Homework 1 Presentation

Brandon Hosley

University of Illinois - Springfield

August 29, 2020

Overview

1. Supervised and Unsupervised Learning

2. Hastie and Tibshirani

Supervised and Unsupervised Learning

Q: What's the difference between Supervised and Unsupervised learning?

Supervised and Unsupervised Learning

Supervised Learning

- Labeled Data

Unsupervised Learning

- Unlabeled Data

Remark

Supervised learning requires labeled or pre-classified data.

Caution!

Labeled data often comes with a greater up-front cost, typically through manual classification.

Supervised and Unsupervised Learning

Supervised Learning

- Labeled Data
- Known Features

Unsupervised Learning

- Unlabeled Data
- Unknown Features

Remark

Labeled data implies that the feature of interest is already known.

Examples:

· Training a model to classify pictures of animals
· Training a model for handwriting recognition

Supervised and Unsupervised Learning

Supervised Learning

- Labeled Data
- Known Features
- Leverages Experience

Unsupervised Learning

- Unlabeled Data
- Unknown Features
- Discovers New Patterns

Remark

Supervised learning will model already established patterns; unsupervised may discover new patterns, or new ways to group or cluster data.

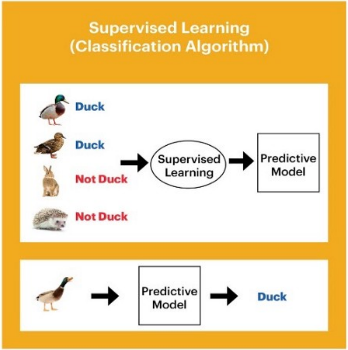
Supervised and Unsupervised Learning

Supervised Learning

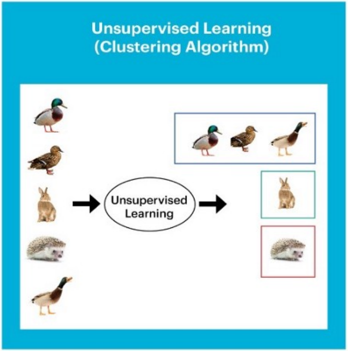
- Labeled Data
- Known Features
- Leverages Experience

Unsupervised Learning

- Unlabeled Data
- Unknown Features
- Discovers New Patterns



Supervised Learning (Classification Algorithm)



Unsupervised Learning (Clustering Algorithm)

Western Digital.

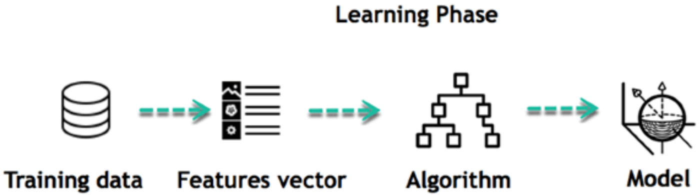
Common Uses: Supervised Learning

Supervised Learning

- Predictive Modeling

Remark

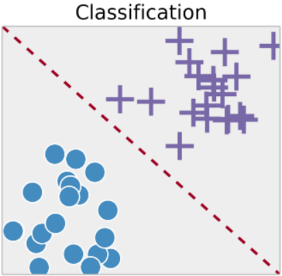
Using known data to predict results



Common Uses: Supervised Learning

Supervised Learning

- Predictive Modeling
- Classification



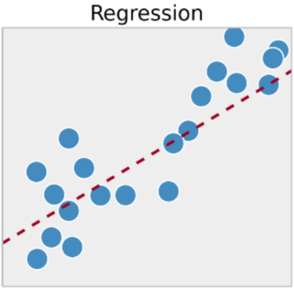
Remark

Classification into known group types based on the features provided to the model

Common Uses: Supervised Learning

Supervised Learning

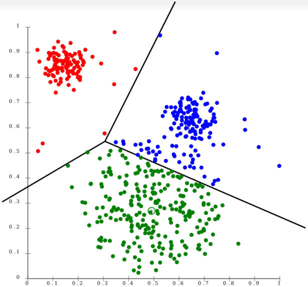
- Predictive Modeling
- Classification
- Regression



Remark

Regression analysis based on the provided data

Common Uses: Unsupervised Learning



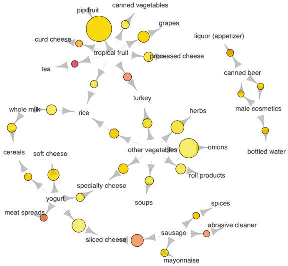
Unsupervised Learning

- Clustering

Remark

Breaking data points into groups on the basis of similar features

Common Uses: Unsupervised Learning



Unsupervised Learning

- Clustering
- Association

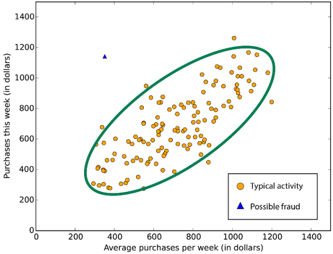
Remark

Determining possible associations between data points

Example: Market Basket Analysis

Determining items frequently purchased together; for customer recommendations, item placement, or supply management

Common Uses: Unsupervised Learning



Unsupervised Learning

- Clustering
- Association
- Anomaly Detection

Remark

Detecting anomalous, unusual, or novel data

Example: Bank Fraud

Using known purchasing patterns (values, locations, times, weather, etc.) to determine probability of a new transaction being fraudulent

Summary

Supervised Learning

Distinctions:

- Labeled Data
- Known Features
- Leverages Experience

Common Uses:

- Predictive Modeling
- Classification
- Regression

Unsupervised Learning

Distinctions:

- Unlabeled Data
- Unknown Features
- Discovers New Patterns

Common Uses:

