① Introduction to Machine Learning. Spring 2013

Note Title 9/4/2012

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

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

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) # Introduction 2013 Spring

Note Title 3/26/2008

Chp 1. 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.
- Medicine – Medical Diagnosis.
- Telecomunication – Network optimization / service.
- Science – Data Physics, Biology.
- Web – Search, Analysis.
- A.I – Vision, language, robotics.

Spring 2013

(3) <u>Machine Learning</u>. (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 experiences.

The model may be
<u>predictive</u>. to make predictions in future.
or <u>desriptive</u> to gain knowledge from data.

Machine Learning involves
<u>Statistics</u> → build mathematical models with uncertainty. Make inferences from samples
<u>Computer Science</u> → efficient algorithms for optimization problems of learning., storing and process data.
After learning → algorithms for inference, storage.
<u>Mathematics</u> → optimization, geometric formulations.

Spring 2013

(4) 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.

Spring 2013

(5) Classification

Credit Scoring.

Bank loans 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. Attribute
Information — income, savings, collateral,
profession, age, financial history.

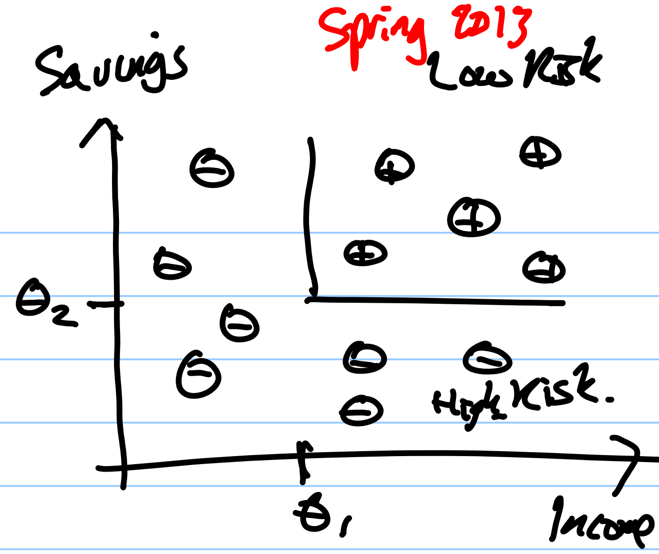
Bank has records of past loans including defaults

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

Classification problem — two classes:
(a) low-risk, (b) high-risk.

Information about customer's attributes are
input to a classier whose task is to assign
the customer to one of these two classes.

Spring 2013



(6) Classification (cont)

⊕ & ⊖ are data instances
"+" classified as low risk
"−" classified as high risk.

Example: two attributes (for simplicity)
savings, income

Example classification rule:
IF income > θ_1 AND savings > θ_2
THEN low-risk
ELSE high-risk.

An example of decision trees

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).

Spring 2013

(7) Pattern Recognition Examples:

Optical Character Recognition (OCR)

recognize character codes from images.

variability in writing styles

exploit redundancy of language - successive characters are not independent but are constrained by the words of the language.

Medical Diagnosis

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

Inputs - Age, gender, medical history, symptoms

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

Tests are expensive, take time and we only want to apply them to gain valuable info

Wrong decision is very bad ~ classifier must take this into account.

Speech recognition.

(8) Spring 2013

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 know the law of addition, you don't need to remember the sum of all possible pairs of numbers)

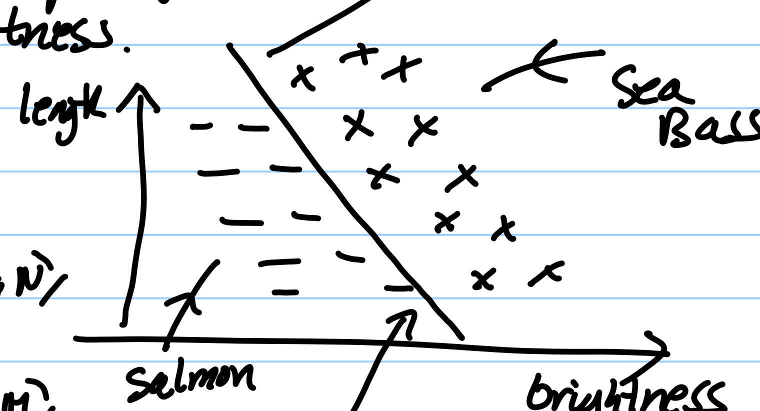
Outlier detection - find instances which do not obey the rule and are exceptions.

-> e.g. to detect anomalies requiring attention (e.g. fraud).

(9) **Task:** Classify fish. Spring 2013

Sea Bass
Salman
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
$\{(\text{length}_i, \text{brightness}_i) : i = 1 \text{ to } N\}$, Sea Bass
$\{(\text{length}_i, \text{brightness}_i) : i = 1 \text{ to } M\}$, Salman.

Want simple rule to discriminate between salmon and sea bass.

