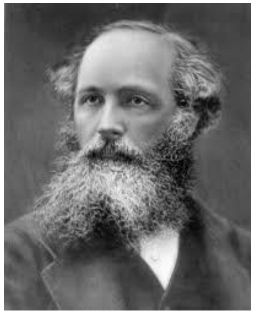
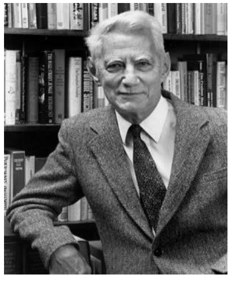
Lecture No. 3 – The Role of Information Theory in Data Mining

- Information Theory Overview
  - Basic Concepts
  - Data Compression
  - Communication Channel
- Information-Theoretic Approaches to Data Mining
  - The Uncertainty Approach
  - The Data Compression Approach
  - Minimum Description Length (MDL) Principle
- Summary





Claude Elwood Shannon (1916-2001)

Motivation: Why Information Theory?

- Data mining objectives – reminder
  - Identify valid, novel, potentially useful, and ultimately understandable <u>patterns</u> in data
- Any large dataset contains a potentially infinite amount of patterns
  - Most of them are <u>not</u> valid (random), completely useless, or too complex to be understood properly
- We need objective criteria for inducing the most <u>informative</u> patterns from data
- **Information theory** provides a nice formal framework for finding and evaluating patterns

Information and Uncertainty

- What is information?
  - Attneave (1959): Information is that which removes or reduces *uncertainty*
- Uncertainty is our limited knowledge about the *outcome* of some (future) event
- Examples of uncertain events
  - Credit card transaction (legitimate / fraudulent)
  - A patient clinical condition (disease = ???)
  - Final grade point average of a student admitted to the university (outcome = value between 0 and 100)
  - More?

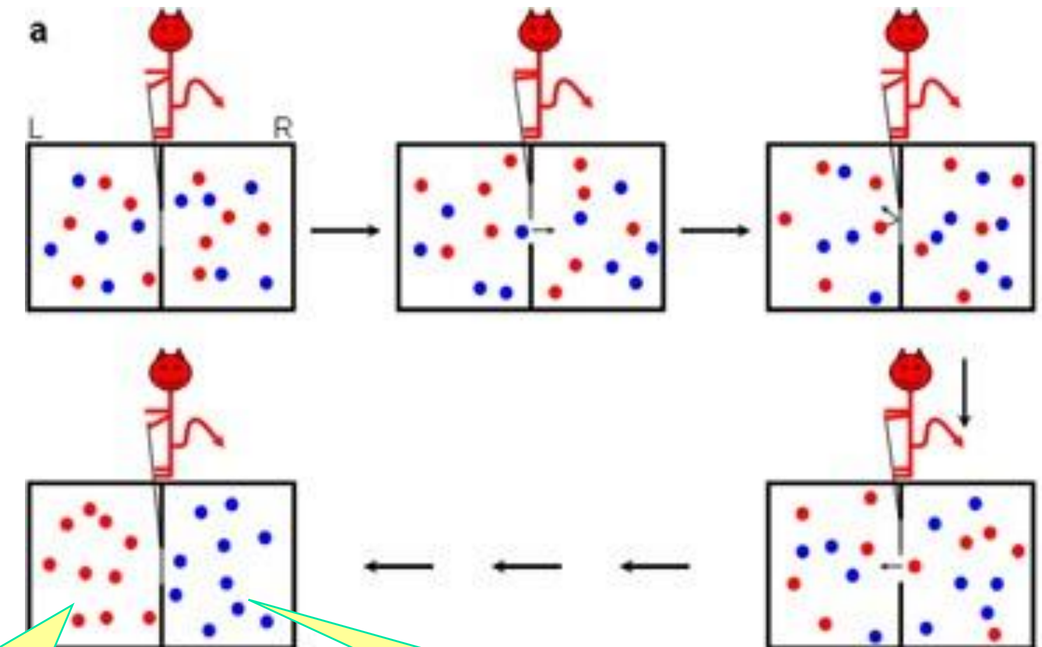
How to measure uncertainty?

- A quantitative measure of uncertainty should have at least the following properties
  - If the outcome of an event can be predicted with a 100% accuracy, then the uncertainty of an event is zero
  - The uncertainty of an event increases with the number of possible outcomes (cc vs. student)
  - For the same number of outcomes, the uncertainty is maximal if each outcome has the same probability (examples?)

Thermodynamics: Maxwell's Demon

(an air-conditioner that needs no power supply)

Background slide



A free app?

