# ① Introduction to Machine Learning   Spring 2014

Note Title   9/4/2012

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

Pre-requisities
- Linear Algebra
- Calculus
- Probability Theory
- Algorithms.
- Geometry

**Books:** Alpaydin. "Introduction to Machine Learning." (2nd Ed)

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."

(2) Note Title 3/26/2008

# Introduction

Why Machine Learning? Big Data!

There is an enormous amount of data
We need to store it, analyze it, understand it
and exploit it.

Where does this data come from?

There are many sources:

Financial – Credit / Fraud / Stock Market.
Manufacturing – Optimization / Troubleshoot / Control.
Medical – Medical Diagnosis.
Telecomunication – Network optimization / service.
Science – Data Physics, Biology.
Web – Search, Analysis.
A.I – Vision, language, robotics.

(3) What does Machine Learning involve?

- Training Data
- A Model with paramaters.

Learning is performed by applying an algorithm to the training data to estimate the parameters.

We want this model to be:

(i) predictive, to make predictions on new data.

(ii) descriptive, to efficiently describe the data.

and, if possible, to give understanding of the data and perform knowledge extraction.

Machine Learning is interdisciplinary. It uses techniques from several disciplines

Statistics – probabilites, modeling uncertainty, make inference from samples.

Computer Science – algorithms for learning, data structures.

Mathematics – optimization, geometry, analysis.

Engineering –

and others.

## (4) Examples of Machine Learning.

Learning Associations between products bought by customers.

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

Want a probabitistic association:

conditional probability $P(Y|X)$ (learnt from data)

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

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

More advanced, make distinctions between customers.

$P(Y|X, D)$

D - customer attribute, e.g. gender, age, marital status.

Also applies to buying books online,

This illustrates the probabilistic approach to machine learing. Requires learning probabilities.

An alternative approach is classification: learn a decision rule to classify data.

(5) <u>Classification Example</u>: credit scoring.

Bank lends money at interest.
What risk is associated with a bank loan?
Which <u>types</u> of customers have high probability of paying back the loan? Low-Risk customers.
Which <u>types</u> of customers have low probability. High-Risk customers.

Bank wants a rule to classify customers as high-risk or low-risk based on data from previous customers

We assume that the data consists of

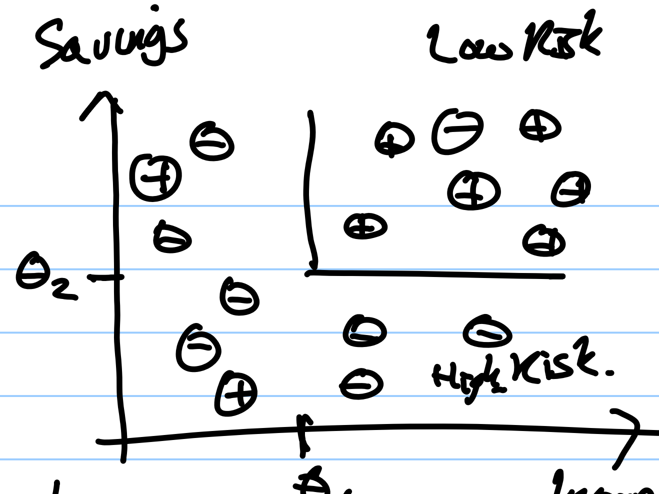
- customer savings $x_1$
- customer income $x_2$
- customer pay back/default $y$

$x_1, x_2$ are continuous numbers,
$y$ is binary, $y = 1$, customer paid back. low-risk
$y = -1$, customer defaulted high-risk

In practice, banks will consider other customer attributes: age, criminal record, college major.

<u>Data</u>: $\{ (x_1^\mu, x_2^\mu, y^\mu) : \mu = 1 \text{ to } N \}$
N customers.

## (6) Classification




⊕ positive examples $y=1$, low-risk.

⊖ negative examples $y=-1$, high-risk.

Example of classification rule:

IF income $> \theta_1$
AND savings $> \theta_2$
THEN low-risk
ELSE high-risk.

Note: this classifier will make some mistakes, but not many. It is typically impossible to find a classifier which makes no mistakes.

This is an example of a decision tree classifier. This one of the three classic machine learning methods.

