Information Theory

Thursday 1st December, 2016

Motivation - Aviation Mechanics

- recent research finds
  - most delays caused by duration of maintenance

Motivation - Aviation Mechanics

Israeltech challenge

- recent research finds
    - most delays caused by duration of maintenance
- 60% of maintenance time
    - "could you pass me that 8cm star shaped screwdriver with a diameter of 1.5cm and electrical insulation rating 3"
    - a technicians toolbox contains over 200 different tools

Motivation - Aviation Mechanics

- recent research finds
  - most delays caused by duration of maintenance
- 60% of maintenance time
  - "could you pass me that 8cm star shaped screwdriver with a diameter of 1.5cm and electrical insulation rating 3"
  - a technicians toolbox contains over 200 different tools
  - 99.99999% a technician uses:
    - two types of screwdriver
    - one wrench
    - one set of pliers

Motivation - Aviation Mechanics

- save time - efficient terminology
  - screwdriver A, screwdriver B
  - wrench A
  - pliers A

Motivation - Aviation Mechanics

Motivation - Aviation Mechanics

Israeltech challenge

The english language is not the boss of me

- who said we have to use english
- we can use a made up word for each tool
- let's say we want to find a word S_A as a name for screwdriver A
- what's the optimal length for all words?

Optimizing Code Length

various measures can be used

- maximal code length for a word
- minimum variance

in our case

- average code length L for words $w_1, \ldots, w_M$
- length $l_1, \ldots, l_M$

$$\begin{aligned} L &= E\left[l\right] \\ &= \sum_{j=1}^{M} p_j \cdot l_j \end{aligned}$$

Optimizing Average Code Length

The optimal code length

$$L^* = \min_{l_1,\ldots,l_M} \sum_{j=1}^{M} p_j \cdot l_j$$

first attempt:

Optimizing Average Code Length

The optimal code length

$$L^* = \min_{l_1,\ldots,l_M} \sum_{j=1}^{M} p_j \cdot l_j$$

first attempt:

- $l_j = 1/p_j$

$$L = \sum_{j=1}^{M} p_j \cdot 1/p_j = \sum_{j=1}^{M} 1 = M$$

Are there no rules?

there should be some rules

Are there no rules?

there should be some rules

Creative Naming

a couple has 3 daughters:

- Sarah
- Leah
- Sarah Leah

Creative Naming

"Sarah Leah come here please"

Uniquely Decodable Codes

Uniquely Decodable Codes
single meaning for each sentence

Instanteneous/Prefix Codes
no word is a prefix of another

- decoding can start as soon as a word is completed

Uniquely Decodable Codes

Kraft-McMillan inquality

a uniquely decodable code with M words over D symbols must satisfy

$$\sum_{i=1}^{M} D^{-l_i} \leq 1$$

board - binary

The Shortest Average Code Length

a uniquely decodable code L satisfies

$$L \geq L^*$$

The Shortest Average Code Length

a uniquely decodable code L satisfies

$$L \geq L^*$$

where L^* is given by

$$\forall i = 1, \ldots, M : \;\; l_i^* = -\log_D p_i$$

The Shortest Average Code Length

Israeltech challenge

a uniquely decodable code L satisfies

$$L \geq L^*$$

where L^* is given by

$$\forall i = 1, \ldots, M : \; l_i^* = -\log_D p_i$$

the mean code length L^* is the **Entropy**

$$\boxed{H_D(X) = E_x\left[-\log_D(p_x)\right]}$$

The Shortest Average Code Length

Israeltech challenge

a uniquely decodable code L satisfies

$$L \geq L^*$$

where L^* is given by

$$\forall i = 1, \ldots, M : \; l_i^* = -\log_D p_i$$

the mean code length L^* is the $\boxed{\textbf{Entropy}}$

$$\boxed{H(X) = E_x\left[-\log\left(p_x\right)\right]}$$

binary

Digital World

- in theory we can achieve L^*
- building the code - sometimes hard
- sometimes impossible

$$D = 2, \;\; p_i = 2/3, \;\; l_i^* = -log\,(p_i) = 0.58496..$$

Entropy

- intuitively
  the amount of information in a probability distribution

Entropy

- intuitively
  the amount of information in a probability distribution

- what's the optimal code length for stating
  - the obvious
  - something that always happen $P\left(X=x\right)=1$
  - something you already know

