

Ch0. Introduction and Overview

ST4240, 2014/2015

Alexandre Thiéry

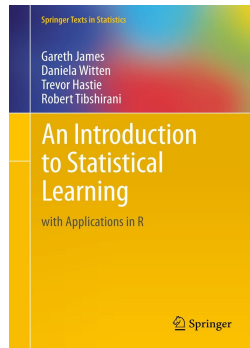
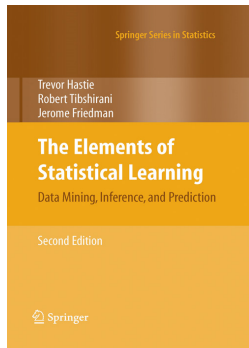
Department of Statistics and Applied Probability

Outline

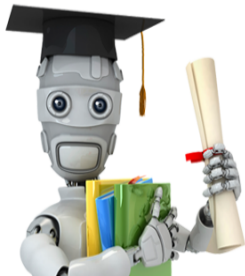
1 Ressources

2 Some applications

3 In this course...



Machine Learning online lecture by Andrew Ng



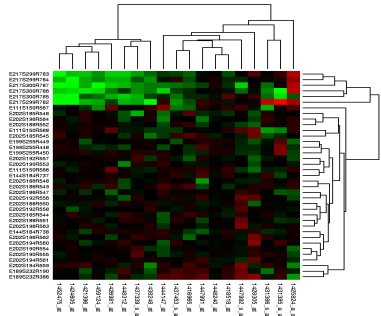
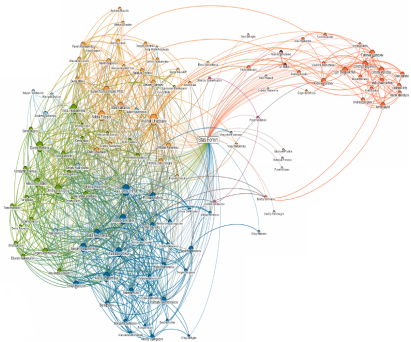
Outline

1 Ressources

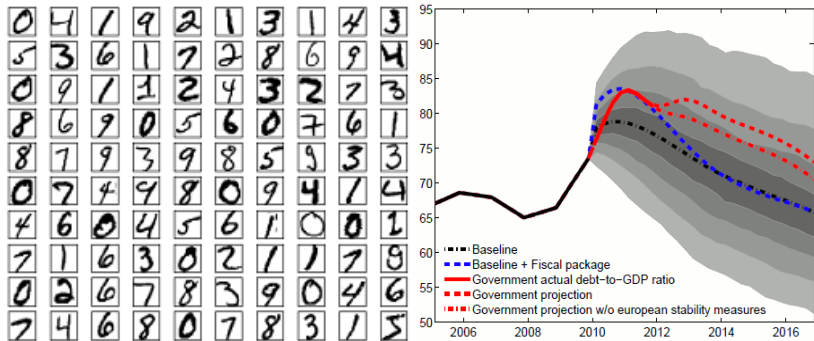
2 Some applications

3 In this course...

- **Descriptive data mining:** Search massive data sets and discover the locations of unexpected structures or relationships, patterns, trends, clusters, and outliers in the data.



- **Predictive data mining:** Build models and procedures for regression, classification, pattern recognition, or machine learning tasks, and assess the predictive accuracy of those models and procedures when applied to fresh data.



Avalanche of data

- Financial transactions : billions of transactions per year
- Analysis of internet traffic data
- Human Genome Project has to deal: gigabytes
- remote-sensing satellite systems:gigabytes per hour
- U.S. census file: $\geq 10^{12}$ bytes

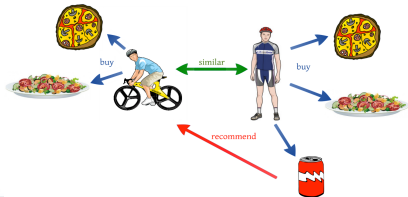


- **Marketing:** Predict new purchasing trends. Given customers who have purchased product A, B, or C, identify those who are likely to purchase product D.

Grant, Welcome to Your Amazon.com (If you're not Grant Ingersoll, [click here.](#))

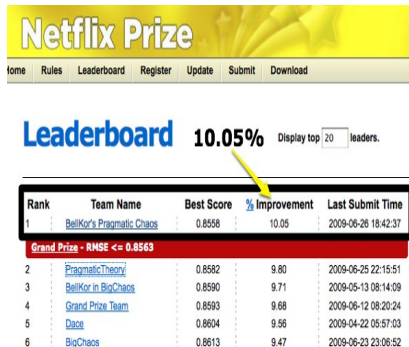
Today's Recommendations For You

Here's a daily sample of items recommended for you. Click here to [see all recommendations](#).