Outside temperature

Inside temperature

The Bad News: Maxwell's Demon is impossible

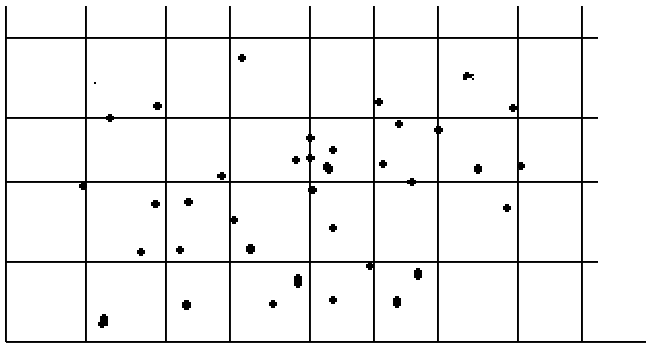
Background slide

- The demon would still need to use energy to observe the molecules (in the form of photons for example).
- Leo Szilard (1929)
  - The Demon has to process information in order to make his decisions, and, in order to preserve the first and second laws (of conservation of energy and of entropy), the energy requirement for processing this information is always greater than the energy stored up by sorting the molecules.
- Shannon (1948)
  - All transmissions of information require a physical channel,

Background slide

Thermodynamics: Boltzmann's Entropy

Represent system in a space whose coordinates are positions and momenta = *mv* (phase space).



Subdivide space into *B* bins.

p_k = fraction of particles whose positions and momenta are in bin *k*.

Amount of uncertainty, or missing information, or randomness, of the distribution of the p_k's, can be measured by $H_B = \sum p_k \log(p_k)$ (also called Gibbs *H*)

Entropy: $S = - k_B H_B$ (k_B - Boltzmann's constant)

Shannon's Information Theory
Basic Concepts

- **Entropy**
- **Goal**: measure of uncertainty of *X*
- $H(X) = \Sigma - p(x) \log_2 p(x)$

  Where

  X - a discrete random variable

  x - value of *X*

  p (x) - probability of *x*

Properties of Entropy

1. $H(X) = 0$ if and only if the outcome is deterministic (all $p(x)$ but one are zero): $-0*\log 0 - 1*\log 1 = 0$
2. $H(X) \leq \log$ [number of outcomes]. $H(X)$ is a maximum, when all outcomes are equiprobable: $\max H(X) = \log$ [number of outcomes]
3. If all the outcomes have the same probability, then $H(X)$ is a monotonic increasing function of the number of outcomes

Entropy – Example

Calculating *H (Test)* and *H (Disease)*

Data	Disease = Yes	Disease = No	Total
Test = Negative	1	3	**4**
Test = Positive	4	2	**6**
Total	**5**	**5**	**10**

Max H (Test) = ?

Max H (Disease) = ?

$H(X) = \Sigma - p(x) \log_2 p(x)$

H (test)	Test = Negative	Test = Positive	Total
p(test)	0.4	0.6	
-log p(test)	1.322	0.737	
-p(test)*log p(test)	0.529	0.442	**0.971**

Entropy of *Test*

H (disease)	Disease = Yes	Disease = No	Total
P(disease)	0.5	0.5	
-log p(disease)	1.000	1.000	
-p(disease)*log p (disease)	0.500	0.500	**1.000**

Entropy of *Disease*

Information Theory

Basic Concepts (cont.)

- **Conditional Entropy**
- **Goal**: measure of uncertainty of *Y*, when *X* is given.
- $H(Y/X) = - \Sigma p(x,y)*\log p(y/x)$

  Where

  X, Y – discrete random variables

  p (x,y) – joint probability of *x* and *y*

  p (y/x) – conditional probability of *y* given *x*

  Properties of Conditional Entropy

1. *If* $Y = f(X)$ *then* $H(Y/X) = 0$
2. *The uncertainty of Y is never increased by knowledge of X:* $H(Y/X) \leq H(Y)$
3. *If X and Y are independent, then* $H(Y/X) = H(Y)$

Conditional Entropy – Example 1

Calculating *H (Disease/Test)*

Data	Disease = Yes	Disease = No	Total
Test = Negative	1	3	**4**
Test = Positive	4	2	**6**
Total	**5**	**5**	**10**

$H(Y/X) = - \Sigma\, p(x,y)*\log p\,(y/x)$

Max H (Disease/Test) = ?

