Knowledge Engineering and Expert Systems

Lecture Notes on Machine Learning

Matteo Mattecci

matteucci@elet.polimi.it

Department of Electronics and Information

Politecnico di Milano

Course Info

- Course Material on Machine Learning (up-to-now)
  - *Machine Learning*, T. Mitchell, McGraw Hill, 1997
  - *The Elements of Statistical Learning*, T. Hastie, R. Tibshirani, J. Friedman, Springer, 2001
  - *Neural Networks and Pattern Recognition*, C. Bishop, Oxford University Press, 1995
  - *Reinforcement Learning: An Introduction*, R.S. Sutton, A.G. Barto, Bradford Books, 1998

Course Info

- Course Material on Machine Learning (up-to-now)
  - *Machine Learning*, T. Mitchell, McGraw Hill, 1997
  - *The Elements of Statistical Learning*, T. Hastie, R. Tibshirani, J. Friedman, Springer, 2001
  - *Neural Networks and Pattern Recognition*, C. Bishop, Oxford University Press, 1995
  - *Reinforcement Learning: An Introduction*, R.S. Sutton, A.G. Barto, Bradford Books, 1998

- Evaluation and Grading
  - Homeworks (every 15 days) and/or Midterm
  - Course Final Project

Dartmouth 1955 – Conception of AI

A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence

J. McCarthy, Dartmouth College
M.L. Minsky, Harvard University
N. Rochester, I.B.M. Corporation
C.E. Shannon, Bell Telephone Laboratories

August 31, 1955

We propose that a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.

http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html

Dartmouth 1956 – The AI Program

1. Automatic Computers
2. How Can a Computer be Programmed to Use a Language
3. Neuron Nets
4. Theory of the Size of a Calculation
5. Self-Improvement
6. Abstractions
7. Randomness and Creativity

Dartmouth 1956 – The AI Program

1. Automatic Computers
2. How Can a Computer be Programmed to Use a Language
3. Neuron Nets
4. Theory of the Size of a Calculation
5. Self-Improvement
6. Abstractions
7. Randomness and Creativity

We'll look at least to 3 of these points:

- Self-Improvement $\longrightarrow$ Learning
- Neuron Nets $\longrightarrow$ Artificial Neural Networks
- Randomness $\longrightarrow$ Genetic Algorithms

Self-Improvement and Learning

A computer program is said to **learn** from experience **E** with respect to some class of **task T** and a **performance measure P**, if its performance at tasks in **T**, as measured by **P** improves because of experience **E**.

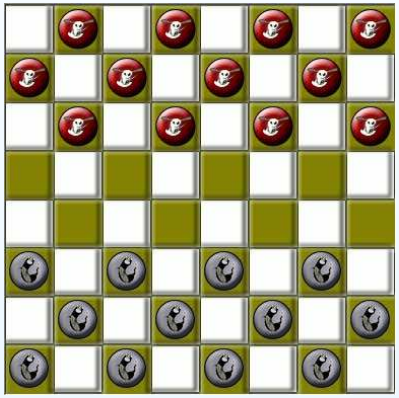
Self-Improvement and Learning

A computer program is said to **learn** from experience **E** with respect to some class of **task T** and a **performance measure P**, if its performance at tasks in **T**, as measured by **P** improves because of experience **E**.

