JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Applied and Computational Mathematics

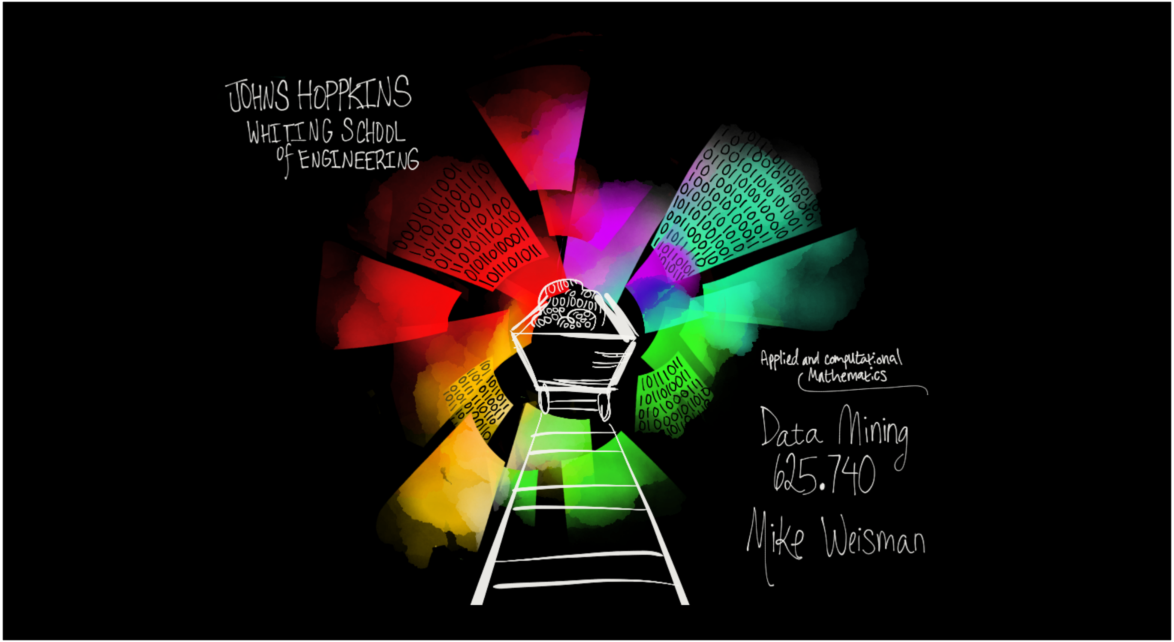
Data Mining

625.740

Introduction to Data Mining

Mike Weisman

email: `data.mining.625.740@gmail.com`


JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING
Applied and computational Mathematics
Data Mining
625.740
Mike Weisman

What is Data Mining?

Data mining is an interdisciplinary subject incorporating elements of statistics, machine learning, artificial intelligence, and data processing.

- Partitioning data into related subsets
- Search and retrieval of useful information from databases
- Extracting patterns from data using computer algorithms or statistical techniques

In this course, we will explore methods for preprocessing, visualizing, and making sense of data, focusing not only on the methods but also on the mathematical foundations of many of the algorithms of statistics and machine learning.

References

1. E. Alpaydin, **Introduction to Machine Learining**
2. S. Ben-David and S. Shalev-Shwartz, **Understanding Machine Learning: From Theory to Algorithms**
3. R. Duda, P. Hart, and Stork, **Pattern Classification**
4. K. Fukunaga, **Introduction to Statistical Pattern Recognition**
5. J. Friedman, R. Tibshirani, and T. Hastie, **Elements of Statistical Learning**
6. G. James, D. Witten, T. Hastie, and R. Tibshirani, **An Introduction to Statistical Learning: with Applications in R**
7. K. Murphy, **Machine Learning: A Probabilistic Approach**
8. P. Tan, M. Steinbach, and V. Kumar, **Introduction to Data Mining**
9. S. Theodoridis and K. Koutroumbas, **Pattern Recognition**
10. L. Torgo, **Data Mining with R: Learning with Case Studies**
11. L. Wasserman, **All of Statistics**

Piazza

https://piazza.com/signup

piazza 625.740 Q & A Resources Statistics Manage Class

hw1 hw2 hw3 hw4 hw5 class_notes latex linear_algebra project exam logistics other data

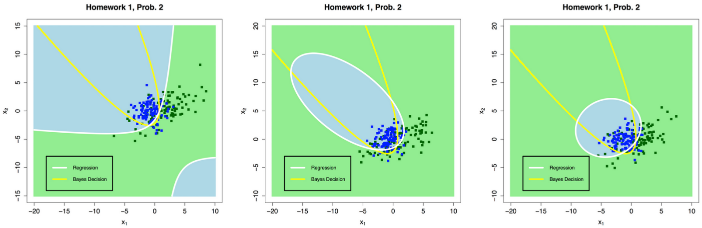
Note History: History Disabled By Instructor

Homework 1, Problem 2

In the figure, the yellow line is the Bayes Decision Boundary found by setting the probability distributions of each class equal. The white line is the Regression Boundary found from linear regression with the terms $\{1, x_1, x_2, x_1x_2, x_1^2, x_2^2\}$.

The Bayes decision boundary is a parabola and regression boundaries that are hyperbolae and ellipses have been found.

Output images showing data, regression boundary, and Bayes decision boundary:



R Code:

```
library(mvtnorm)
mu1 <- c(-1,0)
mu2 <- c( 1,0)
P<-matrix(c(cos(pi/6),sin(pi/6),-sin(pi/6),cos(pi/6)),ncol=2);
Sigma1 <- matrix(diag(c(2,2)),ncol=2)
Sigma2 <- P %*% matrix(diag(c(9,2)),ncol=2) %*% t(P)
N<-100
a<-rmvnorm(n=N,mean=mu1,sigma=Sigma1)
b<-rmvnorm(n=N,mean=mu2,sigma=Sigma2)
x<-rbind(b,a)
```

Student Assessment Criteria

Homework	20%
Class Project	30%
Exams	50%

Project

- An interesting data mining topic
- More in-depth treatment than what we have done in class
- A project proposal will be due around mid-semester
- Students will have an opportunity to give a presentation near the conclusion of the semester
- Fun!

