

EBUS537: Data Mining & Machine Learning -- Overview

Prof. Dongping Song

University of Liverpool Management School

Email: Dongping.song@liv.ac.uk

Module Introduction

- Module aim
 - understand and apply the concepts, theories, techniques and developments associated with DM and ML
- Learning and teaching strategies
 - Lectures + discussions + 126 hours independent learning
- Assessment (100% coursework)
- **Learning outcomes**
- **Planned schedule**
- **Readings**

Module Learning Outcomes

1. “Gain an in depth **knowledge and principles** in the areas of data mining, machine learning, and data warehousing”
2. “Critically assess the **strengths and weaknesses** of various data mining and machine learning techniques from a practitioner/ user perspective”
3. “Be able to **identify, formulate and solve problems** arising from practical applications using data mining and machine learning principles and techniques”

- Focus on DM & ML overall process, key concepts, techniques, and their applications in the management context
- Not about the use of a specific software tool

Note: You can find the module handbook in the VITAL (in the folder ‘About This Module’).

Planned Schedule

- Data mining and machine learning
- Data-related issues
- Data exploration
- Classification
- Association learning
- Clustering
- Alternative classification/ Data warehousing

Recommended Readings

- Tan, P.N., Steinbach, M., and Kumar, V. (2014). Introduction to Data Mining, Pearson.
- Alpaydın, E. (2010). Introduction to Machine Learning, 2ed, The MIT Press. (online)
- Dean J. (2014), Big data, data mining, and machine learning: value creation for business leaders and practitioners, Wiley. (online)
- Han, J., Kamber, M. & Pei, J., (2011). Data Mining: Concepts and Techniques. 3rd ed. Burlington: Elsevier Science. (online)
- Prabhu, S. (2007), Data Mining and Warehousing, New Age International. (online)
- Giudici, P. & Figini, S. (2009), Applied Data Mining for Business and Industry, 2nd Edition, Wiley.

Assignment & Report Writing

- Assignment
- Avoid **plagiarism**!
 - If you use other people's **words** you must put them in quotation marks and refer to the source.
 - If you use other people's **idea** you must acknowledge it and refer to the source.
- Use Harvard referencing style: Name (YEAR).
- Do not use bibliography.

ELC's Academic Language & Skills Support for MSc Business Analytics & Big Data students
8 (x 2hrs) sessions from Week 1 (W/C 28th Jan 2019).

Examples: Power of Data Analysis

- Business, politics, sports and academia have increasingly placed in the **power of data**.
- It is a behind-the-scenes technology that quietly drives the **value of data** and **the potential** to mine it for cost-saving and profit-making insights.
- Examples:
 - **Microsoft** is paying \$26 billion for LinkedIn largely for its database of personal profiles and business connections on more than 400 million people.
 - **General Electric** is betting big that data-generating sensors and software can increase the efficiency and profitability of its jet engines and other machinery.

Analysing big data can accurately predict events and human behaviours

Example: Did Big Data get wrong?

- An interesting example is the American president election in 2016
- Virtually all the major vote forecasters, including Nate Silver's **FiveThirtyEight** site, The New York Times **Upshot** and the **Princeton Election Consortium**, put Mrs. Clinton's chances of winning in the 70% to 90% range.
- Donald Trump won on 8/11/2016 (Tue)
- Did big data get wrong?



Source: PredictWise, FiveThirtyEight, The UpShot (New York Times) on 8/11/2016 (Tue)

Discuss: Did Big Data get wrong?

Learning Outcomes

- Understand data mining and machine learning
- Understand the CRISP-DM Process Model
- Appreciate their relationships with statistics, PR, database, AI
- Understand main data mining techniques
- Discuss applications of data mining and machine learning
- Introduce relevant concepts in supervised learning
- Appreciate noise and model complexity
- Explain model selection and generalization
- Understand main decisions of a supervised learner

Data and Big Data

- Volume:
- Velocity:
- Variety:
- Veracity

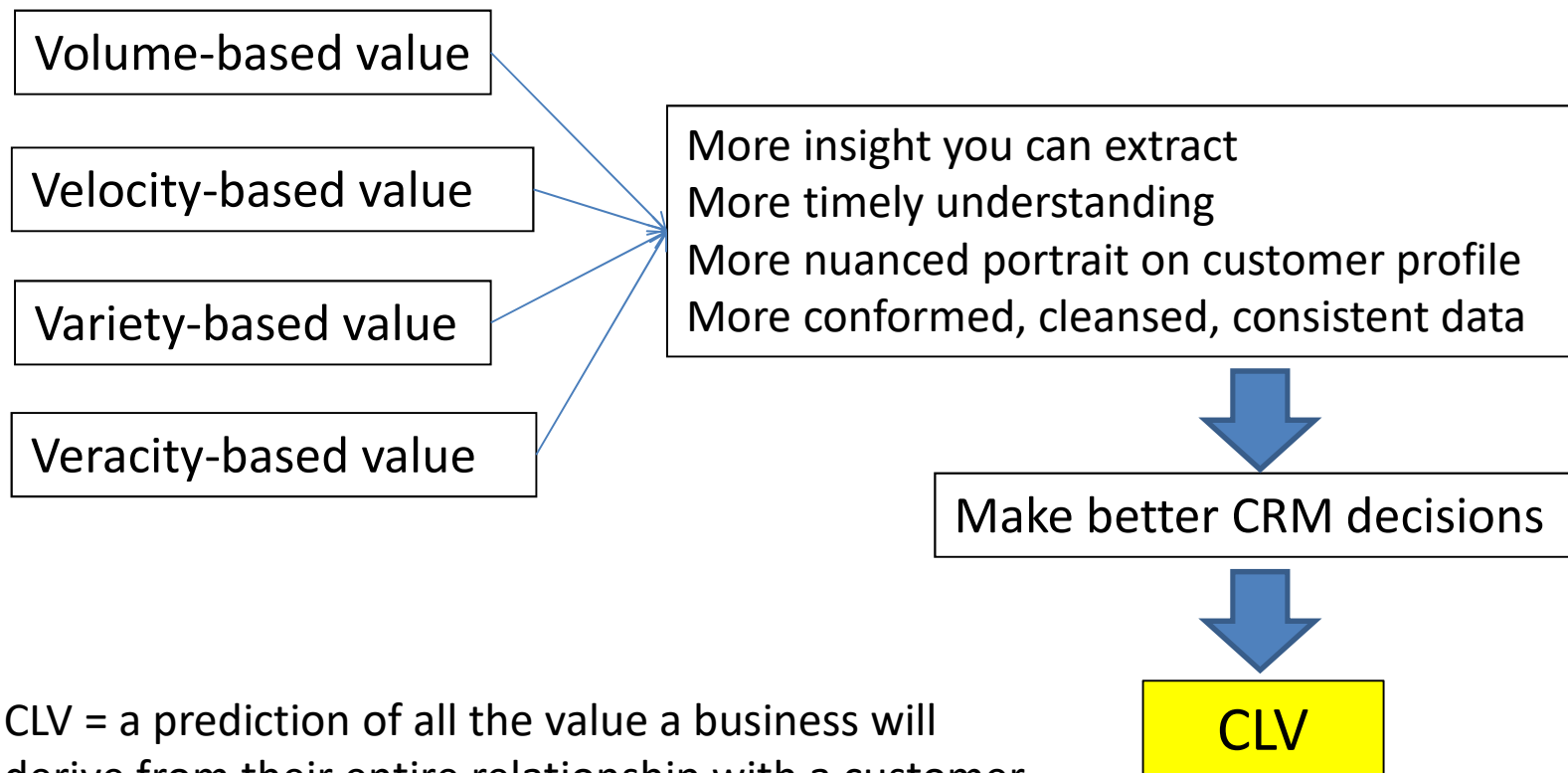
- Computer technologies →
 - Store and process large amounts of data
 - Access it from physically distant locations



- Most data acquisition devices are digital and reliable, e.g.
 - The **point of sale terminals** record the details of each transaction: date, customer info, goods bought and their amount, total money spent, etc.
 - This typically amounts to gigabytes of data every day.

Business Value of Big Data

- To measure the **customer lifetime value** (CLV) impact of big data used for customer relationship management (CRM):



Usefulness of Data → Technique

- What the supermarket chain wants is to be able to **predict** who are the likely customers for a product → need **algorithms** to make prediction → make better decisions.
- However, the algorithm for this is not evident;
- It changes in time and by geographic location.

The stored data becomes **useful** only when it is analyzed and turned into information that we can make use of.

"Big data isn't about the data. It's about analytics" – Prof Gary King, Harvard University

Data and Algorithms

- An **algorithm** is a *sequence of instructions* that should be carried out to transform the input to output, e.g. sorting.
- For some tasks, we have data but do not have an algorithm, e.g. to tell spam emails from legitimate emails.
- We want to “learn” what constitutes spam from them, i.e., we would like the computer (machine) to **extract automatically the algorithm** for this task.

There are many applications for which we **do not have an algorithm** but do have example data → machine learning

Need of Machine Learning & Data Mining

More examples: which people are likely to buy this ice cream flavor, or the next book of this author, or see this new movie, or visit this city, or click this link.

