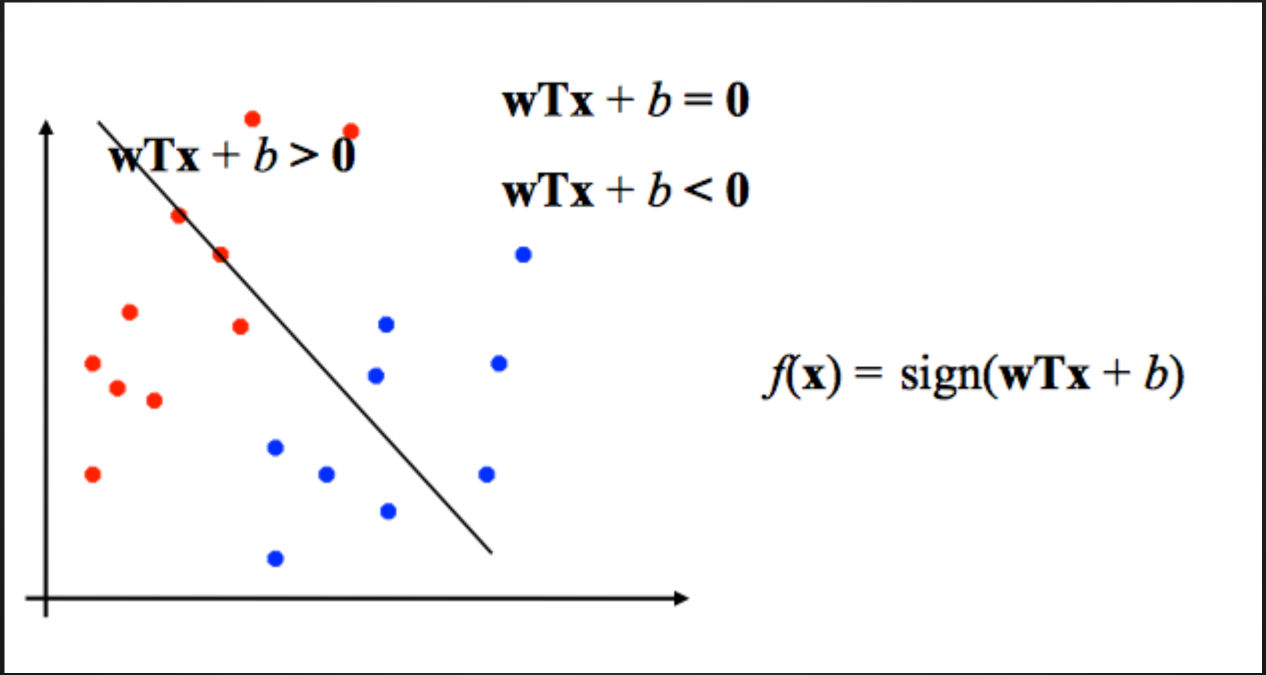
Novelty Detection

SVM

Huh... What is it good for?

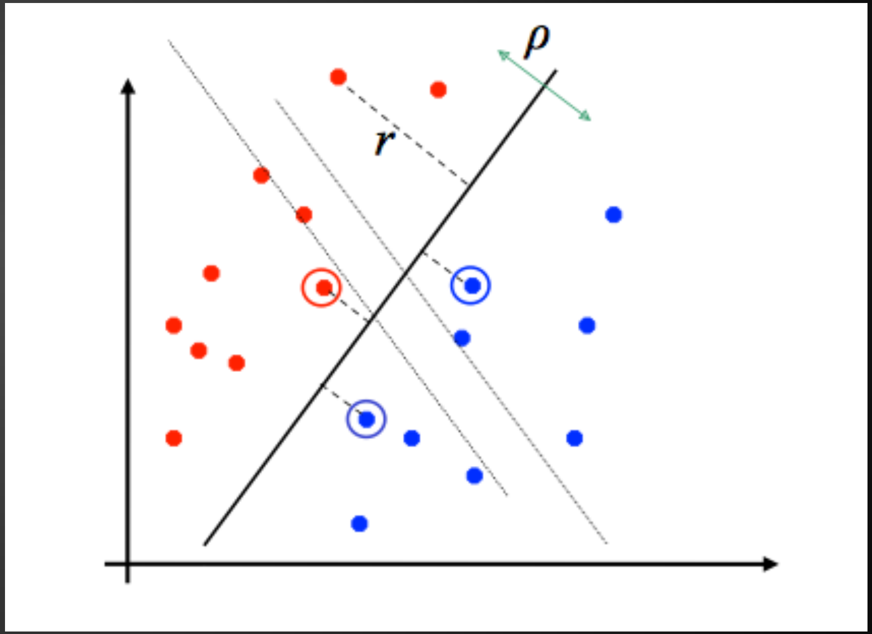
- Particle Identification.
- Face Identification.
- Text Categorization.
- Bioinformatics.

SVM - Linear Separators



SVM - Classification Margin

- Examples closest to the hyperplane are ***support vectors***.
- Maximizing the margin is good according to intuition and learning theory.