P(test,disease)	Disease = Yes	Disease = No	Total
Test = Negative	0.10	0.30	**0.4**
Test = Positive	0.40	0.20	**0.6**
Total	**0.5**	**0.5**	**1.00**

Conditional entropy of Disease/Test

P(disease/test)	Disease = Yes	Disease = No	Total
Test = Negative	0.25	0.75	**1.00**
Test = Positive	0.67	0.33	**1.00**

-log p(disease/test)	Disease = Yes	Disease = No
Test = Negative	2.000	0.415
Test = Positive	0.585	1.585

-p (test,disease)*log p(disease/test)	Disease = Yes	Disease = No	Total
Test = Negative	0.200	0.125	**0.325**
Test = Positive	0.234	0.317	**0.551**
Total H(disease/test)	**0.434**	**0.442**	**0.875**

Conditional Entropy – Example 2

Calculating *H (Test/Disease)*

Data	Disease = Yes	Disease = No	Total
Test = Negative	1	3	**4**
Test = Positive	4	2	**6**
Total	**5**	**5**	**10**

*H(Y/X) = - Σ p(x,y)*log p (y/x)*

Max H (Test/Disease) = ?

P(test,disease)	Disease = Yes	Disease = No	Total
Test = Negative	0.10	0.30	**0.4**
Test = Positive	0.40	0.20	**0.6**
Total	**0.5**	**0.5**	**1.00**

Conditional entropy of Test/Disease

P(test/disease)	Disease = Yes	Disease = No
Test = Negative	0.20	0.60
Test = Positive	0.80	0.40
Total	**1.00**	**1.00**

-log p(test/disease)	Disease = Yes	Disease = No
Test = Negative	2.322	0.737
Test = Positive	0.322	1.322

-p (test,disease)*log p(test/disease)	Disease = Yes	Disease = No	Total
Test = Negative	0.232	0.221	**0.453**
Test = Positive	0.129	0.264	**0.393**
Total H (test/disease)	**0.361**	**0.485**	**0.846**

Information Theory
Basic Concepts (cont.)

- **Mutual Information** (of variables *X* and *Y*)

Goal: the reduction in the uncertainty of *Y* as a result of knowing *X*

$$I(X;Y) = H(Y) - H(Y/X) = \sum_{x,y} p(x,y) \bullet \log \frac{p(y/x)}{p(y)}$$

Properties of Mutual Information (MI)

1. *Symmetry: $I(X;Y) = I(Y;X) = H(X) - H(X/Y)$*
2. *Mutual information is a non-negative quantity: $I(X;Y) \geq 0$*
3. *Maximum MI: If $Y = f(X)$ then $I(X;Y) = H(Y)$*
4. *Minimum MI: If X and Y are independent, then $I(X;Y) = 0$*

Mutual Information – Example

Calculating *I (Test; Disease)*

P(test,disease)	Disease = Yes	Disease = No	Total
Test = Negative	0.10	0.30	**0.4**
Test = Positive	0.40	0.20	**0.6**
Total	**0.5**	**0.5**	**1.00**

$$I(X;Y)=\sum_{x,y} p(x,y)\bullet \log \frac{p(y/x)}{p(y)}$$

P(disease/test)	Disease = Yes	Disease = No	Total
Test = Negative	0.25	0.75	**1.00**
Test = Positive	0.67	0.33	**1.00**

Mutual information of Disease and Test

P(disease/test) / P(disease)	Disease = Yes	Disease = No
test=0	0.50	1.50
test=1	1.33	0.67

p(test,disease)*log p(disease/test) / p(disease)	disease= 0	disease= 1	Total
test=0	-0.100	0.175	**0.075**
test=1	0.166	-0.117	**0.049**
Total I (test;disease)	**0.066**	**0.058**	**0.125**
H (test) - H (test/disease)			**0.125**
H(disease) - H(disease/test)			**0.125**

Information Theory
Additional Concepts (to be re-visited)

- **Conditional Mutual Information**

$$I(X;Y/Z) = H(X/Z) - H(X/Y,Z) = \sum_{x,y} p(x,y,z) \bullet \log \frac{p(x,y/z)}{p(x/z) \bullet p(y/z)}$$

Interpretation: the decrease in entropy of X as a result of knowing Y, when Z is given

- **Chain Rule** $I(X_1,\ldots, X_n;Y) = \sum_{i=1}^{n} I(X_i;Y / X_{i-1},\ldots, X_1)$

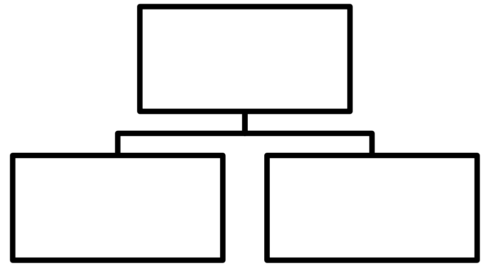
Interpretation: The decrease in entropy of *Y* as a result of knowing *n* variables ($X_1,\ldots, X_n$)