- We hope to extract the answers from data
- We believe there is a process that explains the data
- People do not go to supermarkets and buy things at random. There are certain patterns in the data
- We can construct a good and useful approximation, detect certain patterns or regularities
- This leads to **data mining** and **machine learning**.

Data Mining

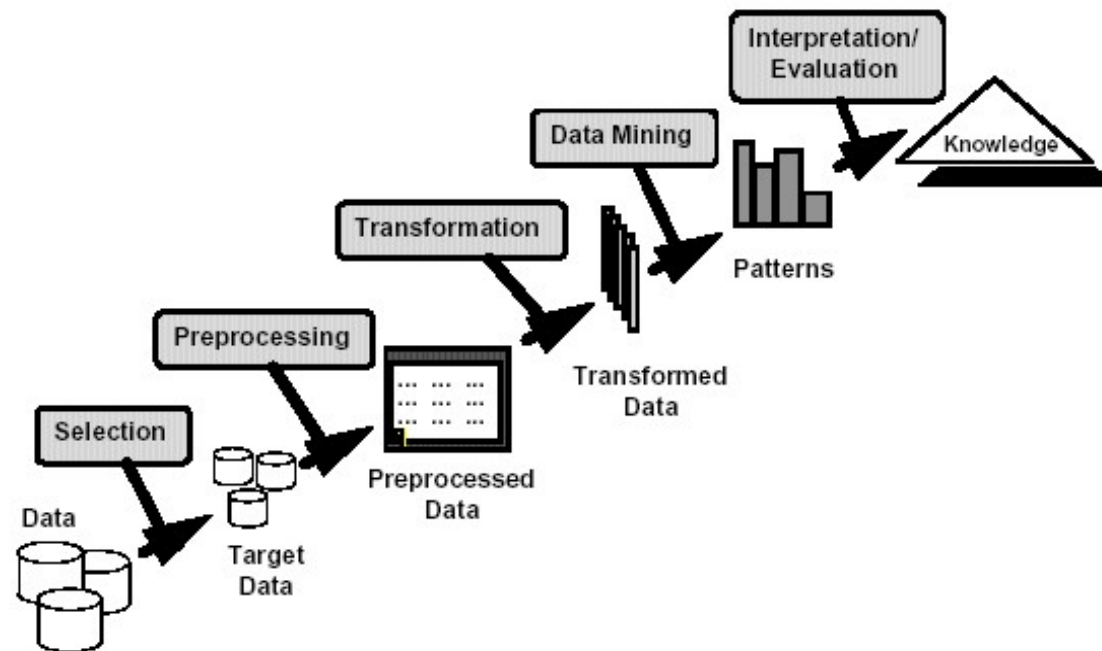
- **Data mining**: application of machine learning methods to extract value information from data
- Similarity to mining
- Application areas
 - **Retail**: Market basket analysis, CRM
 - **Finance**: Credit scoring, fraud detection
 - **Manufacturing**: Control, robotics, troubleshooting
 - **Medicine**: Medical diagnosis
 - **Telecommunications**: Spam filters, intrusion detection
 - **Web mining**: Search engines

Data Mining Definitions

- DM is the **non-trivial extraction** of implicit, previously unknown and potentially **useful information** from data
- DM is the **exploration & analysis**, by automatic or semi-automatic means, of large quantities of data in order to discover **meaningful patterns**
- DM is the search for **relationships** and global **patterns** that exist in large databases but are '**hidden**' among the vast amount of data
- DM is using a variety of **techniques** to identify nuggets of information or decision-making **knowledge** in bodies of data, and extracting these in such a way that they can be put to use in the areas such as decision support, prediction, forecasting and estimation.

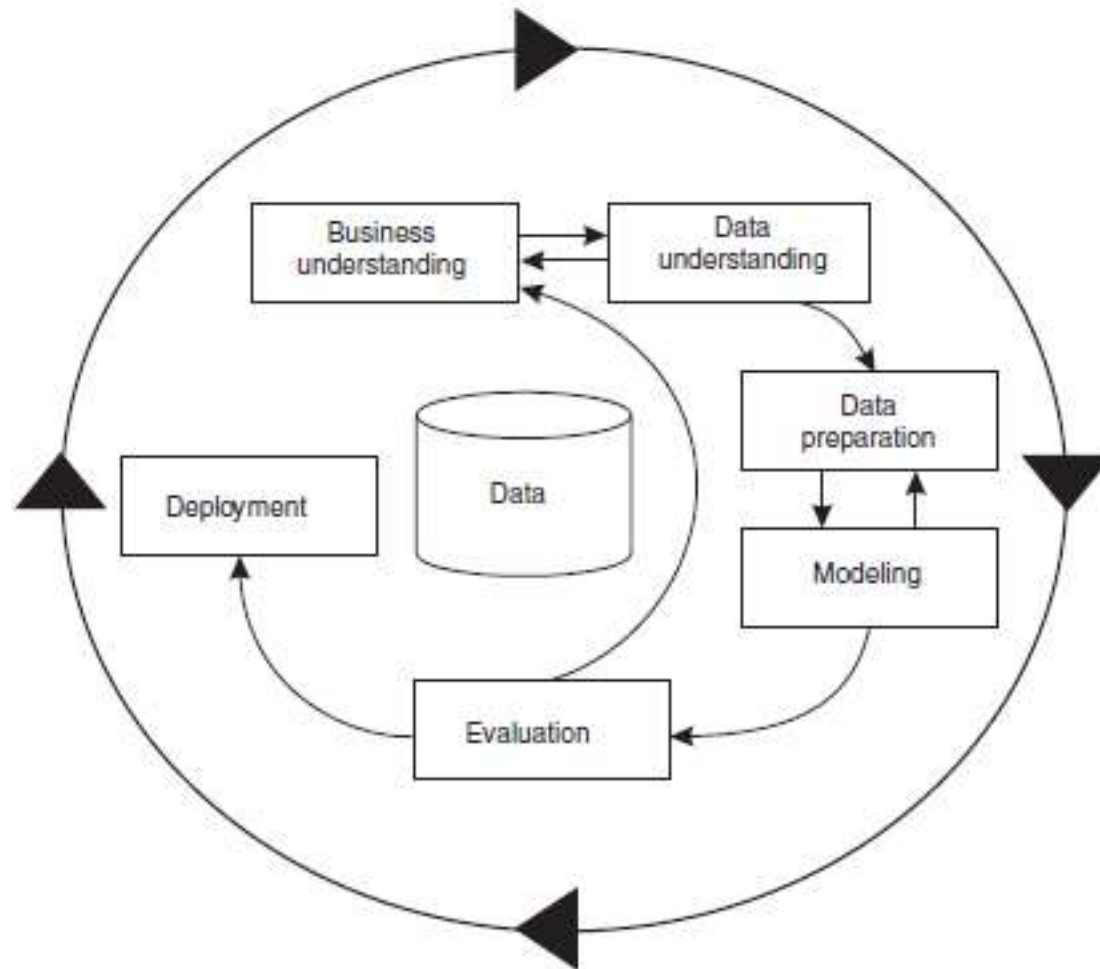
Overall Process of Data Mining

- Broadly, data mining is not just a data analysis, it is a **process** involving many activities to discover knowledge



Extraction of interesting information or patterns (knowledge) from data in large databases

The CRISP-DM Process Model



CRISP=Cross-Industry Standard Process

Mariscal et al (2010). A survey of data mining and knowledge discovery process models and methodologies, Knowledge Engineering Review, 25, 137-166.

The CRISP-DM Process Model

1. Business understanding
2. Data understanding
3. Data preparation
4. Modeling
5. Evaluation
6. Deployment

Machine Learning

- **Machine learning** is not just a database problem; it is also a part of artificial intelligence
- A system that is in a changing environment should have the ability to learn → adapt to such changes
- **Machine learning** is programming computers to optimize a performance criterion using example data or past experience

We have a model defined up to some parameters, and learning is the execution of a computer program to optimize the parameters of the model using the training data or past experience. The model may be predictive to make **predictions** in the future, or descriptive to gain **knowledge** from data, or both.