- Clustering
- Association
- Anomaly Detection

Hastie and Tibshirani

Statistical Learning (Machine Learning)
Statistics as a career:

- The Sexy New Career
- Working for big Tech
- Political and Economic Analysis

Hastie and Tibshirani

Statistical Learning (Machine Learning)
Statistics as a career:

- The Sexy New Career
- Working for big Tech
- Political and Economic Analysis

Statistics
Statistician

Hastie and Tibshirani

Statistical Learning (Machine Learning)

Statistics as a career:

- The Sexy New Career
- Working for big Tech
- Political and Economic Analysis

Statistics → Machine Learning

Statistician

Hastie and Tibshirani

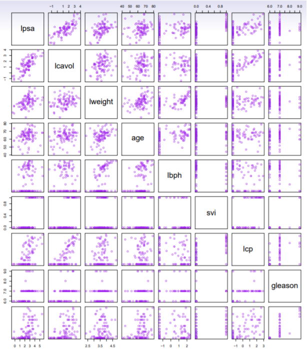
Statistical Learning (Machine Learning)
Statistics as a career:

- The Sexy New Career
- Working for big Tech
- Political and Economic Analysis

Statistics → Machine Learning
Statistician → Data Scientist

Hastie and Tibshirani - Applications of Statistical Learning

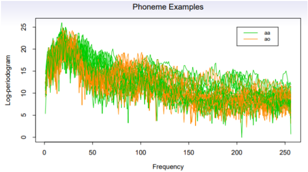
- Health risk factors
- Classify Phonemes
- Myocardial Infarction Prediction
- Email Spam Detection
- Handwritten Number Recognition
- Tissue Oncology Class
- Salary and Demographic relations
- Classify Satellite imagery



Scatter Plot Matrix of Dr. Stamey's Prostate Cancer Research.

Hastie and Tibshirani - Applications of Statistical Learning

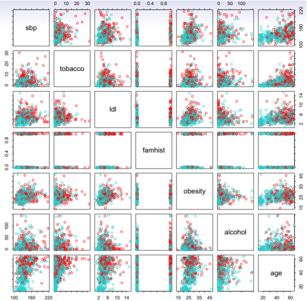
- Health risk factors
- Classify Phonemes
- Myocardial Infarction Prediction
- Email Spam Detection
- Handwritten Number Recognition
- Tissue Oncology Class
- Salary and Demographic relations
- Classify Satellite imagery



AA vs AO Phoneme Log-periodogram.

Hastie and Tibshirani - Applications of Statistical Learning

- Health risk factors
- Classify Phonemes
- Myocardial Infarction Prediction
- Email Spam Detection
- Handwritten Number Recognition
- Tissue Oncology Class
- Salary and Demographic relations
- Classify Satellite imagery



Heart attack risk study in South Africa.

Hastie and Tibshirani - Applications of Statistical Learning

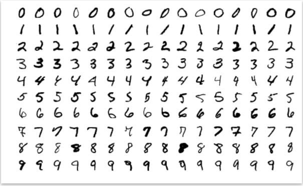
- Health risk factors
- Classify Phonemes
- Myocardial Infarction Prediction
- Email Spam Detection
- Handwritten Number Recognition
- Tissue Oncology Class
- Salary and Demographic relations
- Classify Satellite imagery

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B|A)\Pr(A) + \Pr(B|\neg A)\Pr(\neg A)}$$

Spam detection is often done using some form of Bayesian analysis.

Hastie and Tibshirani - Applications of Statistical Learning

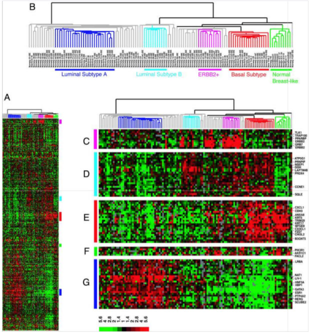
- Health risk factors
- Classify Phonemes
- Myocardial Infarction Prediction
- Email Spam Detection
- Handwritten Number Recognition
- Tissue Oncology Class
- Salary and Demographic relations
- Classify Satellite imagery



The MNIST data set is often used as a practice problem for students to develop a model capable of reading hand-written numbers.

Hastie and Tibshirani - Applications of Statistical Learning

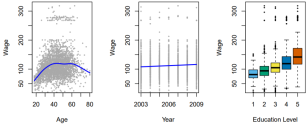
- Health risk factors
- Classify Phonemes
- Myocardial Infarction Prediction
- Email Spam Detection
- Handwritten Number Recognition
- Tissue Oncology Class
- Salary and Demographic relations
- Classify Satellite imagery



Gene expression data used to classify oncological class of histological samples.

Hastie and Tibshirani - Applications of Statistical Learning

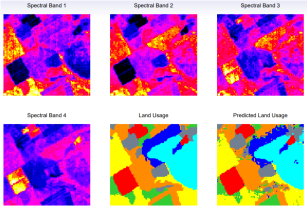
- Health risk factors
- Classify Phonemes
- Myocardial Infarction Prediction
- Email Spam Detection
- Handwritten Number Recognition
- Tissue Oncology Class
- Salary and Demographic relations
- Classify Satellite imagery



Examining factors contributing to income levels in the central Atlantic demographic region.

Hastie and Tibshirani - Applications of Statistical Learning

- Health risk factors
- Classify Phonemes
- Myocardial Infarction Prediction
- Email Spam Detection
- Handwritten Number Recognition
- Tissue Oncology Class
- Salary and Demographic relations
- Classify Satellite imagery



Classification of geographic features shown in satellite imagery (Taken in Southern Australia).

Image Credits

Guru99.com
nvidia.com
Introduction Lecture by Hastie and Tibshirani