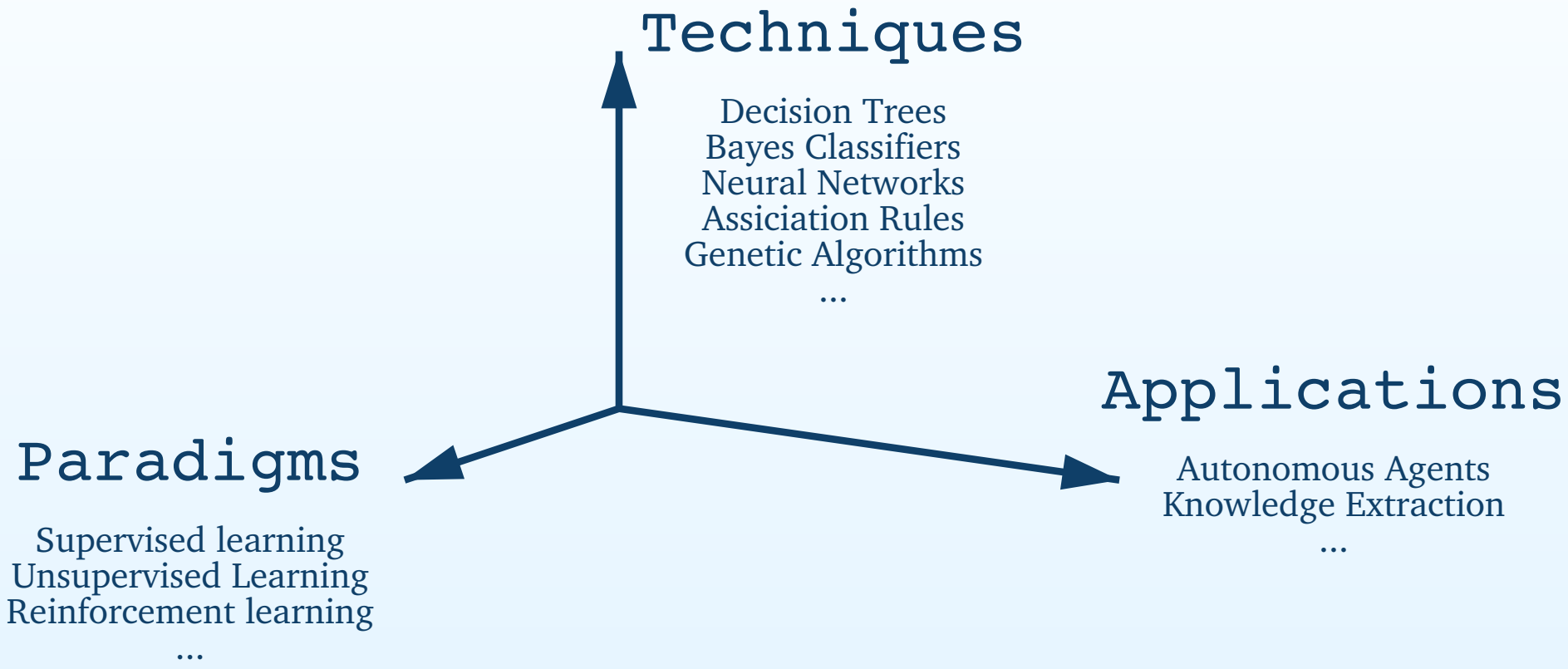
Machine Learning: the study or development of models and algorithms that make systems automatically improve their performance during execution.

Example: Playing Checkers

- Task **T**
  - Play checkers
- Experience **E**
  - Games played with other players
  - Games played against itself
- Performance **P**
  - Percentage of games won



Apps – Paradigms – Techniques



Applications: “What’s your flava?”

- Self customizing programs
  - Newsreader that learns user interests
  - Email anti-spam filters
  - . . .
- Data mining
  - medical records $\longrightarrow$ medical knowledge
  - using historical data to improve decisions
  - . . .
- Software applications we can’t program by hand
  - autonomous driving
  - speech recognition
  - . . .

Learning Paradigms

Imagine an organism or machine that experiences a series of sensory inputs:

$$\mathbf{E} = x_1, x_2, x_3, x_4, \ldots$$

Learning Paradigms

Imagine an organism or machine that experiences a series of sensory inputs:

$$\mathbf{E} = x_1, x_2, x_3, x_4, \ldots$$

- *Supervised learning*: given the **desired outputs** $y_1, y_2, \ldots$, learn to produce the correct output given new input

Learning Paradigms

Imagine an organism or machine that experiences a series of sensory inputs:

$$\mathbf{E} = x_1, x_2, x_3, x_4, \ldots$$

- *Supervised learning*: given the **desired outputs** $y_1, y_2, \ldots$, learn to produce the correct output given new input
- *Unsupervised learning*: **exploit regularities in** $\mathbf{E}$ to **build a representation** that can be used for reasoning or prediction