- **Fano's Inequality**: $H(Y/X_1\ldots X_n) \le H(P_e) + P_e \log_2 (m-1)$
    - $H(P_e) = -P_e \log P_e - (1-P_e)\log(1-P_e)$

Interpretation: Relationship between the minimum prediction error P_e, the conditional entropy of the target *H*, and the number of classes *m* ("upper bound of predictability" - Song et al., 2010)

Fano's Inequality - Example

The classification model:

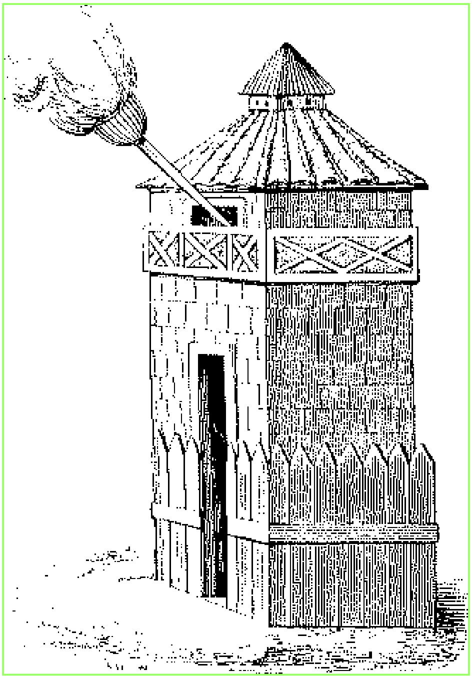


Data	Disease = Yes	Disease = No	Total
Test = Negative	1	3	4
Test = Positive	4	2	6
Total	5	5	10
Error	**1**	**2**	**3**
P_e			0.30
H(P_e)			0.881
H(disease/test)			0.875

$$H\,(Y/\,X_1 \ldots X_n) \le H\,(P_e) + P_e \log_2 (m-1)$$

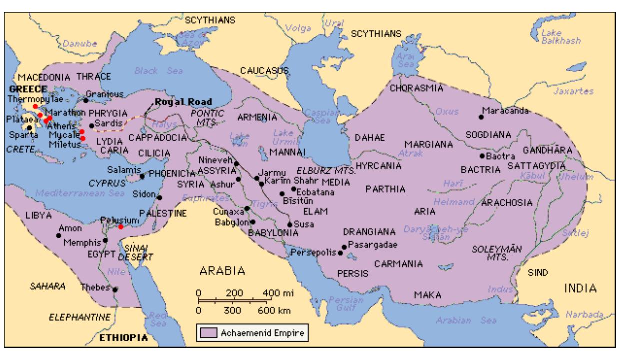
A Brief History of Communication

Background slide



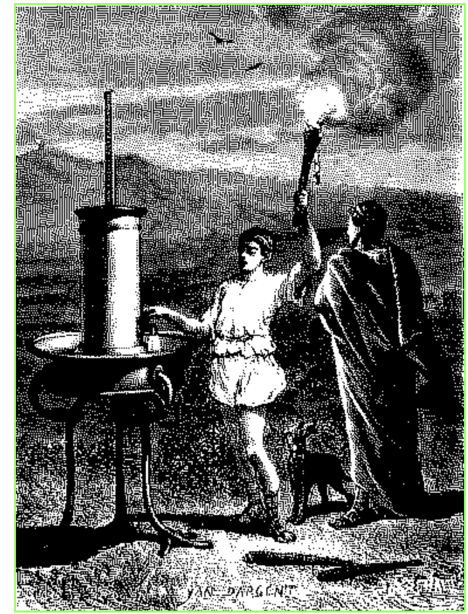
Persian Empire

Around 500 BC



Roman Empire

Around 100 AD



Information Theory and Data Compression

- The Fundamental Problem of Communication (Shannon, 1948)
  - To reproduce at one point (“destination”) either exactly or approximately a message (outcome) selected at another point (“data source”)
- Data Compression
  - Minimizing the number of binary digits (“description length”) required to encode a random message sent by the data source

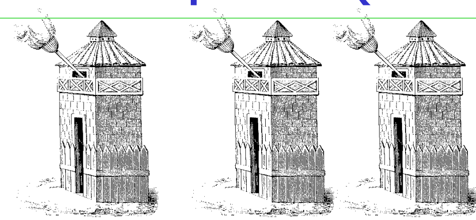
Information Theory and Data Compression (cont.)