SVM - Formulation

- The Lagrangian:

$$L(w, b, \xi, \alpha, \mu) = J(w, b, \xi) + \sum_t \alpha_t[1 - \xi_t - y_t(wx_t + b)] - \sum_t \mu_t \xi_t$$

$$= \frac{\|w\|^2}{2} + C\sum_{t=1}^{T} \xi_t + \sum_t \alpha_t[1 - \xi_t - y_t(wx_t + b)] - \sum_t \mu_t \xi_t$$

$$(\alpha_t \geq 0 \quad \text{and} \quad \mu_t \geq 0)$$

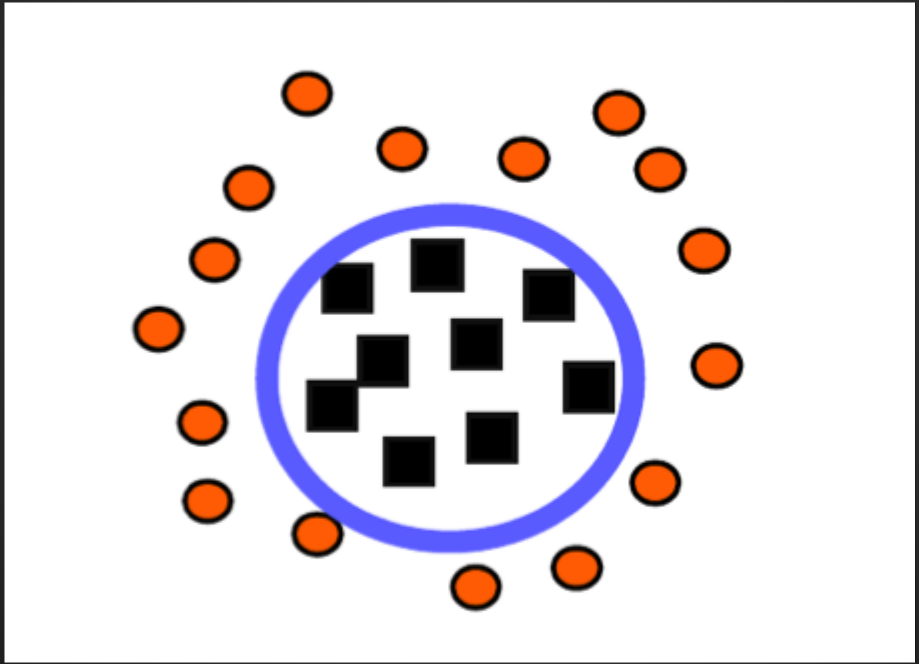
- Look for (w, b, ξ) minimum of L:

$$\frac{\partial L}{\partial w} = 0 \quad \Leftrightarrow \quad w = \sum_t \alpha_t y_t x_t$$

$$\frac{\partial L}{\partial b} = 0 \quad \Leftrightarrow \quad \sum_t \alpha_t y_t = 0$$

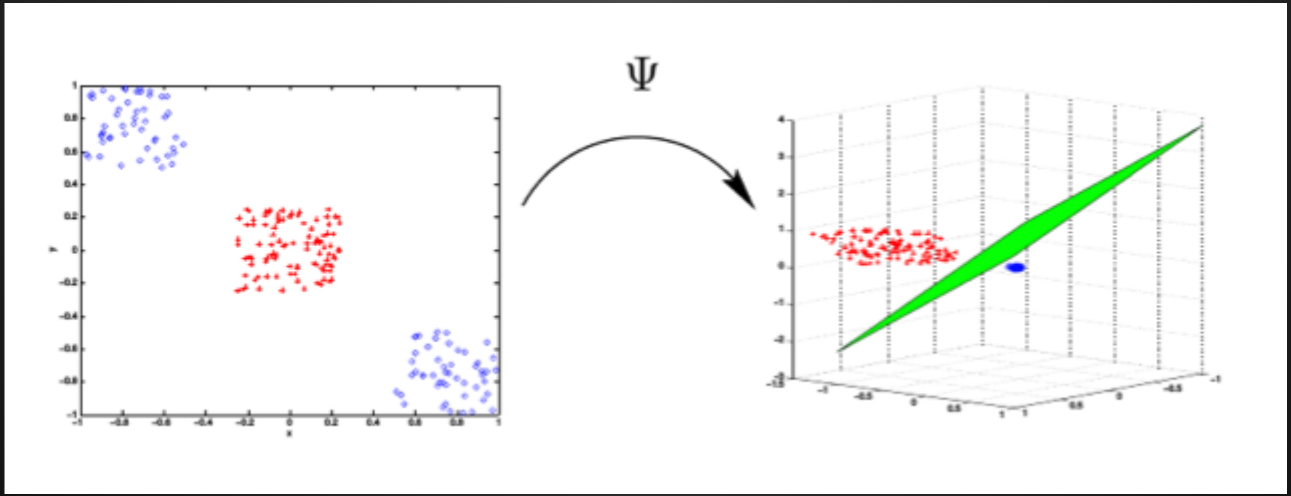
$$\frac{\partial L}{\partial \xi} = 0 \quad \Leftrightarrow \quad C - \alpha_t - \mu_t = 0$$

SVM - Kernels



SVM - Kernels

Project data into a higher dimensional space.
Oooooh, aaaaah.



SVM - Kernels

- High Dim. Space: Dot product == expensive
- Kernels: function that represents the dot product.

Example: instead of

$$\Psi(x) = \begin{pmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \end{pmatrix}$$

use

$$k(x,\, z) = (xz)^2$$

SVM - Using them

- Choose a kernel.
- Minimize the margin.
- Use the decision function:

$$x \mapsto \operatorname{sign}\left(\sum_t \alpha_t y_t \, k(x_t, x) + b\right)$$

SVM - Summary

- SVMs maximize the margin.
- Projects data into a higher dimensional space.
- Kernels simplify the computation.

No Coffee Break For You

Artificial Neural Networks

Some other time…