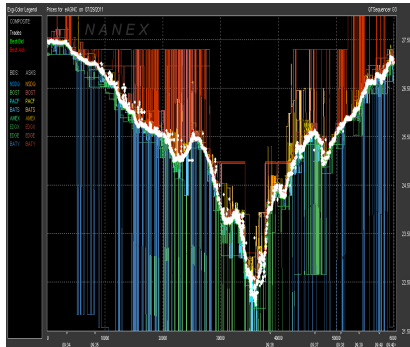
■ Netflix Challenge:

- 10^7 ratings for 17700 movies
- Goal: film recommendation
- One million dollar prize for a 10% improvement

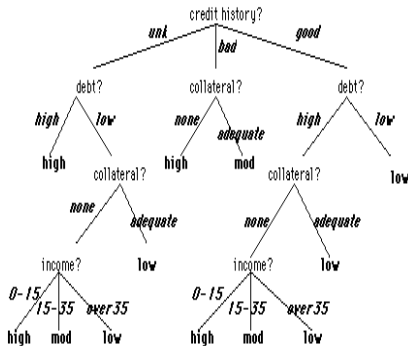


Rank	Team Name	Best Score	% Improvement	Last Submit Time
1	BellKor's Pragmatic Chaos	0.8558	10.05	2009-06-26 18:42:37
Grand Prize - RMSE <= 0.8563				
2	PragmaticTheory	0.8582	9.80	2009-06-25 22:15:51
3	BellKor in BigChaos	0.8590	9.71	2009-05-13 08:14:09
4	Grand Prize Team	0.8593	9.68	2009-06-12 08:20:24
5	Dace	0.8604	9.56	2009-04-22 05:57:03
6	BigChaos	0.8613	9.47	2009-06-23 23:06:52

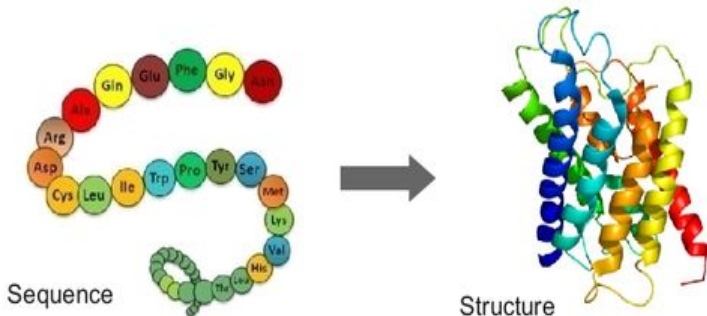
- **Financial Markets:** Analyse volatility patterns in high-frequency stock transactions using volume, price, and time of each transaction.



- **Insurance/credit scoring:** Identify characteristics of buyers of new policies. Find unusual claim patterns. Find "risky" customers.



- **Molecular Biology:** Analyse amino acid sequences and DNA microarrays. Predict protein structure and identify related proteins.