- Optimal Code (Shannon, 1948)
  - Assigning $-\log p_i$ bits to encode message *i*
    - p_i – probability of the message (outcome) *i*
    - $-\log p_i$ – the *informational value* ("surprisal") of the outcome *i*
- Interpretation
  - Assign shorter codes (descriptions) to more frequent messages and vice versa
- Conclusion
  - The shortest average *description length* of a random message ("minimum description length") is the entropy of the data source $H(X) = \sum -p(x_i)\log p(x_i)$

Communication Systems

Communication goal: To reproduce at one point ("destination") either exactly or approximately a message (outcome) selected at another point ("data source")

Data Source



Transmitted Messages

(Sequences of symbols)

Communication Channel

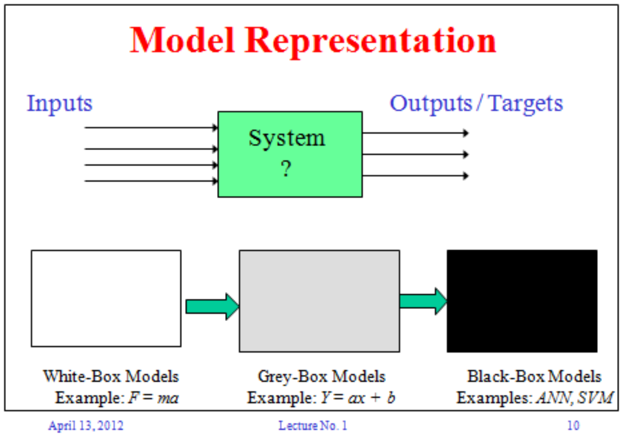
Received Messages

Noise source

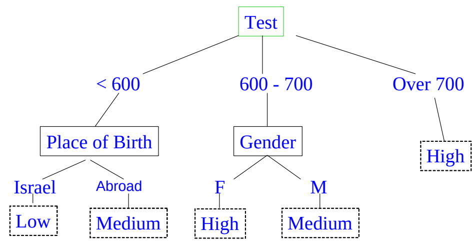
Destination

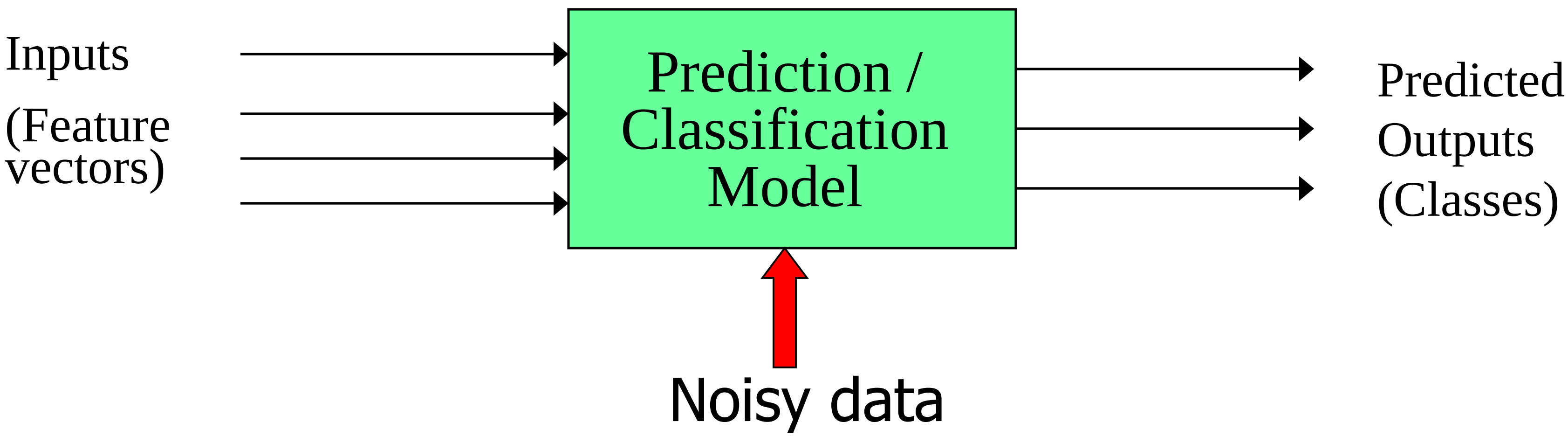
Data Mining Models

Reminder:



Classification goal: To predict either exactly or approximately an outcome of the actual system ("data source")





Background slide

Information Theory
Noisy Communication Channel
(based on Shannon, 1948)