discuss

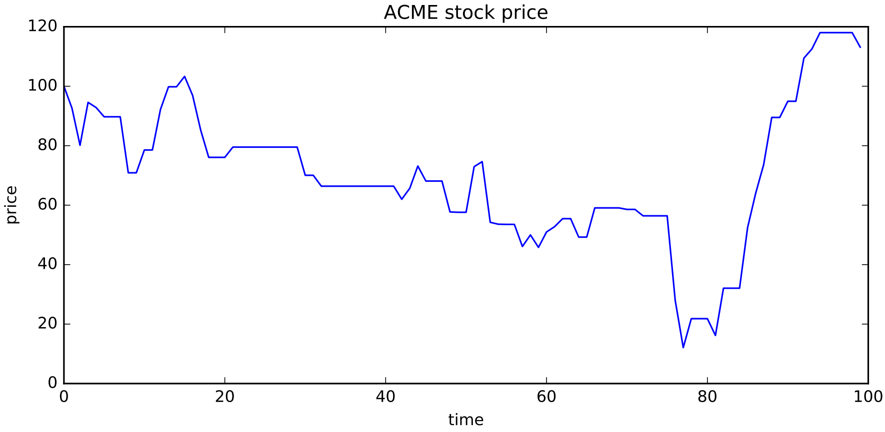
Entropy

Stock price change monitor

- change = up, down, no change

Entropy

Stock price change monitor



Entropy

Stock price change monitor

- change = up, down, no change
- how many bits to encode?

Entropy

IsraelTech challenge

Stock price change monitor

- change = up, down, no change
- how many bits to encode?
- 3 options $\Rightarrow \log_2(3) = 1.585$ board - binary 3
- 2 bits

Entropy

IsraelTech challenge

Stock price change monitor

- change = up, down, no change
- how many bits to encode?
- 2 bits
- $P(up) = P(down) = 0.25$
- $P(no - change) = 0.5$

Entropy

Stock price change monitor

- change = up, down, no change
- how many bits to encode?
- 2 bits
- $P\left(up\right) = P\left(down\right) = 0.25$
- $P\left(no - change\right) = 0.5$
- encode

no change	0
down	10
up	11

Entropy

IsraelTech challenge

Stock price change monitor

- change = up, down, no change
- $P\left(up\right) = P\left(down\right) = 0.25$
- $P\left(no - change\right) = 0.5$
- encode

no change	0
down	10
up	11

- average code length $0.5 \cdot 1 + 0.25 \cdot 2 + 0.25 \cdot 2 = 1.5$

Entropy

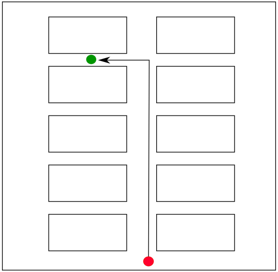
Stock price change monitor

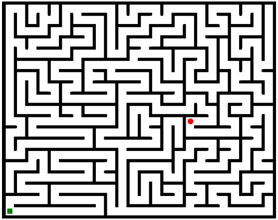
- change = up, down, no change
- $P(up) = P(down) = 0.25$
- $P(no - change) = 0.5$
- encode

no change	0
down	10
up	11

- average code length $0.5 \cdot 1 + 0.25 \cdot 2 + 0.25 \cdot 2 = 1.5$
- binary entropy $H(\{0.5, 0.25, 0.25\}) = 1.5$ board

Entropy





Copilot 1	only speaks when you need to turn	only speaks when you have a turn to take
Copilot 2	"stright, straight, ..., straight, turn left"	"continue, continue, turn left here, ..."

Sample Entropy

you are given samples $X_1, \ldots, X_n$

- how do you compute $H(X)$?

Sample Entropy

Israeltech challenge

you are given samples $X_1, \ldots, X_n$

- how do you compute $H(X)$?
- MLE estimates of the probabilities

$$\forall x : \hat{p}_x = \frac{1}{n} \sum_{i=1}^{N} \mathbb{1}\{X_i = x\}$$

$$H(X) = -\sum_{x} \hat{p}_x(x) \log(\hat{p}_x(X))$$

Entropy - Lower Bound

absolute certainty about the data X	$=$	information is obvious

$$P(X = x) = 1$$

$$H(X) = \sum_x p_x \log(1/p_x)$$
$$= 0$$

Entropy - Upper Bound

all outcomes are equally likely

- X has M unique values (words)
- $\forall i = 1, \ldots, M : \;\; p_i = 1/M$

$$
\begin{aligned}
H(X) &= -\sum_{i=1}^{M} 1/M \log(1/M) \\
&= \log(M) \sum_{i=1}^{M} 1/M \\
&= \log(M)
\end{aligned}
$$