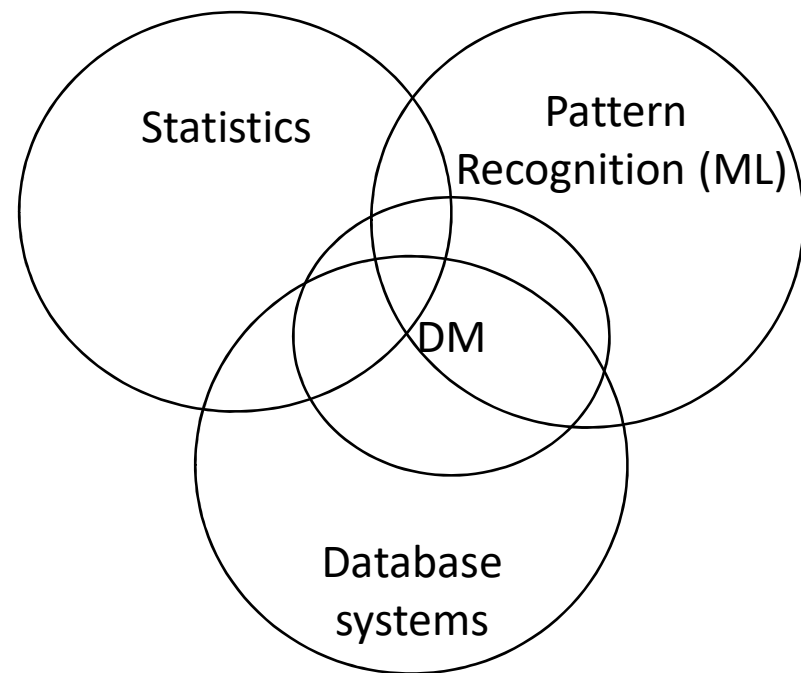
Machine Learning

- There is no need to “learn” to calculate payroll
- Learning is used when:
 - Human expertise does not exist (navigating on Mars),
 - Humans are unable to explain their expertise (speech recognition)
 - Solution changes in time (routing on a computer network)
 - Solution needs to be adapted to particular cases (user biometrics, e.g. fingerprint verification)

ML is a study that gives computers the ability to learn without being explicitly programmed.

Data Mining & Machine Learning

- DM vs ML?
- DM and ML are closely related to statistics, pattern recognition, database systems, AI, DS
- Traditional techniques may be unsuitable due to
 - Enormity of data
 - High dimensionality of data
 - Heterogeneous, distributed nature of data



Data Mining Tasks

- **Predictive** tasks
 - Use some variables to predict unknown or future values of other variables.
- **Descriptive** tasks
 - Find human-interpretable patterns that describe the data.

Correlations, trends,
clusters, and anomalies

ML and DM Techniques

- Supervised Learning
 - Unsupervised Learning
 - Semi-supervised Learning
-
- Association
 - Clustering
 - Classification
 - Regression
 - Reinforcement Learning

Supervised Learning

- The aim of supervised learning is to **map the input to an output** whose correct values are provided by a supervisor.

Learn a mapping: $X \rightarrow Y$; or $Y = f(X)$

- The process of an algorithm learning from the training dataset can be thought of as a teacher supervising the learning process
- Divided into two groups: **regression** and **classification**.
- **Popular examples**: Linear regression; Decision tree; Random forest; Rule-based classifier; Bayes classifier; Artificial neural network; Support vector machines

Unsupervised Learning

- The aim of unsupervised learning is to find the **regularities** and **structure** in the input. You only have input data and no output variable
- There is no correct answers. Algorithms are left to their own devices to discover and present the interesting structure in the data.
- Can be grouped into: **association** and **clustering** problems.
- **Popular algorithms**: k-means and Apriori algorithm

Certain patterns occur more often than others;
Correlations, trends, clusters, and anomalies;

Semi-Supervised Learning

- **Semi-supervised learning** problems: we have a large amount of input data (X) and only some of the data is labeled (Y)
- In real world, expensive or time-consuming to label data
- Use unsupervised learning techniques to discover and learn the structure in the input variables.
- Use supervised learning techniques to make best guess predictions for the unlabeled data, as training data to build a predictive model

Association: Definition

- **Definition:** Association rule learning is to discover **interesting relations** between seemingly unrelated data in a database (e.g. transactional records)
- Mathematically, finding an association rule is to learn a **conditional probability**.

$P(Y | X)$ = probability that somebody who buys X also buys Y , where X and Y are products/services.

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

Association Rule Discovery: Definition

- Given a set of records each of which contain some number of items from a given collection;
 - Produce **dependency rules** which will predict **occurrence** of an item based on occurrences of other items.

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:

$\{\text{Milk}\} \rightarrow \{\text{Coke}\}$

$\{\text{Milk}\} \rightarrow \{\text{Beer}\}$

$$P(Y | X) = ?$$

Association Rule Discovery:

Application 1

- Marketing and Sales Promotion:
 - Let the rule discovered be
{Bagels, ... } --> {Potato Chips}
 - *Potato Chips as consequent* => Can be used to determine what should be done to boost its sales.
 - *Bagels in the antecedent* => Can be used to see which products would be affected if the store discontinues selling bagels.
 - *Bagels in antecedent and Potato chips in consequent* => Can be used to see what products should be sold with Bagels to promote sale of Potato chips!

Association Rule Discovery:

Application 2

- Supermarket shelf management:
 - Goal: To identify items that are bought together by sufficiently many customers.
 - Approach: Process the point-of-sale data collected with barcode scanners to find **dependencies** among items.
 - A classic rule --
 - *If a customer buys diaper and milk, then he is very likely to buy beer.*
 - We may put six-packs of beers stacked next to diapers!

Association Rule Discovery:

Application 3

- Inventory Management:
 - Goal: A consumer appliance repair company wants to anticipate the **nature of repairs** on its consumer products and keep the service vehicles equipped with **right parts and tools** to reduce on number of visits to consumer households.
 - Approach: Process the data on **tools and parts** required in previous repairs at different consumer locations and discover the **co-occurrence patterns**.

Discussion: Association Rule

Clustering: Definition

- **Definition:** given a set of data points, each having a set of attributes, and a **similarity measure** among them, find **clusters** such that
 - Data points in one cluster are more similar to one another.
 - Data points in separate clusters are less similar to one another.
- **Similarity Measures:**
 - Euclidean Distance if attributes are continuous.
 - Other Problem-specific Measures.

Feature-based similarity measures
Probabilistic similarity measures

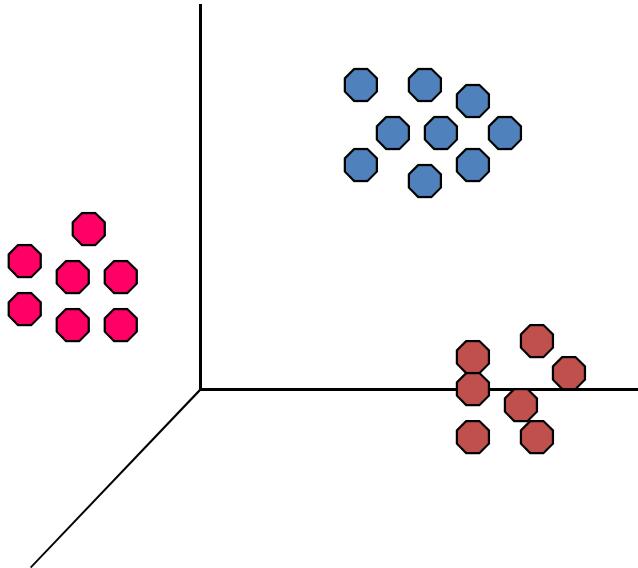
Perceptions concerning the taste of products, views on political issues, or opinions about people, are not deterministic.

Illustrating Clustering

Euclidean Distance Based Clustering in 3-D space.

Intra-cluster distances
are minimized

Inter-cluster distances
are maximized



Clustering: Application 1

- **Market Segmentation:**
 - Goal: a clustering model allocates customers similar in their attributes to the same group.
 - Approach:
 - Collect different attributes of customers based on their geographical and lifestyle related information.
 - Find clusters of similar customers.
 - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Segment customers into a small number of groups for additional analysis and marketing activities

Clustering: Application 2

- **Document Clustering:**

- Goal: To find groups of documents that are similar to each other based on the important terms appearing in them.
- Approach: To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.
- Gain: Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents.

- Relate a new document to a category from Financial, Foreign, National, Metro, Sports, Entertainment.
- Cluster WWW query results in hierarchical structure

Illustrating Document Clustering

- Clustering Points: 3204 Articles of Los Angeles Times.
- Similarity Measure: How many words are common in these documents (after some word filtering).