Learning Paradigms

Imagine an organism or machine that experiences a series of sensory inputs:

$$\mathbf{E} = x_1, x_2, x_3, x_4, \ldots$$

- *Supervised learning*: given the **desired outputs** $y_1, y_2, \ldots$, learn to produce the correct output given new input
- *Unsupervised learning*: **exploit regularities in** $\mathbf{E}$ to **build a representation** that can be used for reasoning or prediction
- *Reinforcement learning*: **producing actions** $a_1, a_2, \ldots$ which affect the environment, and **receiving rewards** $r_1, r_2, \ldots$ learn to act in a way that **maximises rewards** in the long term

Course Outline [*Tentative*]

- Probability for Dataminers
  - Information Gain
  - Probability Basics
- Model Selection Techniques
  - Cross-Validation
  - Model Complexity
- Optimization Techniques
  - Gradient Based
  - Genetic Algorithms
- Supervised Learning
  - Decision Trees
  - Bayes Classifi ers
  - Neural Networks
- Unsupervised Learning
  - Clustering
  - Association Rules
- Reinforcement Learning
  - MDP & Q-Learning

Supervised Learning

– Basics & Examples –

Supervised Learning

- **The Task T:** extract from a finite set of examples a *model* of the observed phenomenon to be used in the future for prediction or decision making about it

Supervised Learning

- **The Task T:** extract from a finite set of examples a *model* of the observed phenomenon to be used in the future for prediction or decision making about it
- **The Experience E:** a set of *examples* for the desired “behaviour” pre-processed by an expert [the supervisor] as pairs input / output