Entropy - Bounds

Entropy is bounded

$$\boxed{0 \leq H\left(X\right) \leq \log\left(M\right)}$$

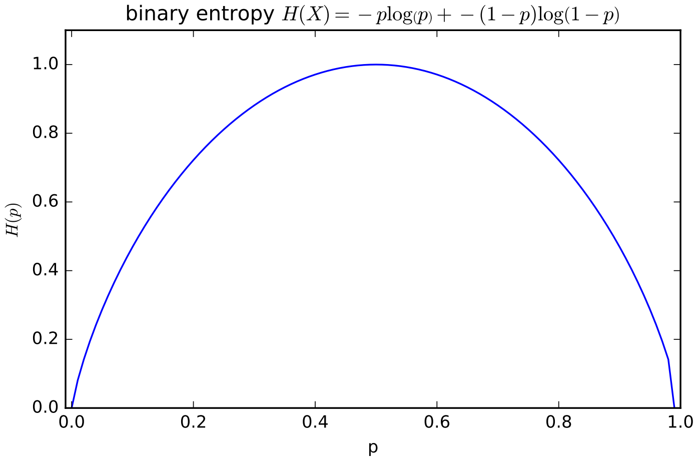
Binary Entropy

X is a random variable (a coin)

$$\begin{aligned} H(X) &= \sum_x p_x \log\left(\frac{1}{p_x}\right) \\ &= -p\log(p) - (1-p)\log(1-p) \end{aligned}$$

Binary Entropy

X is a random variable (a coin)



What about variance

both measure the amount of variation

- variance is a function of the **data**
- entropy is a function of the **data probabilities**

What about variance

both measure the amount of variation

- variance is a function of the **data**
- entropy is a function of the **data probabilities**

toy example

- data is weather on Mercury
- temperature is $-173C/427C$ with $p_{hot} = 0.6$

$$V[X] = 86400, \quad H(X) = 0.971$$

What about variance

toy example

- data is weather on Mercury
- temperature is $-173C/427C$ with $p_{hot} = 0.6$

$$V\left[X\right] = 86400, \quad H\left(X\right) = 0.971$$

- a mission to mercury - variance
- predicting weather in mercury - entropy
  - temprature on planet X $-1.73C/4.27C$ with $p_{hot} = 0.6$

Joint Entropy

X, Y discrete random variables

$$H\left(X,Y\right) = -\sum_{x}\sum_{y} p\left(x,y\right)\log\left(p\left(x,y\right)\right)$$

Conditional Entropy

X, Y discrete random variables

$$H\left(Y \mid X\right) = \sum_{x} P\left(X\right) H\left(Y \mid X = x\right)$$

Conditional Entropy

Israeltech challenge

X, Y discrete random variables

$$H(Y \mid X) = \sum_x P(X)\, H(Y \mid X = x)$$
$$= -\sum_x P(x) \sum_y P(y \mid x) \log P(y \mid x)$$

Conditional Entropy

X, Y discrete random variables

$$\begin{aligned} H(Y \mid X) &= \sum_x P(X)\, H(Y \mid X = x) \\ &= -\sum_x P(x) \sum_y P(y \mid x) \log P(y \mid x) \\ &= -\sum_x \sum_y P(X, Y) \log P(y \mid x) \end{aligned}$$

Conditional Entropy

X, Y discrete random variables

$$
\begin{aligned}
H(Y \mid X) &= \sum_x P(X)\, H(Y \mid X = x) \\
&= -\sum_x P(x) \sum_y P(y \mid x) \log P(y \mid x) \\
&= -\sum_x \sum_y P(X, Y) \log P(y \mid x) \\
&= -E_{XY}\left[\log P(Y \mid X)\right]
\end{aligned}
$$

Conditional Entropy

Information can't hurt

$$H\left(X\right) \geq H\left(X \mid Y\right)$$

Conditional Entropy

Information can't hurt

$$H\left(X\right) \geq H\left(X \mid Y\right)$$

when is

$$H\left(X\right) = H\left(X \mid Y\right)$$

Conditional Entropy

Information can't hurt

$$H(X) \geq H(X \mid Y)$$

when is

$$H(X) = H(X \mid Y)$$

X and Y are independent

Chain Rule

- X an email
- Y a follow up email

what is the minimal average message length $H\left(X,Y\right)$

Chain Rule

- X an email
- Y a follow up email

what is the minimal average message length $H\left(X,Y\right)$

- Y - "p.s. the party is at 5"
- Y - "I almost forgot. In an unrelated matter..."
- Y - "" (no need for a follow up)