<i>Category</i>	<i>Total Articles</i>	<i>Correctly Placed</i>
<i>Financial</i>	555	364
<i>Foreign</i>	341	260
<i>National</i>	273	36
<i>Metro</i>	943	746
<i>Sports</i>	738	573
<i>Entertainment</i>	354	278

Clustering: Application 3

- **Image compression:**
 - the input instances are image pixels represented as RGB values;
 - quantize colour and use average for the colours in the same group
- **Bioinformatics:**
 - DNA in our genome is a sequence of bases.
 - A protein is a sequence of amino acids
 - One application area of clustering in molecular biology is *alignment*, which is - matching one sequence to another

Classification: Definition

- **Definition:** Given a collection of records
 - Each record contains a set of **attributes**, one of the attributes is the **class label**.
- Find a **model** for class label as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
 - A test set is used to determine the accuracy of the model.
 - Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Classification is for discrete target variables

Classification Example

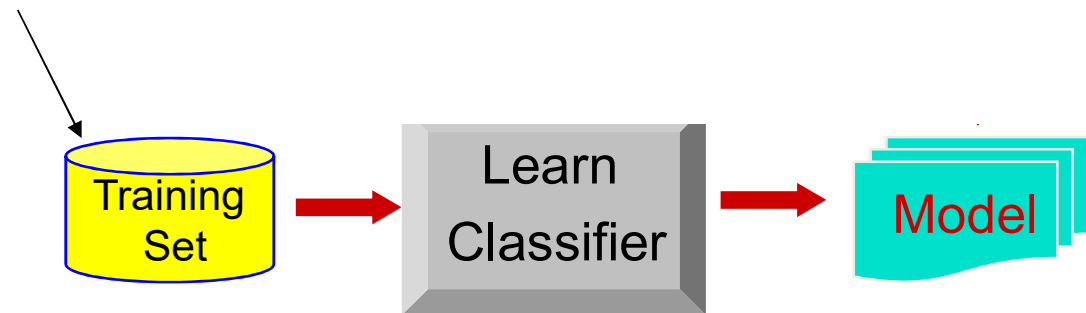
<i>Tid</i>	<i>Refund</i>	<i>Marital Status</i>	<i>Taxable Income</i>	<i>Cheat</i>
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

categorical

categorical

continuous

class



Classification: Application 1

- **Direct Marketing:**
 - Goal: Reduce cost of mailing by *targeting* a set of consumers likely to buy a new cell-phone product.
 - Approach:
 - Use the data for a similar product introduced before.
 - We know which customers decided to buy and which decided otherwise.
 - Collect various demographic, lifestyle, and company-interaction related information about all such customers.
 - Type of business, where they stay, how much they earn, etc.
 - Use this information as input attributes to learn a classifier model.

Collect data; Identify attributes; define class label; build a model;
➔ Make prediction.

Classification: Application 2

- **Fraud Detection:**
 - Goal: Predict fraudulent cases in credit card transactions.
 - Approach:
 - Use credit card transactions and the information on its account-holder as attributes.
 - When does a customer buy, what does he buy, how often he pays on time, etc
 - Label past transactions as fraud or fair transactions.
 - Learn a model for the class of the transactions.

Collect data; Identify attributes; define class label; build a model;
➔ Make prediction.

Classification: Application 3

- **Customer Attrition/Loyalty:**
 - Goal: To predict whether a customer is likely to be lost to a competitor.
 - Approach:
 - Use detailed record of transactions with each of the past and present customers, to find attributes.
 - How often the customer calls, where he calls, what time-of-the day he calls most, his financial status, marital status, etc.
 - Label the customers as loyal or disloyal.
 - Find a model for loyalty.

Collect data; Identify attributes; define class label; build a model;
➔ Make prediction.

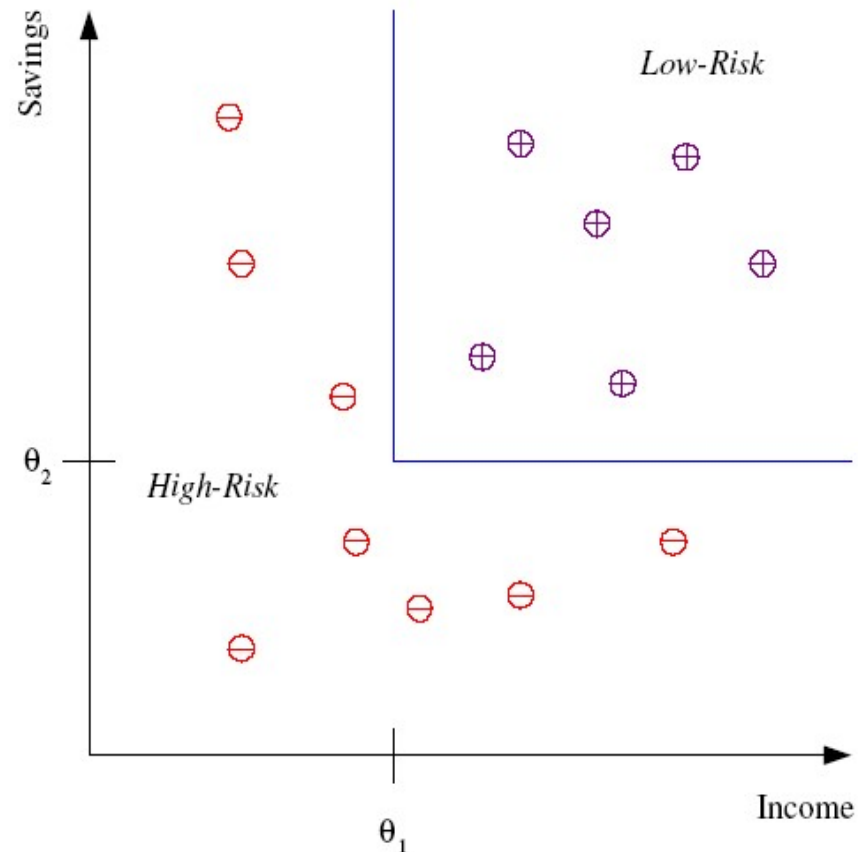
Classification: Application 4

- **Sky Survey Cataloging:**
 - Goal: To predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images.
 - Approach:
 - Segment the image.
 - Measure image attributes (features) - 40 of them per object.
 - Model the class based on these features.
 - Success Story: Could find 16 new high red-shift quasars, some of the farthest objects that are difficult to find!

Collect data; Identify attributes; define class label; build a model;
➔ Make prediction.

Classification: Application 5

- Credit scoring:
 - A credit is an amount of money loaned by a financial institution;
 - Differentiating between **low-risk** and **high-risk** customers from their income and savings



Discriminant: IF $income > \theta_1$ AND $savings > \theta_2$
THEN **low-risk** ELSE **high-risk**

Discussion: Classification

Regression: Definition

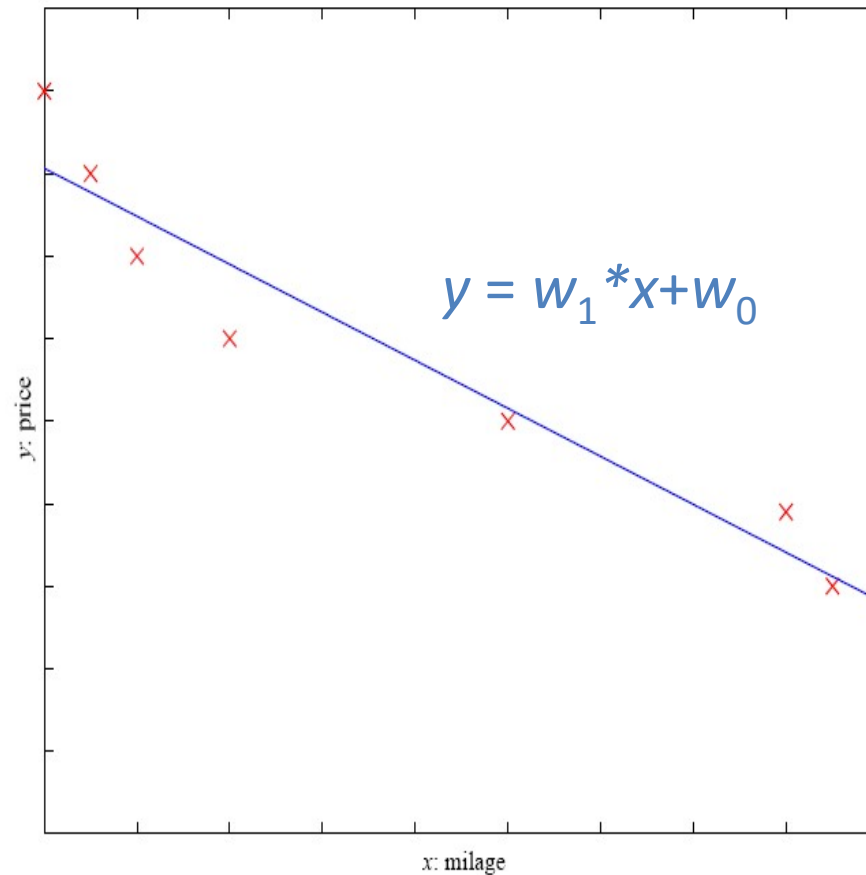
- **Definition:** Regression is a data mining technique to predict a value of a given **continuous valued variable** based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Greatly studied in statistics, neural network fields.
- Examples:
 - Predicting sales amounts of new product based on advertising expenditure.
 - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
 - Time series prediction of stock market indices.

Regression is for continuous valued target variables

Regression Method

Example: Price of a used car

- x : car attributes
- y : price
- $y = g(x \mid \theta)$
- $g(.)$ model,
- θ parameters



Regression Applications

- Navigate a mobile robot (autonomous car)
 - Output: angle of the steering wheel
 - Inputs: provided by sensors on the car
 - Build a machine that roasts coffee
 - Output: quality of coffee
 - Inputs: temperatures, times, coffee bean type,
- ✓ We make a number of experiments and for different settings of these inputs, and measure the quality of the coffee.
 - ✓ To find the optimal setting, we fit a regression model linking these inputs to coffee quality and choose new points to **sample** near the optimum of the current model to look for a better configuration

Reinforcement Learning: Definition

- **Definition:** Reinforcement learning is a machine learning technique, inspired by behaviourist psychology, concerned with how software agents should **take actions** in an environment so as to maximize the **cumulative reward**.
- RL differs from standard supervised learning in that correct input/output pairs are not presented, but **delayed reward**;
- There is a focus on on-line performance, which involves finding a balance between **exploration and exploitation**.
- Learning a policy: to seek a **sequence of actions** to reach the goal.