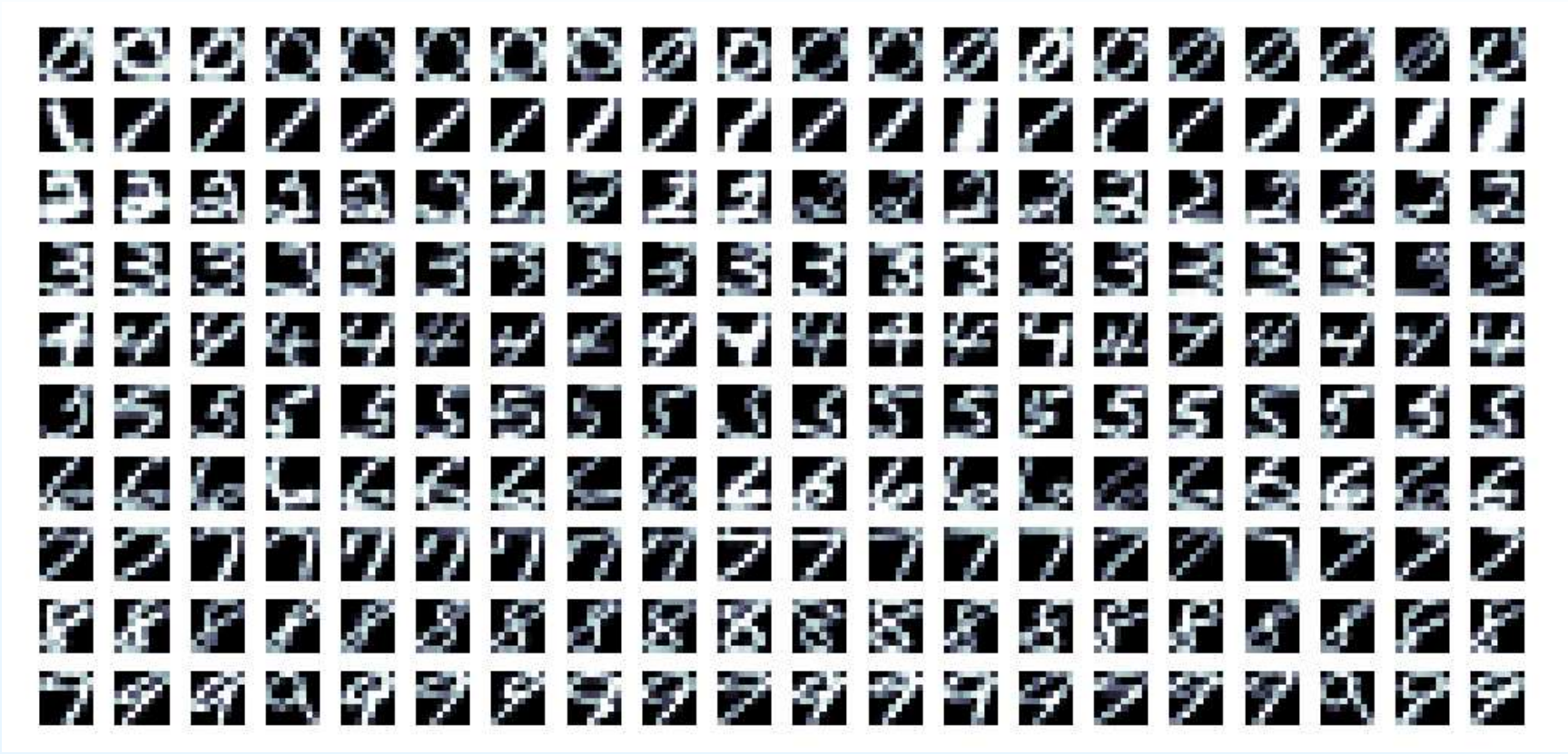
Supervised Learning

- **The Task T:** extract from a finite set of examples a *model* of the observed phenomenon to be used in the future for prediction or decision making about it
- **The Experience E:** a set of *examples* for the desired “behaviour” pre-processed by an expert [the supervisor] as pairs input / output
- **The Performance P:** is the measure of the distance between the desired output for new examples and the output provided by the model

Supervised Learning Examples

- Predict whether a patient, hospitalized due to a heart attack, will have a second heart attack. The prediction has to be based on demographic, diet, and clinical measurements.
- Predict the price of a stock in six months from now, on the basis of company performance measures and economic data.
- Estimate the amount of glucose in the blood of a diabetic person, from the infrared absorption of that person's blood.
- Identify the risk factors for prostate cancer, based on clinical and demographic variables.
- Identify the numbers in handwritten ZIP code, from a digitalized image.

Example: ZIP Codes Images



There are 7291 training observations and 2007 test observations.
Each observation is a 16 x 16 grayscale image

Terminology

Classification: The desired outputs y_i are discrete class labels. The goal is to classify new inputs correctly.

Regression: The desired outputs y_i are continuous valued. The goal is to predict the output accurately for new inputs.

Terminology

Classification: The desired outputs y_i are discrete class labels. The goal is to classify new inputs correctly.

Regression: The desired outputs y_i are continuous valued. The goal is to predict the output accurately for new inputs.

Generalization: The capability of a model to correctly predict new samples never seen during the training.

Terminology

Classification: The desired outputs y_i are discrete class labels. The goal is to classify new inputs correctly.

Regression: The desired outputs y_i are continuous valued. The goal is to predict the output accurately for new inputs.

Generalization: The capability of a model to correctly predict new samples never seen during the training.

Inductive Hypothesis: *A solution that approximate the target function over a suffi ciently large set of training examples will also approximate the target function over unobserved examples*

Probability for Dataminers

– Information Gain –

Information and Bits

Your mission, if you decide to accept it, will be:

"Transmit a set of independent random samples of X over a binary serial link."

Information and Bits

Your mission, if you decide to accept it, will be:

"Transmit a set of independent random samples of X over a binary serial link."

1. Starring at X for a while, you notice that it has olny four possible values: A, B, C, D

Information and Bits

Your mission, if you decide to accept it, will be:

"Transmit a set of independent random samples of X over a binary serial link."

1. Starring at X for a while, you notice that it has olny four possible values: A, B, C, D
2. You decide to transmit the data encoding each reading with two bits:

$$A = 00, B = 01, C = 10, D = 11.$$

Mission Accomplished!

Information and “Fewer Bits”

Your mission, if you decide to accept it, will be:

*“The previous code uses 2 bits for symbol.
Knowing that the probabilities are not equal: P(X=A)=1/2, P(X=B)=1/4, P(X=C)=1/8, P(X=D)=1/8, invent a coding for your transmission that only uses 1.75 bits on average per symbol.”*

Information and “Fewer Bits”

Your mission, if you decide to accept it, will be:

*“The previous code uses 2 bits for symbol.
Knowing that the probabilities are not equal: P(X=A)=1/2, P(X=B)=1/4, P(X=C)=1/8, P(X=D)=1/8, invent a coding for your transmission that only uses 1.75 bits on average per symbol.”*

1. You decide to transmit the data encoding each reading with a different number of bits:

$$A = 0, B = 10, C = 110, D = 111.$$

Mission Accomplished!