Chain Rule

$$H(X, Y) = H(X) + H(Y \mid X)$$

discuss

Chain Rule

a previous email Z

$$H\left(X, Y \mid Z\right) = H\left(X \mid Z\right) + H\left(Y \mid X, Z\right)$$

Language Gaps

explaining

- no shared language/terms
- takes more time
- higher average message length

Cross Entropy

- P data distribution for X
- Q another distribution over X

$$\begin{aligned} H(P,Q) &= E_P\left[-\log Q(X)\right] \\ &= -\sum_x P(X)\log Q(X) \end{aligned}$$

- the average code length when using Q

Relative Entropy

Kullback-Leibler Divergence (KL)

- P data distribution for X
- Q another distribution over X
- expected number of extra **bits**
  - using Q to describe $X \sim P$

$$D\left(P \mid\mid Q\right) = \sum_{x} P\left(x\right) \log \frac{P\left(X\right)}{Q\left(X\right)}$$

Relative Entropy

Kullback-Leibler Divergence (KL)

- P data distribution for X
- Q another distribution over X
- expected number of extra **bits**
  - using Q to describe $X \sim P$

$$D\left(P \mid\mid Q\right) = \sum_{x} P\left(x\right) \log \frac{P\left(X\right)}{Q\left(X\right)}$$

- nonnegative (0 for $P = Q$)
- distance - not a metric

Mutual Information

- measures information about X gained from Y

$$I\left(X;Y\right)=\sum_{x}\sum_{y}P\left(X,Y\right)\log\frac{P\left(X,Y\right)}{P\left(X\right)P\left(Y\right)}$$

Mutual Information

- measures information about X gained from Y

$$I\left(X;Y\right)=\sum_{x}\sum_{y}P\left(X,Y\right)\log\frac{P\left(X,Y\right)}{P\left(X\right)P\left(Y\right)}$$

- nonnegative
- $I\left(X;X\right)=H\left(X\right)$

Mutual Information

conditional entropy $H\left(Y \mid X\right)$

Mutual Information

conditional entropy $H(Y \mid X)$

- unexplained information after knowing X

$$H(Y \mid X) = H(Y) - I(Y; X)$$

Mutual Information

conditional entropy $H(Y \mid X)$

- unexplained information after knowing X

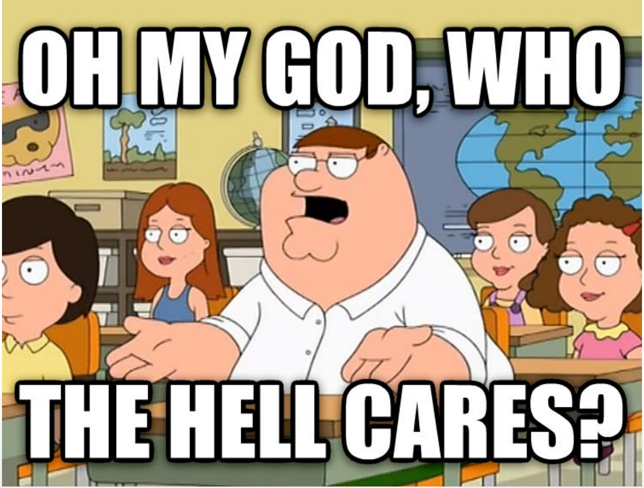
$$H(Y \mid X) = H(Y) - I(Y; X)$$

relative entropy

- mutual information = KL divergence(joint, marginals)

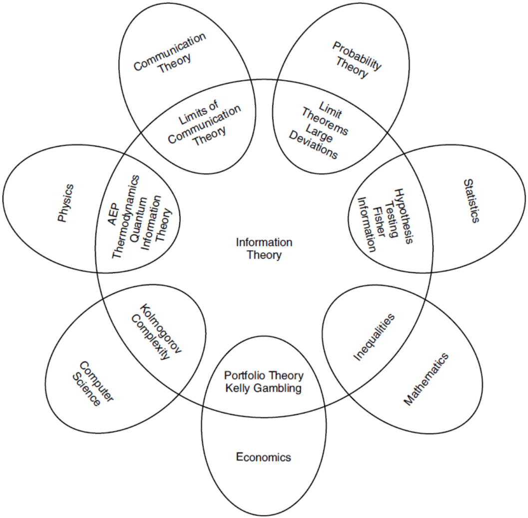
$$I(X; Y) = D(P(X, Y) \,||\, P(X) P(Y))$$