Reinforcement Learning: Applications

- Credit assignment problem
- Game playing to achieve economic utility
- Robot in a maze to reach the goal state
- Robot with partial sensory info
- Multiple agents to accomplish a common goal

Assign credit for the success among the multitude of decisions

A single move by itself is not that important; it is the sequence of right moves that is good

Learn the correct sequence of actions to reach to the goal state from an initial state

Discussion: Association, Classification, Clustering

Supervised Learning

- Start with learning a class from a simple case
- Introduce concepts of training set and hypothesis class
- Explain specific, general hypothesis and margin
- Appreciate noise and model complexity
- Extend to learning multiple classes
- Discuss regression learning for continuous outputs
- Explain model selection and generalization
- Understand main decisions of a supervised learner

An Example

- We have a set of examples of cars, and we have a group of people that we survey to whom we show these cars.
- The people look at the cars and label them;
- The cars that they believe are family cars are ***positive examples***, and the other cars are ***negative examples***.
- **Class learning** is finding a description that is shared by all positive examples and none of the negative examples.

Prediction: Given a unseen car before, by checking with the description learned, we will be able to say whether it is a family car or not.

Knowledge extraction: This study may be sponsored by a car company, and the aim is to understand what people expect from a family car.

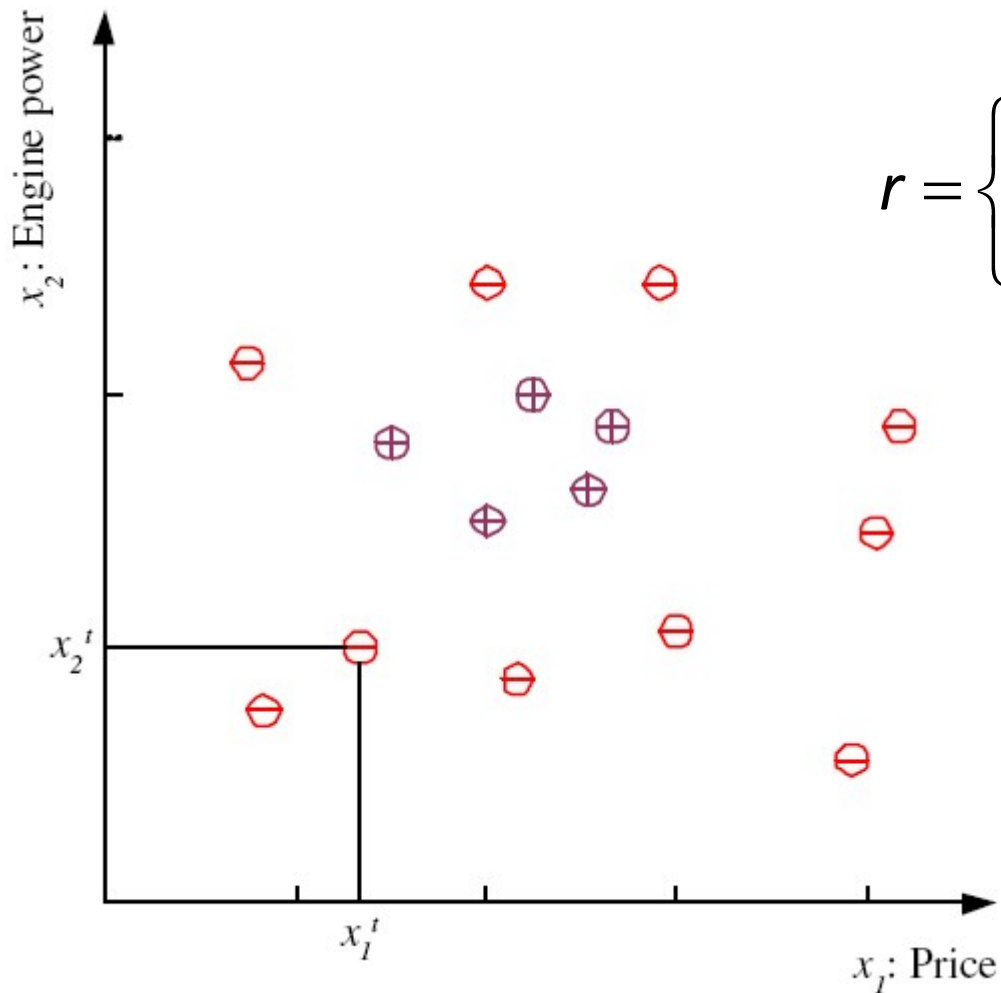
Learning a Class from Examples

- Class C of a “family car”
 - Prediction: Is car x a family car?
 - Knowledge extraction: What do people expect from a family car?
- Output:
 - Positive (+) and negative (–) examples
- Input representation:
 - x_1 : price, x_2 : engine power

Map Training set \mathcal{X}

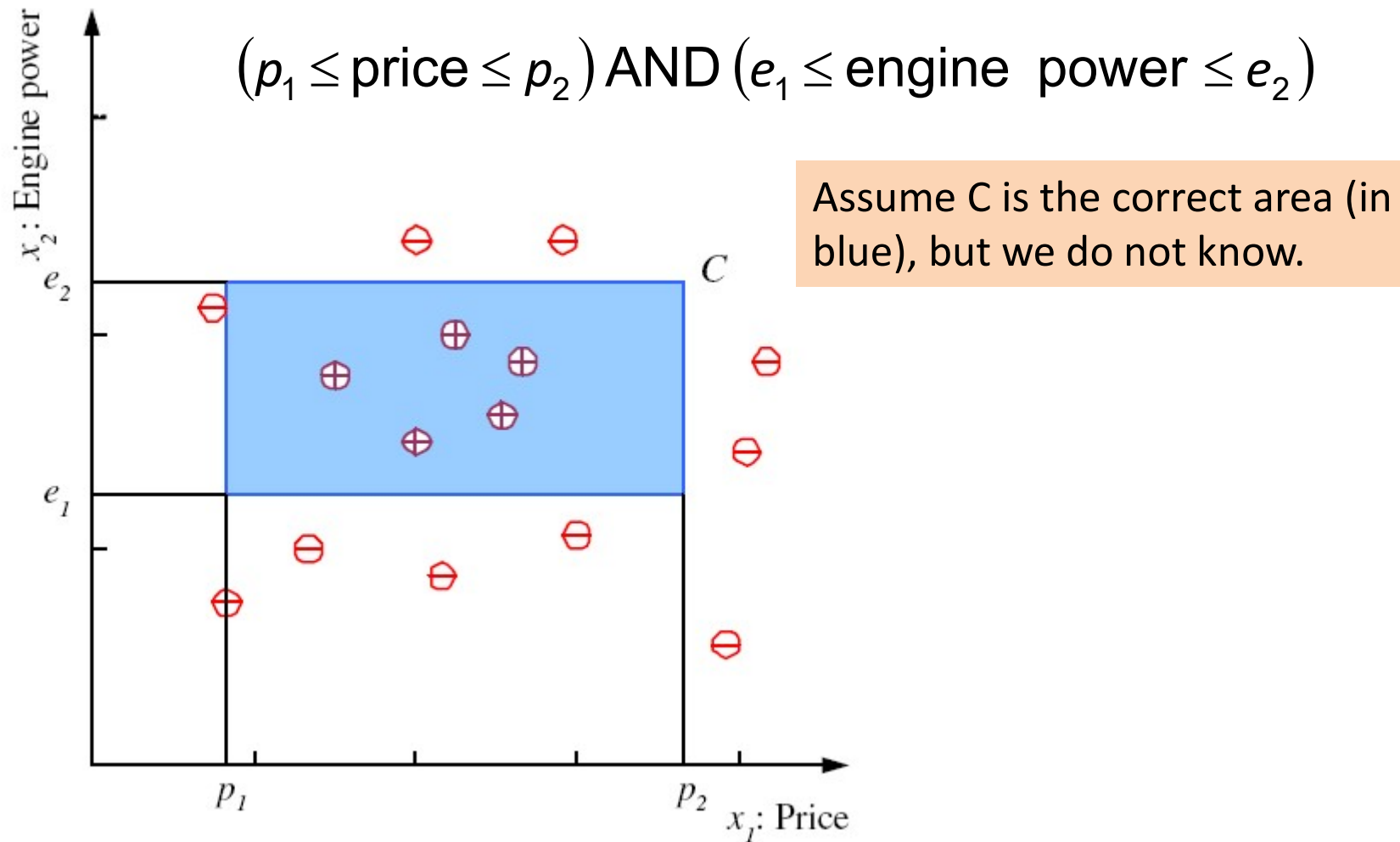
$$\mathcal{X} = \{\mathbf{x}^t, r^t\}_{t=1}^N$$

$$r = \begin{cases} 1 & \text{if } \mathbf{x} \text{ is positive} \\ 0 & \text{if } \mathbf{x} \text{ is negative} \end{cases}$$