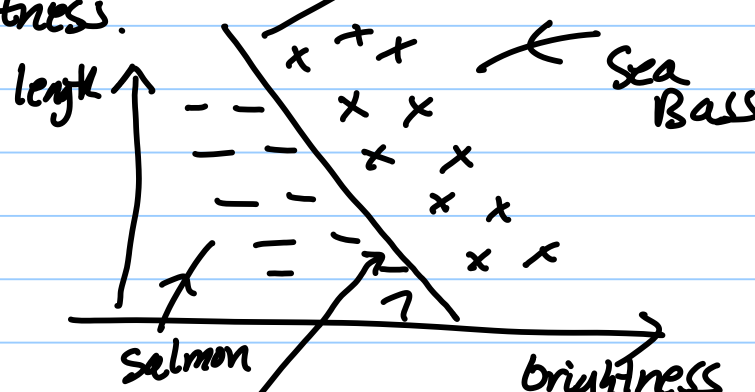
Note: We assume that this classifier — learnt on past data — will be able to predict future data.
Key assumption — the future is similar to the past. Note: not always true, financial crises, wars, etc.

## (7) Classification. Linear Classifier

Task: classify fish as salmon or sea bass

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

Use length and brightness.




Training Data

$\{(x_1^\mu, x_2^\mu, y^\mu) : \mu = 1 \text{ to } N\}$

$y = +1$, sea bass
$y = -1$, salmon
$x_1$ - brightness, $x_2$ - length.

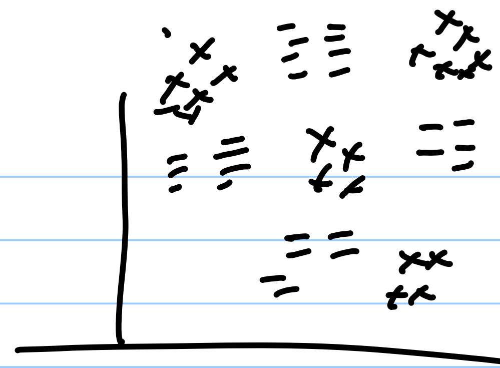
Want simple rule to discriminate between salmon and sea bass.
linear classifier sea bass on one side
salmon on the other.

Note: linear classifier / perceptron is another of the three classic machine learning methods.

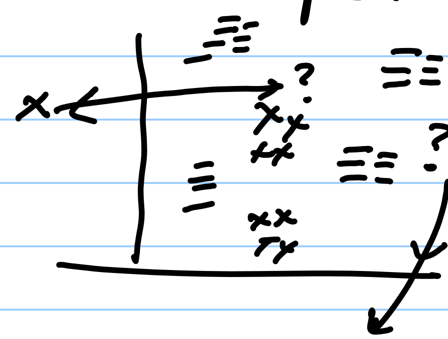
The third classic machine learning method is the nearest neighbor classifier.

## (8) Nearest neighbor classifier

Suppose we have training data.
we cannot find a linear classifier that separates the ++ and -- examples.



Nearest neighbor classifies a new example ? by the nearest examples.



Note: the three classic methods tend to be good for different types of data. But nearest neighbor is very good in general.

We will discuss these methods, and many others, in the rest of the course.

Now we introduce a key concept in Machine Learning. The difference between

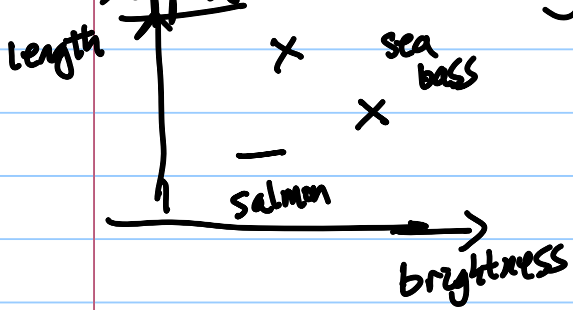
Memorization: finding a classifier that gives good results on the training data.

Generalization: finding a classifier that gives good results on data that we havn't seen. Good prediction.

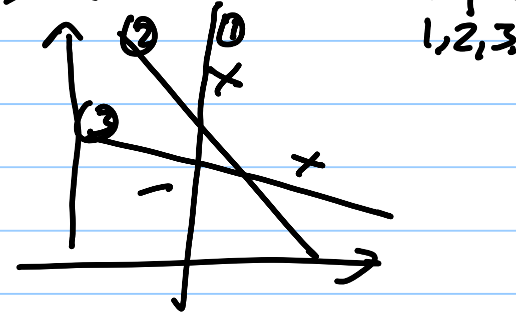
(9) Memorizaton and Generalization.