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

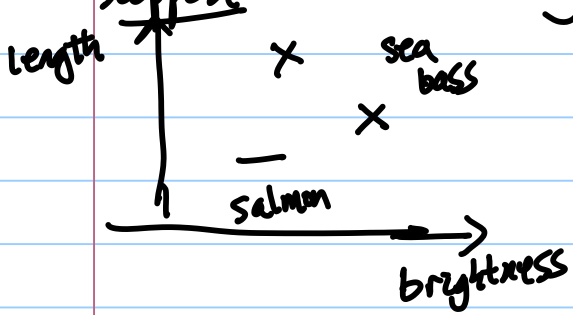
(10)

Spring 2013

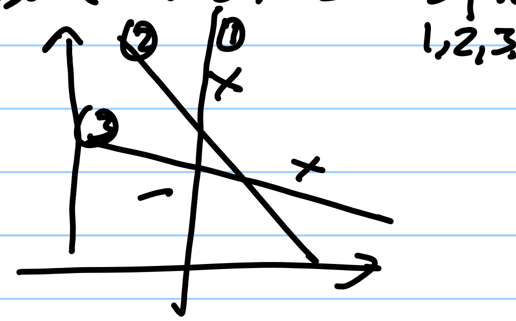
Memorization and Generalization.

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

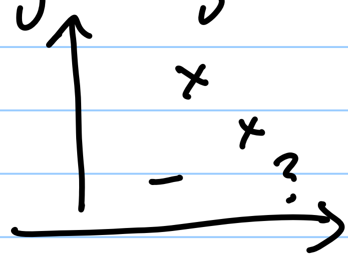
Suppose our training set contains only three examples



Many possible linear classifiers 1,2,3



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?

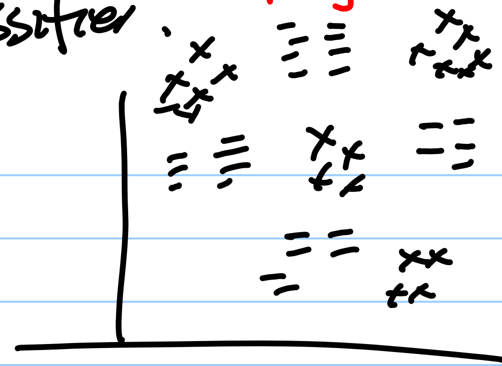
Answer : we do not know. We do not have enough training data to learn the classifier. We need more data.

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

But we want a classifier to predict and correctly classify new data. To generalize to new data.

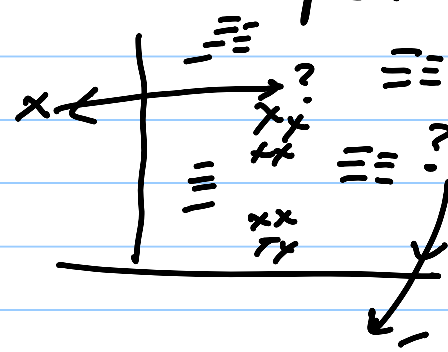
Spring 2013

(11) The 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. 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.

(12) Key Points: Spring 2013

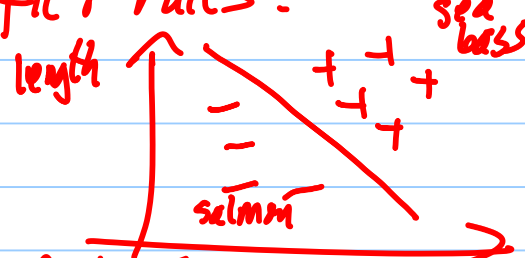
Data → want to learn a classifier (or a more complicated decision - later in course.).

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

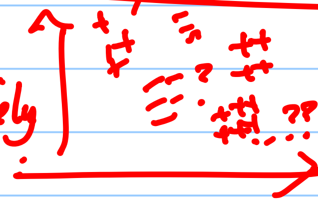
Three Classic types of classifier rules :

(i) linear classifier

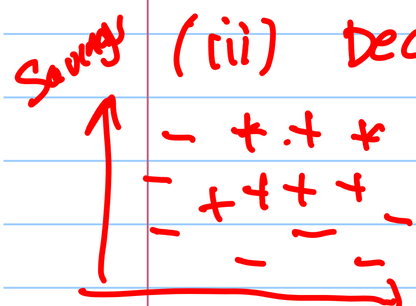


(ii) nearest neighbor classifier.

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



(iii) Decision Trees _ Game of Twenty Questions



+ low-risk

− high-risk

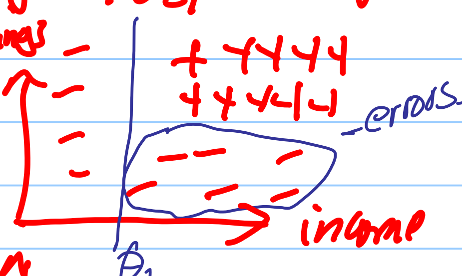
You allowed to ask a sequence of questions?

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) is savings $> \theta_2$

These two questions classify all examples correctly.