Exams

- Theoretical problems
- Computations involving data and algorithms

Course Outline

We will *emphasize* the use of techniques from the fields of machine learning and statistics.

- Review of Statistics
- Parameteric Models
- Unsupervised Learning
- Regression
- Bayesian Classifiers
- Neural Networks
- Support Vector Machines
- Additional Topics

Supervised Learning

- The inputs are $\mathbf{x} \in \mathscr{X}$, the domain set.
- The outputs are $y \in \mathscr{Y}$, the label set.
- Training data $\mathscr{S} = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\}, \quad \mathscr{S} \in \mathscr{X} \times \mathscr{Y}$.

Supervised Learning Tasks

Regression:
Find coefficients β_j to model $y = \sum_j \beta_j x_j + \beta_0$.

Classification:
y is a member of a finite set, for example, in digit classification $y \in \{0, \cdots, 9\}$

Measures of Success

Let us assume that there is some probability distribution, $\mathscr{D}$, over $\mathscr{X}$ and an underlying correct labeling $f : x \to y$.
To produce the training data, we sample $\{x_i\}, \quad i = 1, \ldots, m$ over $\mathscr{D}$ and label it by the function f.
Our goal is to come up with a prediction rule $h : x \to y$.
We define the Risk to be:

$$L = P_{x \sim \mathscr{D}}[h(x) \neq f(x)] = \mathscr{D}(\{x : h(x) \neq f(x)\}).$$

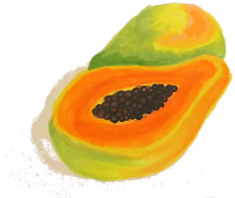
The Training Error or Emperical Risk is

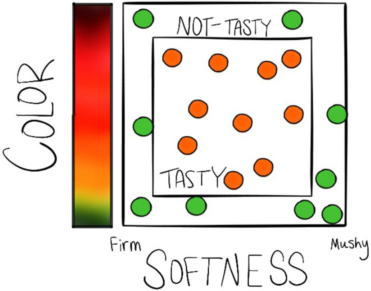
$$L_{\mathscr{S}}(h) = \frac{|\{i \in \{1, \ldots, m\} : h(x_i) \neq y_i\}|}{m}.$$

Finding a predictor h that minimizes $L_{\mathscr{S}}(h)$ is called Emperical Risk Minimization.

Overfitting

Papaya Feature Data





Predictor: $h_s(x) = \begin{cases} y_i, & \text{if } \exists i \in \{i, \ldots, m\} \ni x_i = x \\ 0, & \text{otherwise} \end{cases}$

This is an emperical minimum cost algorithm: $L_{\mathscr{S}}(h_s) = 0$, yet $L_{\mathscr{D}}(h_s) = \frac{1}{2}$.

Unsupervised Learning

- There is no output variable.
- A goal for unsupervised learning may be to cluster the data based on the features of $\mathbf{x} \in \mathscr{X}$.
- Often used as a preprocessing step for supervised learning or where labels do not exist are not available.

Learning and Adaptation*

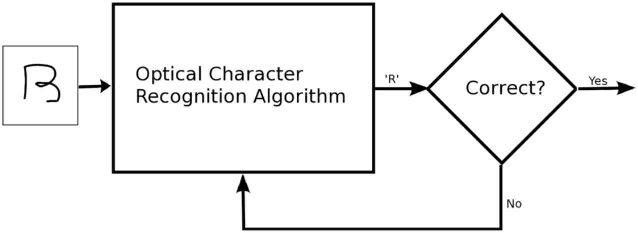
- Reinforcement Learning
  - No desired category is given
  - Feedback given whether tentative category is right or wrong
- Unsupervised Learning
  - System looks for patterns in data
  - Number of categories (or clusters) and what these categories are may be unknown beforehand
- Supervised Learning
  - Category labels are provided with costs for misclassification
  - Learning algorithm seeks to reduce cost

---

*R. Duda and P. Hart, **Pattern Classification and Scene Analysis**

Reinforcement Learning Example

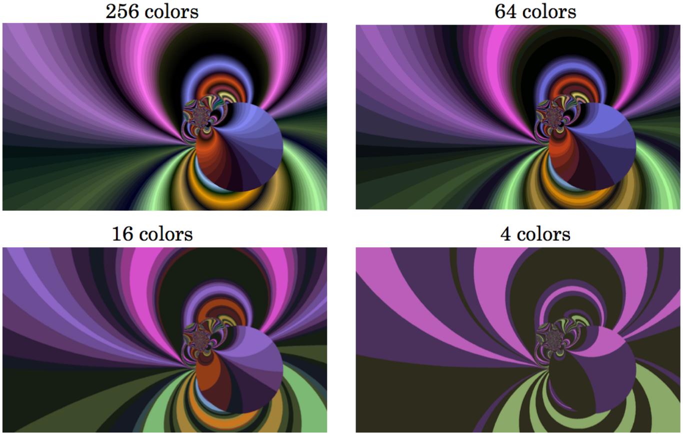
Optical Character Recognition



In reinforcement learning, no desired category signal is given.
The only feedback is ‘correct’ or ‘incorrect’.

Unsupervised Learning Example

Image compression



Supervised Learning Example

JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

Applied and Computational Mathematics

Data Mining

625.740

Introduction to Data Mining