Information and Entropy

Suppose X can have one of m values with probability

$$P(X = V_1) = p_1, \ldots, P(X = V_m) = p_m.$$

What's the smallest possible number of bits, on average, per symbol, needed to transmit a stream of symbols drawn from X's distribution?

Information and Entropy

Suppose X can have one of m values with probability

$$P(X = V_1) = p_1, \ldots, P(X = V_m) = p_m.$$

What's the smallest possible number of bits, on average, per symbol, needed to transmit a stream of symbols drawn from X's distribution?

$$\begin{aligned} H(X) &= -p_1 \log_2 p_1 - p_2 \log_2 p_2 - \ldots - p_m \log_2 p_m \\ &= -\sum_{j=1}^{m} p_j \log_2 p_j = \textit{Entropy} \text{ of } X \end{aligned}$$

Information and Entropy

Suppose X can have one of m values with probability

$$P(X = V_1) = p_1, \ldots, P(X = V_m) = p_m.$$

What's the smallest possible number of bits, on average, per symbol, needed to transmit a stream of symbols drawn from X's distribution?

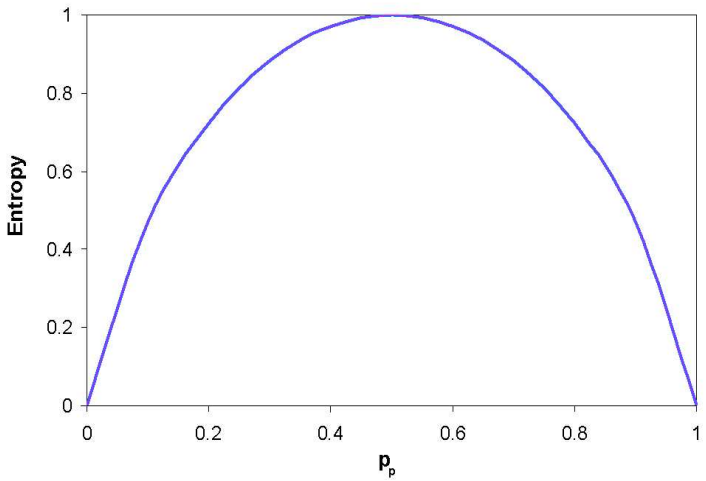
$$\begin{aligned} H(X) &= -p_1 \log_2 p_1 - p_2 \log_2 p_2 - \ldots - p_m \log_2 p_m \\ &= -\sum_{j=1}^{m} p_j \log_2 p_j = \textit{Entropy} \text{ of } X \end{aligned}$$

"Good idea! But what is entropy anyway?"

Entropy: “What is it anyway?”

Simple Case:

- X has 2 values $\oplus$ and $\ominus$
- $p_\oplus$ probability of $\oplus$
- $p_\ominus = 1 - p_\oplus$ probability of $\ominus$

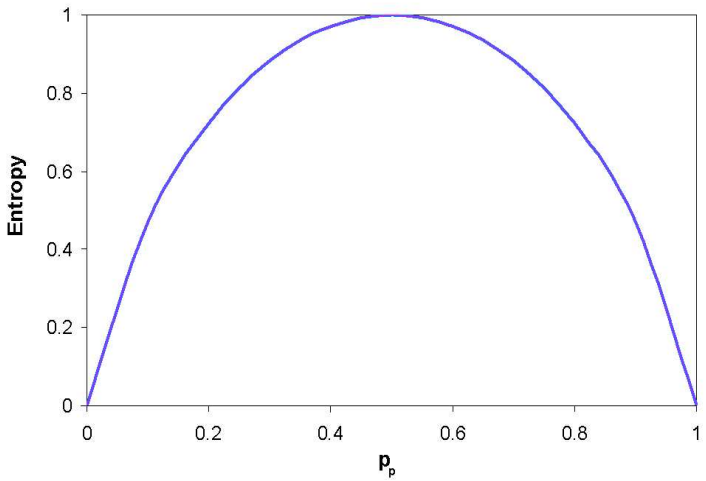


$$H(X) = -p_\ominus \log_2 p_\ominus - p_\oplus \log_2 p_\oplus$$

Entropy: “What is it anyway?”