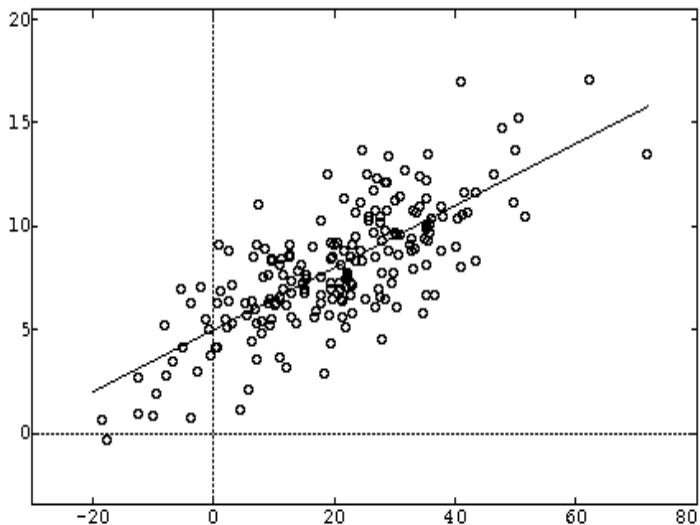
Outline

1 Ressources

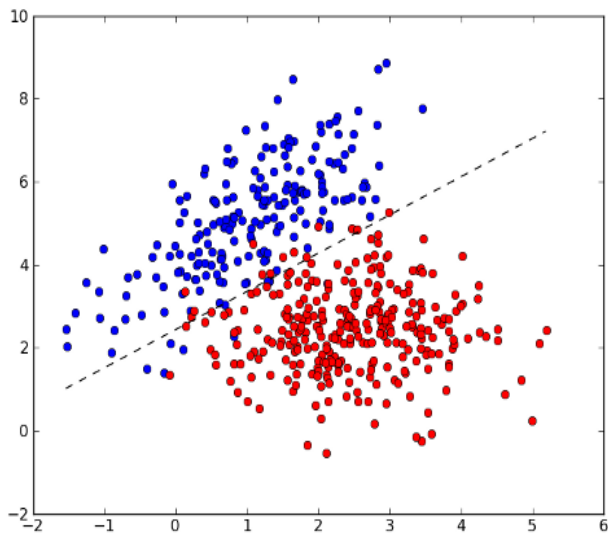
2 Some applications

3 In this course...

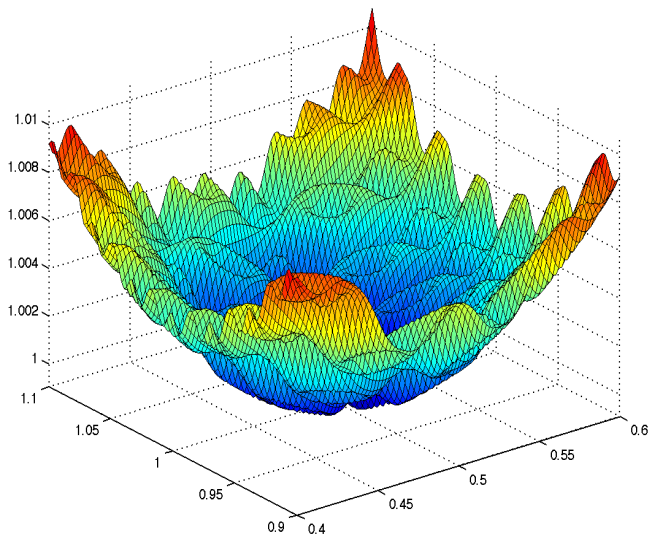
Regression



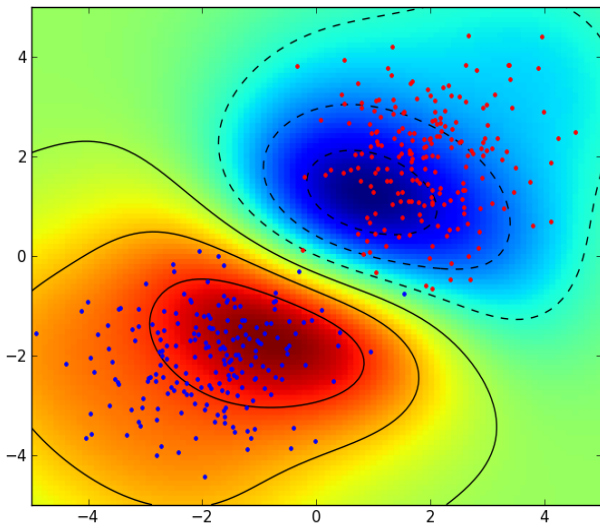
Classification



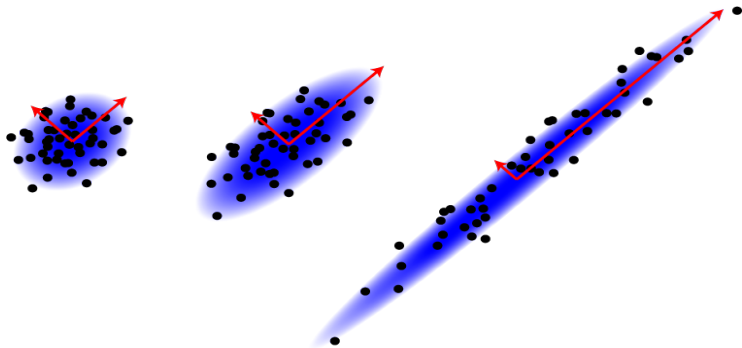
Optimization



Support Vector Machine (SVM) and Kernel Methods



Dimension reduction



Tree methods, Random Forest