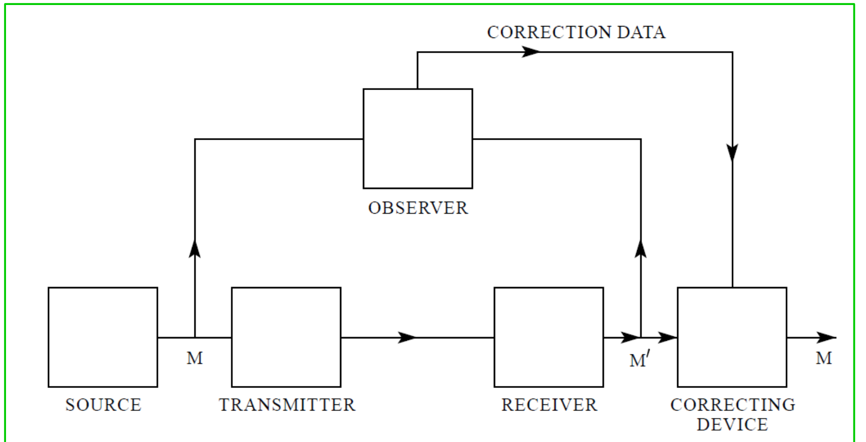


Fig. 8—Schematic diagram of a correction system.

- Received Signal $E = f(S, N)$
  - Where
    - S – Transmitted signal
    - N – Noise
- Q.1: Can we reconstruct with *certainty* the original signal from the received signal?
  - A.: No ☹
- Q.2: Can we eliminate the noise by transmitting the information in a certain way?
  - A.: Yes ☺ (By sending additional information to correct the received signal)

Noisy Communication Channel

Background slide

Example

How much information is transmitted if $P(Y = 1 / X = 0) = P(Y = 0 / X = 1) = 0.5$?

- The information source *X* is sending two symbols (0 and 1) with the same probability ($H(X) = ?$)
- *Gross rate* of transmission: 1000 bits (symbols) /second
- The error rate of the communication channel is about 1% for any input *x:*
  - $P(Y = 1 / X = 0) = P(Y = 0 / X = 1) = 0.01$
- What is the rate of transmission of information?
  - The minimum number of bits required to send the correction information ("1" – incorrect, "0" – correct):
    - $H(X / Y) = -[.99 \log .99 + 0.01 \log 0.01] = 0.081$ *bits* (81 bits per second)
  - The *net rate* of information transmission
    - 1000 – 81 = 919 bits per second (with 1% error frequency!)

Background slide

Channel Capacity

- Rate of transmission *R* - definitions:
  - $R = H(X) - H(X/Y)$
    - Amount of information sent less the uncertainty of what was sent
  - $R = H(Y) - H(Y/X)$
    - Amount received less the noise
  - $R = H(X) + H(Y) - H(X;Y)$
    - Amount sent + amount received less the joint entropy
- Channel Capacity (Maximum possible rate of transmission)
  - $C = Max_X \{ H(X) - H(X/Y) \}$

The Channel Coding Theorem

- If the entropy (uncertainty) per second of the information source H does not exceed the channel capacity C, it is possible to transmit the information over the channel with an arbitrarily small probability (frequency) of error
- Otherwise, the entropy (uncertainty) of the output will be at least $H - C$
- Data mining interpretation
  - A model can be as accurate as the input data itself, but no more

Lecture No. 2 – The Role of Information Theory in Data Mining

- Information Theory Overview
  - Basic Concepts
  - Data Compression
  - Communication Channel
- Information-Theoretic Approaches to Data Mining
  - The Uncertainty Approach
  - The Data Compression Approach
  - Minimum Description Length (MDL) Principle
- Summary

Information-Theoretic Approaches to Data Mining

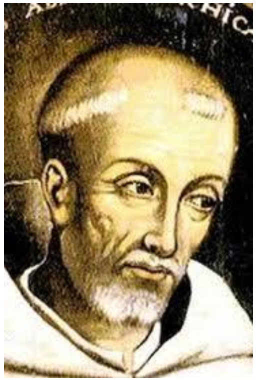
- The Uncertainty Approach
  - Data mining is aimed at reducing uncertainty of the target (predicted) variables
  - Uncertainty can be represented by entropy
  - Data mining algorithms look for models that minimize entropy or maximize mutual information (information gain)
  - Usage: ID3, C4.5, IFN, etc.

Information-Theoretic Approaches to Data Mining (cont.)