Simple Case:

- X has 2 values $\oplus$ and $\ominus$
- $p_\oplus$ probability of $\oplus$
- $p_\ominus = 1 - p_\oplus$ probability of $\ominus$



$$H(X) = -p_\ominus \log_2 p_\ominus - p_\oplus \log_2 p_\oplus$$

Entropy measures “disorder” or “uniformity in distribution”

1. *High Entropy*: X is very “disordered” $\rightarrow$ “boring”
2. *Low Entropy*: X is very “ordered” $\rightarrow$ “interesting”

Specifi c Conditional Entropy

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
Hystory	No
Math	Yes

Suppose we are interested in predicting output Y from input X where

- X = University subject
- Y = Likes the movie “Gladiator”

Specifi c Conditional Entropy

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
Hystory	No
Math	Yes

Suppose we are interested in predicting output Y from input X where

- X = University subject
- Y = Likes the movie “Gladiator”

From this data we can estimate

- P(Y = Yes) = 0.5
- P(X = Math) = 0.5
- P(Y = Yes | X = History) = 0

Specifi c Conditional Entropy

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
Hystory	No
Math	Yes

Suppose we are interested in predicting output Y from input X where

- X = University subject
- Y = Likes the movie “Gladiator”

Definition of Specific Conditional Entropy:

- H(Y|X=v): *the entropy of Y only for those records in which X has value v*
  - H(Y|X=Math) = 1
  - H(Y|X=History) = 0

Conditional Entropy

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
Hystory	No
Math	Yes

Definition of Conditional Entropy H(Y|X):

- *The average Y specifi c conditional entropy*
- *Expected number of bits to transmit Y if both sides will know the value of X*
- $\sum_j P(X = v_j)H(Y|X = v_j)$

Conditional Entropy

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
Hystory	No
Math	Yes

Definition of Conditional Entropy H(Y|X):

- $\sum_j P(X = v_j)H(Y|X = v_j)$

v_j	$P(X = v_j)$	$H(Y|X = v_j)$
Math	0.5	1
Hystory	0.25	0
CS	0.25	0

H(Y|X) = ?

Conditional Entropy

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
Hystory	No
Math	Yes

Definition of Conditional Entropy H(Y|X):

- $\sum_j P(X = v_j)H(Y|X = v_j)$

v_j	$P(X = v_j)$	$H(Y|X = v_j)$
Math	0.5	1
Hystory	0.25	0
CS	0.25	0

H(Y|X)= 0.5 $\times$ 1 + 0.25 $\times$ 0 + 0.25 $\times$ 0 = 0.5

Good, but what about Machine Learning?

Information Gain

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
Hystory	No
Math	Yes

I must transmit Y on a binary serial line.
How many bits on average would it save me if both ends of the line knew X?

$$\begin{aligned} IG(Y|X) &= H(Y) - H(Y|X) \\ &= 1 - 0.5 = 0.5 \end{aligned}$$

Information Gain

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
Hystory	No
Math	Yes

I must transmit Y on a binary serial line.
How many bits on average would it save me if both ends of the line knew X?

$$\begin{aligned} IG(Y|X) &= H(Y) - H(Y|X) \\ &= 1 - 0.5 = 0.5 \end{aligned}$$

Information Gain measures the "information" provided by X to predict Y

This IS about Machine Learning!

Relative Information Gain

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
Hystory	No
Math	Yes

I must transmit Y on a binary serial line.
What fraction of the bits on average would it save me if both ends of the line knew X?

$$\begin{aligned} RIG(Y|X) &= (H(Y) - H(Y|X))/H(Y) \\ &= (1 - 0.5)/1 = 0.5 \end{aligned}$$

Well, we'll find soon Information Gain and Relative Information gain talking about supervised learning with Decision Trees ...

Why is Information Gain Useful?

Your mission, if you decide to accept it, will be:

"Predict whether someone is going live past 80 years."

From historical data you might find:

- IG(LongLife | HairColor) = 0.01
- IG(LongLife | Smoker) = 0.2
- IG(LongLife | Gender) = 0.25
- IG(LongLife | LastDigitOfSSN) = 0.00001

What you should look at?