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

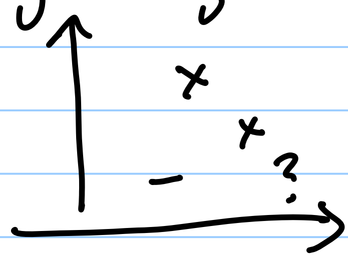
Suppose our training set contains only three example




Many possible linear classifiers




But these three classifiers will not generalize to new data.




How to classify new data?

(1) says ? is sea bass

(2) says ? is sea bass

(3) says ? 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)

Key Factors: amount of training data.
complexity of classifiers.

Need to ensure that complexity of classifiers $\ll$ amount of data.

(10)

## Other Issues

Much of Machine Learning involves learning classifiers, or probabilities.

But we may also want to perform:

Knowledge Extraction — use algorithms to understand the structure of the data.

Compression — learn simple ways to describe the data.

Outlier Detection — find instances which do not obey the rules, which are outliers.

These may signal the onset of fundamental changes — financial crises, wars, ..

They may also signal events like fraud.

Another key issue is the curse of dimensionality.

Most machine learning problem involve high-dimensional data. E.g. $(x_1 \ldots x_{100}, y)$ not $(x_1, x_2, y)$ (as in our examples)

Problem: • our geometric intuitions are bad in high-dimension.

• classifiers in high-dimensions may require an enormous amount of data.

(11) Key Points:

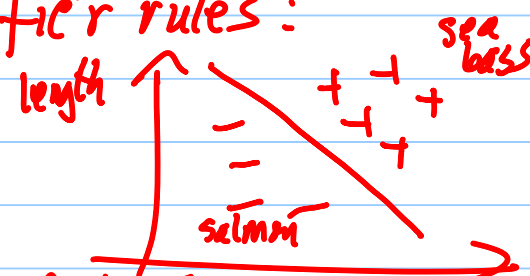
Data → wait to learn a classifier (or a more complicated decision - later in course.).

Data: $\{(\underline{x}_i, y_i) : i = 1 \text{ to } N\}$ $\underline{x}_i$: features e.g. income/savings

$y_i$: classifier, e.g. high-risk, low-risk

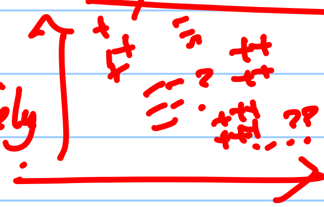
Three Classic types of classifier rules:

(i) linear classifier

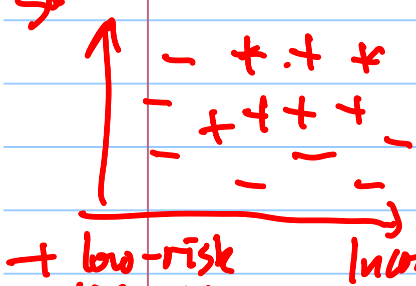



(ii) nearest neighbor classifier.

Classify ? and ?? by majority vote of neighbors. i.e. by − and + respectively




Savings (iii) Decision Trees _ Game of Twenty Questions




+ low-risk

− high-risk

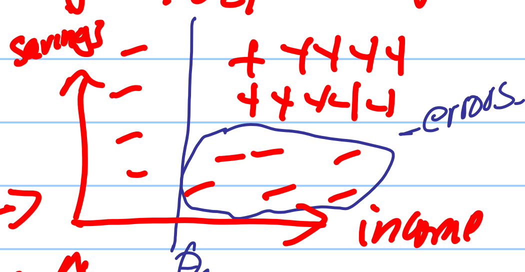
You allowed to ask a sequence of questions?

E.G. is income $> \theta_1$, then high-risk

is savings $> \theta_2$, then high-risk

Strategy: Ask question (1).

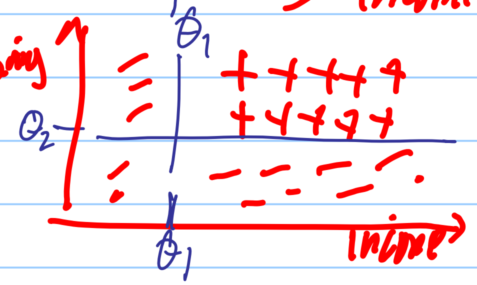
is income $> \theta_1$




This question classifies some examples correctly, but has some errors

So follow-up with question (2) saving

is savings $> \theta_2$




These two questions classify all examples correctly.