- The Data Compression Approach (see Mannila, 2000)
  - Smaller models are more comprehensible to the user
  - The goal of data mining is to *compress the data* by finding some structure (model) for it
  - Data mining algorithms should choose a hypothesis that compresses data the most (the MDL Principle)
  - Usage: Bayesian learning

Occam's Razor

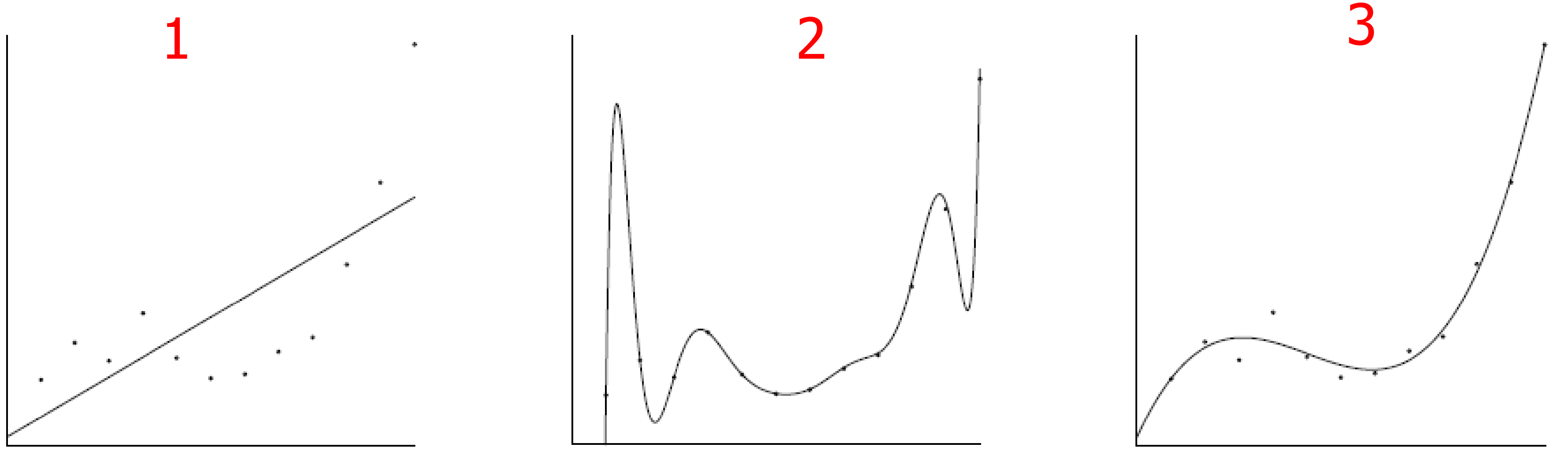
- Commonly attributed to William of Ockham (1290--1349). In sharp contrast to the principle of multiple explanations, it states:
  - *Entia non sunt multiplicanda praeter necessitatem*
  - Entities should not be multiplied beyond necessity.



- Commonly explained as:
  - when have choices, choose the simplest theory.
- Bertrand Russell: ``It is vain to do with more what can be done with fewer.'‘
- Newton (*Principia*):
  - *Natura enim simplex est, et rerum causis superfluis non luxuriat*
    - הטבע פשוט ואין לו עודף של סיבות מיותרות למהות הדברים

The Data Compression Approach

- Example: regression (line fitting)
  - Which model is the best?
    - Model selection and overfitting
    - Complexity of the model vs. Goodness of fit



Source: Grunwald et al. (2005) *Advances in Minimum Description Length: Theory and Applications.*

Minimum Description Length (MDL) Principle

- Problem Statement
  - The attribute values in each case are available to both a *sender* and a *receiver*
  - Only the sender knows the class to which each case belongs
  - The sender must transmit the classification information to the receiver *by using a minimum number of bits*
- Decision Variable
  - The model instance (“hypothesis”) to be used by the “sender” out of a given family of models (e.g., decision trees, info-fuzzy networks, *k*th degree polynomials, etc.)

The MDL Preliminaries

- *L (h)* - the length, in bits, of the description of the hypothesis (the *theory cost*)
  - Also called *parametric complexity* (measure of the model “richness”, related to the number of model parameters)
  - Example – decision tree: *L (h) = f (number of nodes)*
- In the case of noisy data, the *exceptions* to the hypothesis should also be transmitted
- *L (D/h)* - the length, in bits, of the description of the data under the assumption that both the sender and the receiver know the hypothesis (encoded with the help of the hypothesis)
  - Complex hypotheses lead to small *L (D/H)* and vice versa

