Introduction to concepts, theories, and algorithms for pattern recognition and machine learning.

Pre-requisites
- Linear Algebra
- Calculus
- Probability Theory
- Algorithms


Classic Book: Duda, Hart, Stork. “Pattern Classification”

Statistical Perspective: Hastie, Tibshirani, Friedman. “Elements of Statistical Learning” (2nd Ed)

Advanced: Bishop. “Pattern Recognition and Machine Intelligence”

Recent: Murphy. “Machine Learning a Probabilistic Perspective”
Chp. 1. Introduction

Why Machine Learning? Big Data!

Alpaydin.

Data Mining.
ability to store vast amounts of data.

Need to understand regularities in the data. Not perfect understanding.

Good and Useful Approximations.

Examples:

Financial - Credit / Fraud / Stock Market.
Manufacturing - Optimization / Troubleshoot / Control.
Medical - Medical Diagnosis.
Telecommunication - Network optimization / service.
Science - Data Physics, Biology.
Web: Search / Analysis.
Machine Learning. (in practice) programming computers to optimize a performance criterion using example data or past experience.

A model defined up to some parameters. Learning is the execution of a computer program to optimize these parameters using training data (past experience). The model may be predictive to make predictions in future or descriptive to gain knowledge from data.

Machine Learning involves

Statistics \(\rightarrow\) build mathematical models with uncertainty. Make inferences from sample

Computer Science \(\rightarrow\) efficient algorithms for optimization problems of learning, storing and process data.

After learning \(\rightarrow\) algorithms for inference, storage.

Mathematics \(\rightarrow\) optimization/ geometric formulations.
Examples of Machine Learning Applications

Learning Associations

Retail: Basket Analysis

Find associations between products bought by customers.

If people who buy X typically also buy Y — then client who buys X is a potential customer for Y.

Want an association rule

Conditional probability \( P(Y|X) \).

E.g. \( P(\text{chips} | \text{beer}) = 0.7 \)

Then 70% customers who buy beer will also buy chips.

More advanced — make distinctions between customers

\( P(Y | X, D) \)

\( D \) - customer attributes

e.g. gender, age, marital status.

Bookseller - products are books or authors

Web Portal - links to webpages, what links will user click, download pages in advance.
Classification

Credit Scoring.

Bank lends money at interest.

What risk is associated with loan?

What probability that customer will fail/default to pay back all/part of the money.

Credit Scoring — bank calculates the risk given the amount of credit and information about customer. Attributes:

- Income, savings, collateral
- Profession, age, financial history

Bank has records of past loans including defaults.

Bank wants to infer a general rule coding the association between customer attributes and his risk.

Classification problem — two classes:

(a) low-risk, (b) high-risk.

Information about customer’s attributes are input to a classifier whose task is to assign the customer to one of these two classes.
(6) Classification (cont)

Θ be data instances

+" classified as low-risk

-" classified as high-risk.

Example: two attributes (for simplicity)

Savings, Income

Example classification rule:

IF income > Θ₁ AND Savings > Θ₂

THEN low-risk

ELSE high-risk.

This is an example of a discriminant function that separates the data into two classes.

Main application is prediction. If future is similar to the past, then we can make prediction for novel instances

In some cases, instead of 0/1 decision, we may want to calculate prob. P(Y|X) (i.e., learn an association).
Pattern Recognition Examples:

- Optical Character Recognition (OCR)
- recognize character codes from images
- variability in writing styles
- explicit redundancy of language - successive characters are not independent, but are constrained by the nature of the language.

Medical Diagnosis

- inputs are relevant information about the patient and the classes are the illnesses.

Input - Age, gender, medical history, symptoms

Some tests may not have been applied and these inputs are missing.

Tests are expensive, take time and are only cost to apply them to gain valuable info. Wrong decision is very bad - classifier must take this into account.

speech recognition.
Knowledge Extraction.

Learning a rule from data also gives knowledge extraction.

The rule is a simple model that explains the data — looking at this model gives an explanation for the process underlying the data.

This knowledge can be used — e.g. to advertise to low-risk customers for bank loans.

Learning also performs compression, since we get an explanation that is simpler than the data. It requires less memory to store and less computation to process.

(If you knew the law of addition, you don't need to remember the sum of all possible pairs of numbers.)

Outlier detection — find instances which do not obey the rule and are exceptions.

→ E.g. to detect anomalies requiring attention (e.g. fraud).
Task: Classify fish.

Sea Bass
Salmon

What features to use?
choices: length of fish
         width of fish
         brightness (dark/bright)
         texture
         shape of head

Use length and brightness.

Training Data
Examples
( (length_i, brightness_i); i=1 to n) Sea Bass
( (length_i, brightness_i); i=1 to n) Salmon.

Want simple rule to discriminate between salmon and sea bass.

Linear classifier:
sea bass on one side of the line, salmon on the other.
Memorization and Generalization.

We want to learn a classifier that works on data we have not seen yet.

Suppose our training set contains only three examples:

- salmon
- brightness
- length
- sea bass
- sea bass
- sea bass

Many possible linear classifiers:

$$\begin{array}{l}
1, 2, 3
\end{array}$$

But these three classifiers will not generalize to new data.

How to classify new data?

\(1\) says \(\textit{is sea bass}\)

\(2\) says \(\textit{is sea bass}\)

\(3\) says \(\textit{is salmon}\)

Which is right?

Answer: we do not know. We do not have enough training data to learn the classifier. We need more data.

Memorization: All three classifiers \(1, 2, 3\) can classify the training data (i.e. memorize it).

But we want a classifier to predict and correctly classify new data. To generalize to new data.
(1) The nearest neighbor classifier.

Suppose we have training data.

We cannot find a linear classifier that separates the + and - examples.

Nearest neighbor classifiers a new example by the nearest examples. E.g.

This lecture has given examples of three main classification methods:

(1) decision trees,

(2) linear classifier,

(3) nearest neighbor.

Many advanced machine learning techniques are based on these three simple methods.

Also, this lecture has made the distinction between memorizing the training data and generalizing to predict/classify new data.

Machine learning must generalize. This involves a trade-off between the complexity of the classifier and the amount of training data.

If limited training data, then only generalize if you use a simple classifier.
Key Points:
Data $\rightarrow$ want to learn a classifier (or a more complicated decision later in course).

Data: $\{(x_i, y_i) : i = 1, n\}$ features e.g. income/savings
$y_i$ classifier, e.g. high-risk, low-risk

Three classic types of classifier rules:
(i) linear classifier

(ii) nearest neighbor classifier.
Classify $\hat{\tau}$ by majority vote of neighbors. i.e., by $-$ and $+$ respectively.

Savvy (iii) Decision Trees - Game of Twenty Questions

- $-$ $+$ $+$ $+$
- $+$ $+$ $+$ $+$
- $-$ $-$ $-$ $-$

You allowed to ask a sequence of questions?
E.g. is income $> 81$, then high-risk
is savings $> 82$, then high-risk

Strategy: Ask question (1): income

This question classifies some examples correctly, but has some errors

So follow-up with question (2) saying is savings $> 82$

These two questions classify all examples correctly.