$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Identify Class C



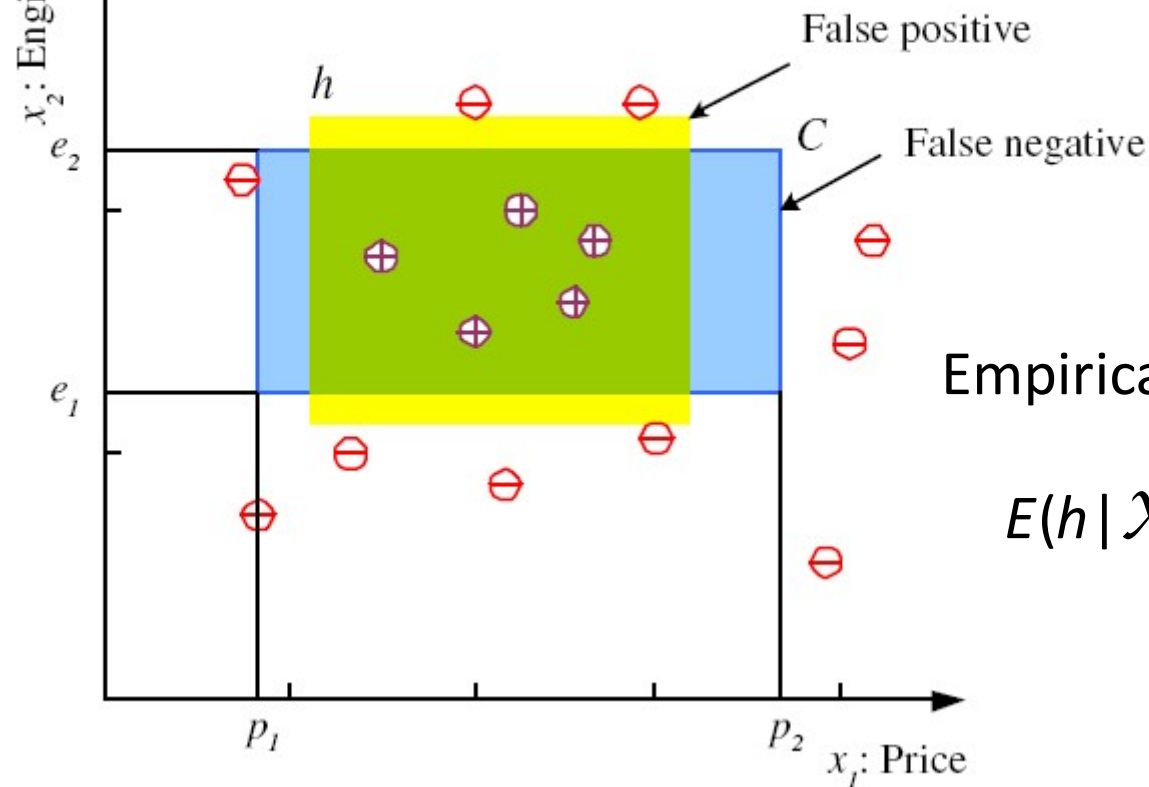
Hypothesis class \mathcal{H}

$$h(\mathbf{x}) = \begin{cases} 1 & \text{if } h \text{ says } \mathbf{x} \text{ is positive} \\ 0 & \text{if } h \text{ says } \mathbf{x} \text{ is negative} \end{cases}$$

Assume:

Correct = Blue area

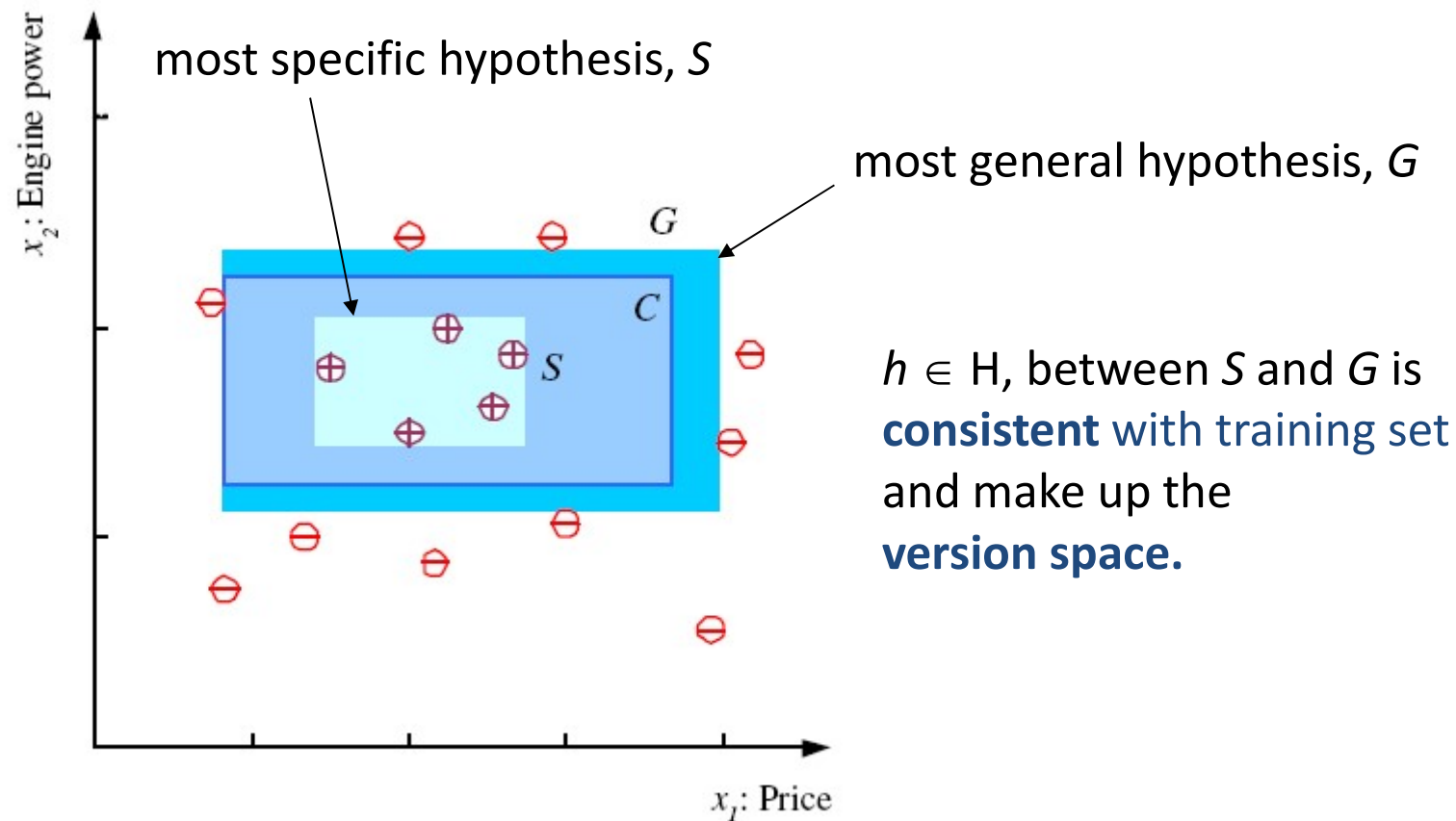
Hypothesis = Yellow area



Empirical Error of h on \mathcal{H} :

$$E(h | \mathcal{X}) = \sum_{t=1}^N 1(h(\mathbf{x}^t) \neq r^t)$$

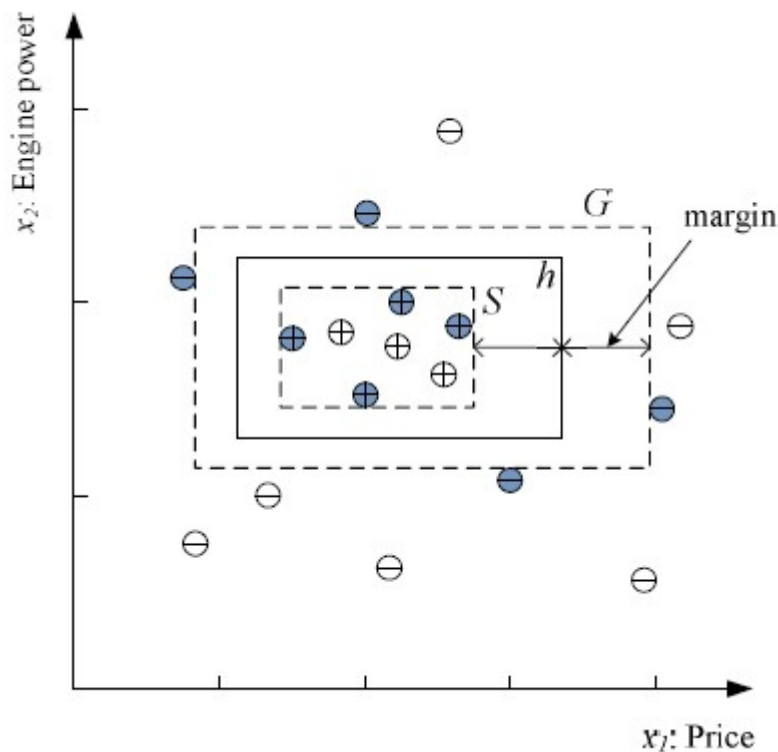
Most Specific, General Hypotheses



Problem: how well our hypothesis will correctly classify future examples that are not part of the training set.