The MDL Principle

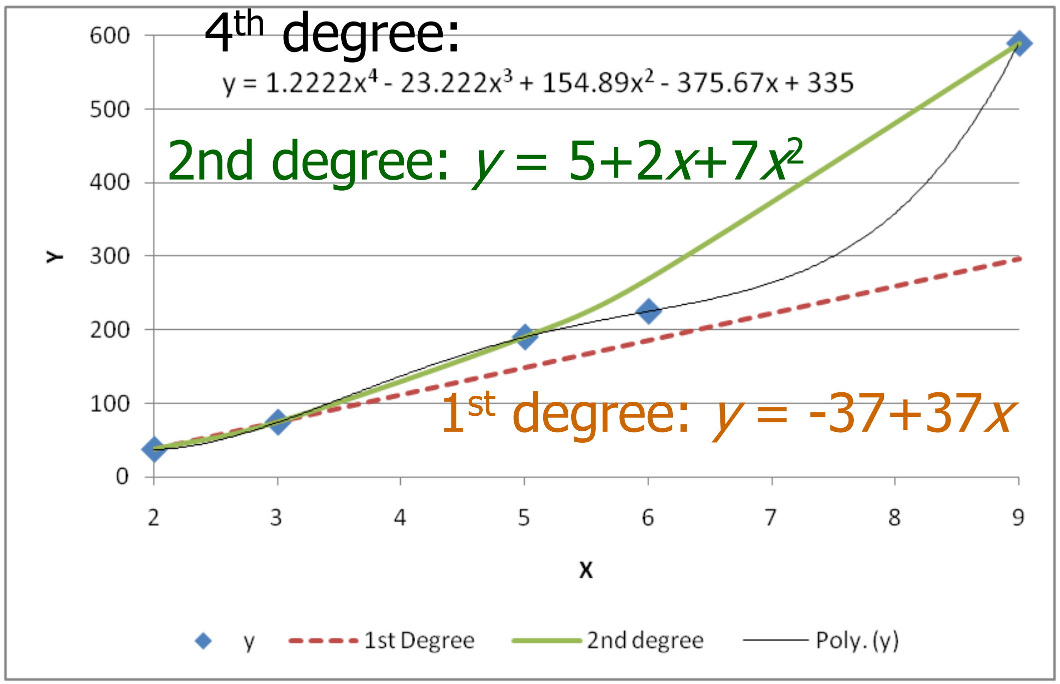
- Choose the hypothesis h_{MDL} which satisfies the following

$$\boxed{h_{MDL} = \arg\min_{h \in H} \{L_{C_1}(h) + L_{C_2}(D/h)\}}$$

  - L –description length (bits)
  - C_1 - the optimal encoding of the hypothesis h
  - C_2 – the optimal encoding of data D given the hypothesis
- Interpretation
  - The MDL principle represents the trade-off between the model complexity and the number of errors committed by it in the training data
- Practical Usage
  - The MDL principle proved to be an efficient tool for dealing with the problem of *overfitting*

MDL Example: Learning a polynomial

No	x	y	1st Degree	2nd degree	4th degree
1	2	37	37	37	37
2	3	74	74	74	74
3	5	190	148	190	190
4	6	225	185	269	225
5	9	590	296	590	590



- d - number of bits required to describe each entry in a polynomial
- Description length of degree k-1 polynomial: kd bits
- Description length of m points not on the polynomial: md bits
- MDL Cost:
  - 4^{th} degree: $5d$
  - 2^{nd} degree: $3d + d = 4d$
  - 1^{st} degree: $2d + 3d = 5d$
- Which model is the best?

Summary

- Information theory provides a nice formal framework for the process of data mining from both the uncertainty reduction aspect and the aspect of data compression
- The usage of information-theoretic heuristics in numerous data mining algorithms has brought satisfactory results in terms of predictive accuracy and model compactness
- Many other aspects of the information theory are still waiting for their implementation by the KDD researchers and practitioners (see Song et al., 2010)

Bibliography

- F. Attneave (1959). Applications of Information Theory to Psychology. Holt, Rinehart and Winston.
- T. M. Cover (1991). Elements of Information Theory. Wiley.
- O. Maimon and M. Last (2000), Knowledge Discovery and Data Mining – The Info-Fuzzy Network (IFN) Methodology, Kluwer Publishers.
- H. Mannila (2000). Theoretical Frameworks for Data Mining. SIGKDD Explorations, 1 (2): 30-32.
- T.M. Mitchell (1997). Machine Learning. McGraw-Hill.
- J.R. Quinlan (1986). Induction of Decision Trees. Machine Learning, 1 (1): 81-106.
- J. R. Quinlan (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann.
- J.R. Quinlan (1996). Improved Use of Continuous Attributes in C4.5. Journal of Artificial Intelligence Research, 4: 77-90.
- C.E. Shannon (1948), A Mathematical Theory of Communication, Bell Syst. Tech. J., 27: 379-423.
- C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, Limits of Predictability in Human Mobility. Science 19 February 2010: 327 (5968), 1018-1021.