Margin in Version Space

- Choose h with largest margin between S and G .



In some applications, a wrong decision may be very costly, we can say that any instance that falls in between S and G is a case of *doubt*.

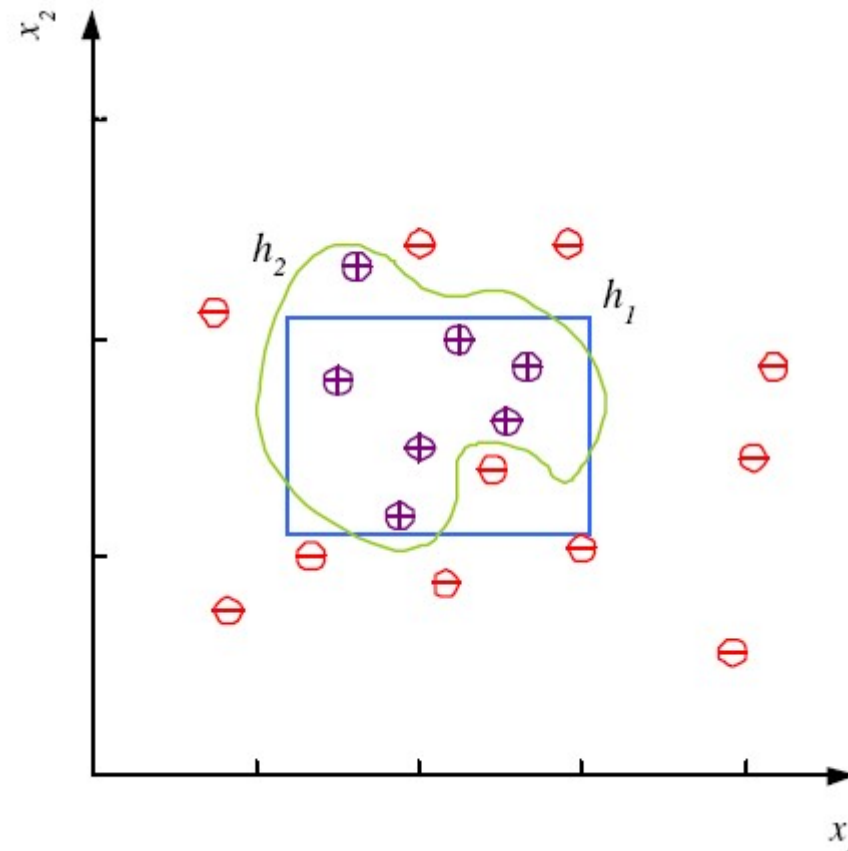
How to formulate the function $h(x)$ if not assuming rectangle H ?

Noise and Model Complexity

Noise is any unwanted anomaly in the data, the class may be more difficult to learn.

- Imprecision in recording the input attributes
- Teacher noise (error in re-label)
- Hidden or latent noise

To define a more complicated shape one needs a **more complex model** with a much larger number of parameters



Noise and Model Complexity

Use the simple rectangle model because

- Simpler to use (lower computational complexity)
- Easier to train (lower space complexity)
- Easier to explain (more interpretable)
- Generalizes better (lower variance)

It is easier to find the corner values of a rectangle than the control points of an arbitrary shape.

Occam's razor: *"simpler explanations are more plausible"*

Einstein's razor: *"Everything should be made as simple as possible, but no simpler"*

Principle: a simple (but not too simple) model would generalize better than a complex model.

Learning Multiple Classes

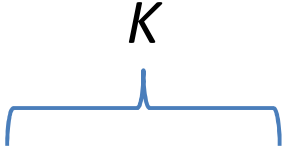
- The family car example is a two-class problem
- In general case, we have **K classes** denoted as $C_i, i = 1, \dots, K$.
- The training set:

$$\mathcal{X} = \{\mathbf{x}^t, \mathbf{r}^t\}_{t=1}^N$$

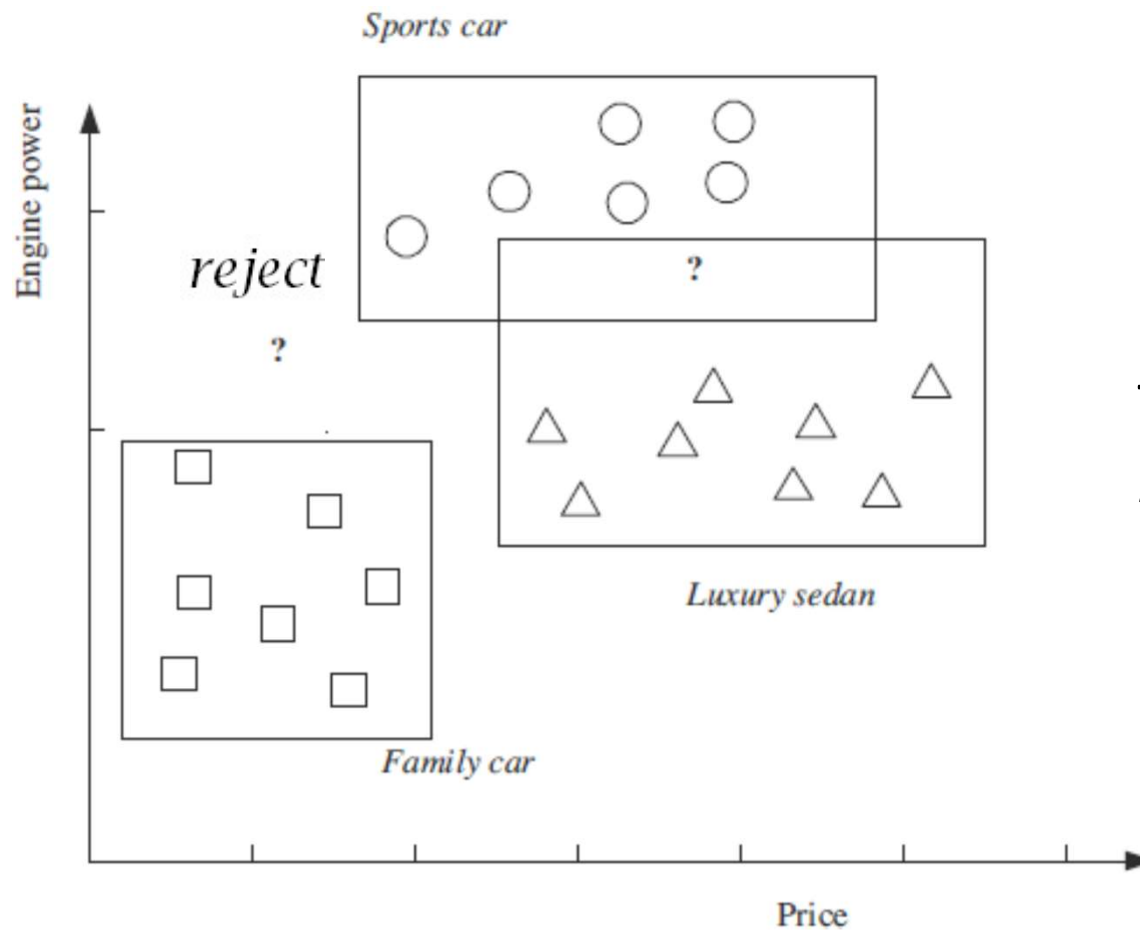
- where

$$r_i^t = \begin{cases} 1 & \text{if } \mathbf{x}^t \in C_i \\ 0 & \text{if } \mathbf{x}^t \in C_j, j \neq i \end{cases}$$

\mathbf{r} has K dimensions, e.g. $\mathbf{r}^t = \{1, 0, 0, \dots, 0\}$



Multiple Classes, C_i $i=1,\dots,K$



Train hypotheses
 $h_i(\mathbf{x})$, $i = 1, \dots, K$:

$$h_i(\mathbf{x}^t) = \begin{cases} 1 & \text{if } \mathbf{x}^t \in C_i \\ 0 & \text{if } \mathbf{x}^t \in C_j, j \neq i \end{cases}$$

Comments on Multiple Classes

- Total empirical error is:

$$E(\{h_i\}_{i=1}^K | \mathcal{X}) = \sum_{t=1}^N \sum_{i=1}^K 1(h_i(\mathbf{x}^t) \neq r_i^t)$$

- For a given \mathbf{x} , ideally only one of $h_i(\mathbf{x})$, $i = 1, \dots, K$ is 1 and we can choose a class.
- When no, or two or more, $h_i(\mathbf{x})$ is 1, we cannot choose a class, and this is the case of **doubt** and the classifier **rejects** such cases.
- In a dataset, if all classes have **similar distribution**, then the same hypothesis class can be used for all classes
- Handwritten digit recognition dataset vs. medical diagnosis data.

All healthy people are alike; each sick person is sick in his/ her own way

Regression

- When the output is a **numeric value**, what we would like to learn is a numeric function.

$$\mathcal{X} = \{x^t, r^t\}_{t=1}^N$$

Training dataset

$$r^t \in \mathbb{R}$$

$$r^t = f(x^t) + \varepsilon$$

where $f(x) \in \mathbb{R}$ is the unknown function and ε is random noise (e.g. hidden variables).

$$E(g | \mathcal{X}) = \frac{1}{N} \sum_{t=1}^N [r^t - g(x^t)]^2$$

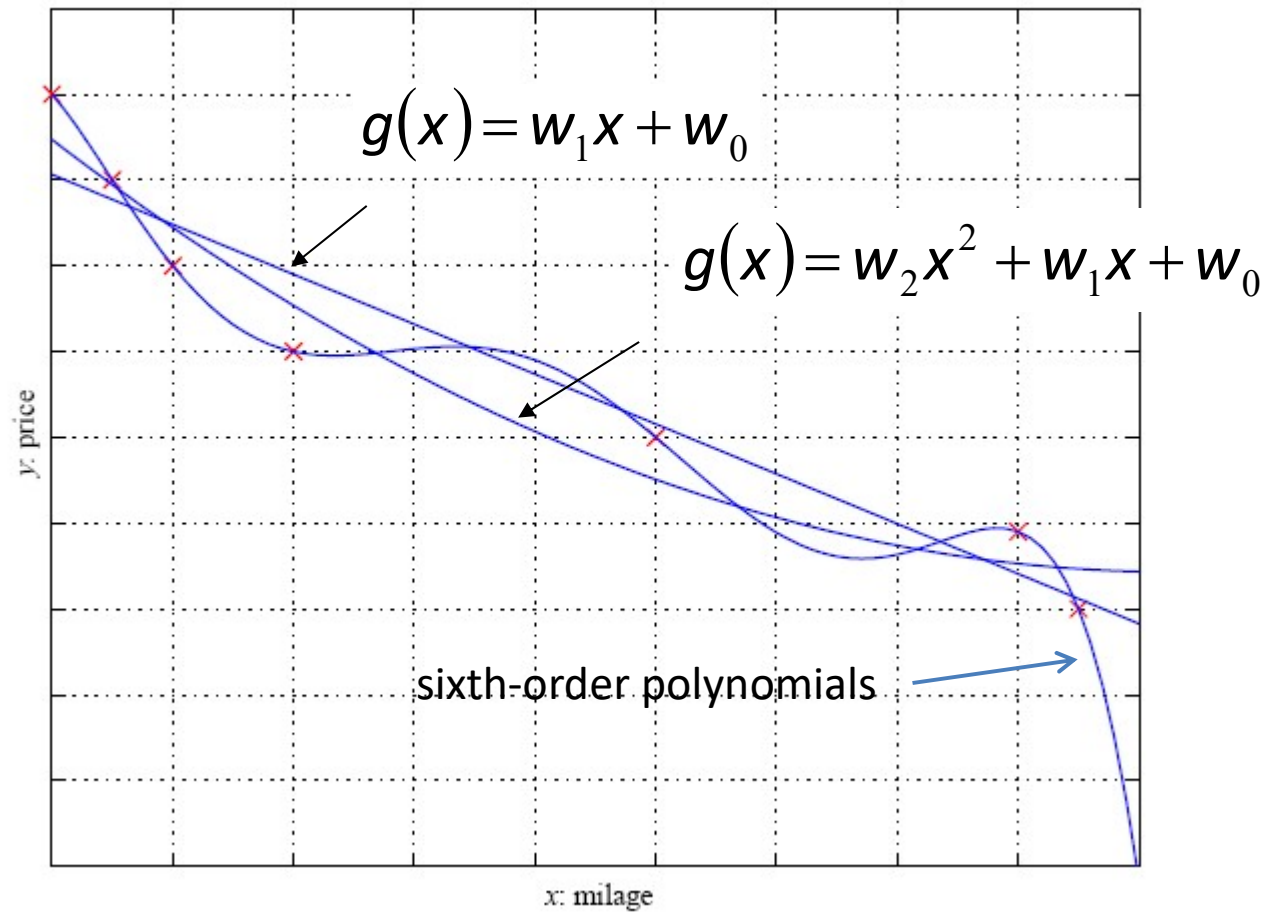
Empirical error

If we assume that $g(x)$ is linear, then

$$g(x) = w_1 x_1 + \dots + w_d x_d + w_0$$

→ How do we determine $g(x)$?

Linear, Second-order, Sixth-order Polynomial Regression



Linear Regression

- We used a single input linear model

$$g(\mathbf{x}) = w_1 x + w_0$$

- where w_1 and w_0 are the **parameters** to learn from data, which should minimize

$$E(w_1, w_0 | \mathcal{X}) = \frac{1}{N} \sum_{t=1}^N [r^t - (w_1 x^t + w_0)]^2$$

- Take the first partial derivatives with respect to w_1 and w_0 ,

$$\frac{1}{N} \sum_{t=1}^N [r^t - w_1 x^t - w_0] x^t = 0$$

$$\frac{1}{N} \sum_{t=1}^N [r^t - w_1 x^t - w_0] = 0$$



$$w_1 = \frac{\sum_t x^t r^t - \bar{x} \cdot \bar{r} \cdot N}{\sum_t (x^t)^2 - N \cdot \bar{x}^2}$$

$$w_0 = \bar{r} - w_1 \bar{x}$$

Non-linear Regression

- If the linear model is too simple and incurs a large approximation error, use a higher-order function of the input, e.g. quadratic:

$$g(\mathbf{x}) = w_2x^2 + w_1x + w_0$$

where similarly we have an **analytical solution** for the parameters.

- When the order of the polynomial is increased, the error on the training data decreases, e.g. the sixth-order polynomial in the figure.
- However, **Occam's razor** also applies in the case of regression

Model Selection & Generalization

- We should make some extra assumptions to have a unique solution with the data we have (e.g. rectangle with the largest margin).
- The set of assumptions we make to make learning possible is called the **inductive bias** of the learning algorithm (about \mathcal{H})
- Each hypothesis class has a certain **capacity** and can learn only certain functions, e.g. hypothesis class with a union of two rectangles has higher capacity
- **Model selection** refers to choosing between possible \mathcal{H} .
- **Generalization** refers to how well a model performs on new data

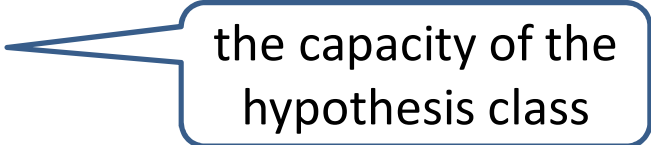
Model Selection & Generalization

- For best generalization, we should match the complexity of the hypothesis class H with the complexity of the function underlying the data
- **Underfitting:** \mathcal{H} less complex than C or f , e.g. trying to fit a line to data sampled from a third-order polynomial
- **Overfitting:** \mathcal{H} more complex than C or f , e.g. when fitting two rectangles to data sampled from one rectangle

Triple Trade-Off

- There is a trade-off between three factors (Dietterich, 2003):

1. Complexity of \mathcal{H} , $c(\mathcal{H})$,



the capacity of the hypothesis class

2. Training set size, N ,

3. Generalization error, E , on new data

- As N increases, $E \downarrow$
- As $c(\mathcal{H})$ increases, first $E \downarrow$ and then $E \uparrow$.

Cross-Validation

- To measure the generalization ability of a hypothesis, we have to access data outside of training set, i.e. validation set.
- Assuming large enough training and validation sets, the hypothesis that is the most accurate on the validation set is the best one.
- This process is called *cross-validation*.
- To estimate generalization error, we need data unseen.
- Resampling when there is few data

Training set (50%)
Validation set (25%)
Test (publication) set (25%)

Decisions of a Supervised Learner

1. Model: $g(\mathbf{x} | \theta)$

2. Loss function: $E(\theta | \mathcal{X}) = \sum_t L(r^t, g(\mathbf{x}^t | \theta))$

3. Optimization procedure:

$$\theta^* = \arg \min_{\theta} E(\theta | \mathcal{X})$$

Analytical methods,
Gradient-based methods,
Simulated Annealing,
Genetic Algorithms

Comments

1. The hypothesis class of $g(\cdot)$ should be large enough
2. There should be enough training data to allow us to pinpoint a good enough hypothesis
3. We should have a good optimization method that finds the correct hypothesis given the training data

Summary

- Concept of data mining and machine learning
- The CRISP-DM Process Model
- Relationships with statistics, pattern recognition, database, AI
- Main data mining techniques
- Applications of data mining and machine learning
- Concepts in supervised learning
- Noise and model complexity
- Model selection and generalization
- Main decisions of a supervised learner