You may be tempted to ask



Information Theory - Importance

Israeltech challenge



Decision Trees

medi-bot - medical diagnosis chat bot

1. while no diagnosis:
   1.1 ask a yes/no question

Decision Trees

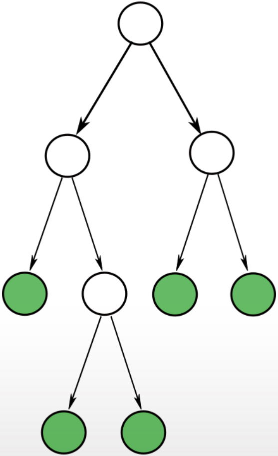
medi-bot - medical diagnosis chat bot

1. while no diagnosis:
    1.1 ask a yes/no question

- assume all conditions are diagnosable
- enough questions - medi-bot is able to diagnose

Decision Trees

medi-bot



green leafs

Another motivation - 20 questions

game

Another motivation - 20 questions

IsraelTech challenge

- is it a person?
- no
- is it a household item?
- yes
- ...

Another motivation - 20 questions

- also a decision tree
- each question divides the possible answers

Under Pressure

- perfect diagnosis
- on average 10 questions
- can't ask more than 5 time constraints, overfitting
- how do we build the tree?

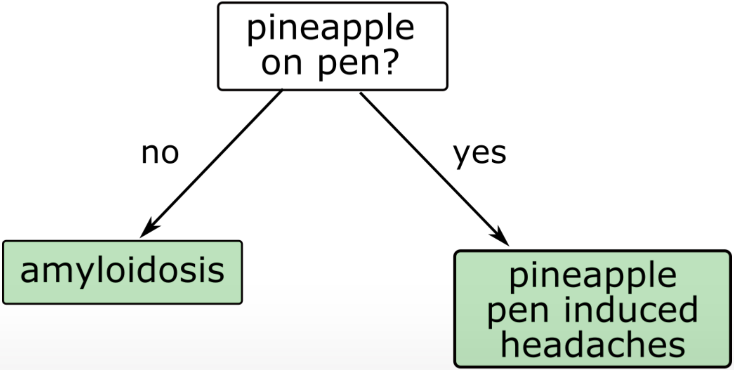
What's the best question?

optimal question = perfect decision (classification)

What's the best question?

Israeltech challenge

optimal question = perfect decision (classification)



What's the best question in practice?

discuss

What's the best question in practice?

idea - best question q

- maximum information

What's the best question in practice?

idea - best question q

- maximum information
- information about the data D which data

What's the best question in practice?

idea - best question q

- maximum information
- information about the data D which data
- the average information after asking?

$$E\left[H\left(D \mid q\right)\right] = P\left(q = yes\right) H\left(D \mid q = yes\right) + P\left(q = no\right) H\left(D \mid q = no\right)$$

What's the best question in practice?

IsraelTech challenge

idea - best question q

- maximum information
- information about the data D which data
- the average information after asking?

$$E\left[H\left(D \mid q\right)\right] = P\left(q = yes\right) H\left(D \mid q = yes\right) + P\left(q = no\right) H\left(D \mid q = no\right)$$

- information = reduction in entropy

$$I\left(D; q\right) = H\left(D\right) - H\left(D \mid q\right)$$

Simple Example

will it rain? (1 question)

index	cloudy (c)	rain forecasted (f)	rain
1	yes	yes	yes
2	no	no	no
3	yes	no	no
4	no	yes	no
5	yes	no	yes

Simple Example

is rain forcasted ?

index	cloudy (c)	rain forecasted (f)	rain
1	yes	yes	yes
2	no	no	no
3	yes	no	no
4	no	yes	no
5	yes	no	yes

rain is forcasted: $rain = \{ \; yes, no \; \}$

rain is not forcasted: $rain = \{ \; no, no, yes \; \}$

Simple Example

is it cloudy?

index	cloudy (c)	rain forecasted (f)	rain
1	yes	yes	yes
2	no	no	no
3	yes	no	no
4	no	yes	no
5	yes	no	yes

it's cloudy: $rain = \{\ yes, no, yes\ \}$

it's not cloudy: $rain = \{\ no, no\ \}$

Simple Example

Israeltech challenge

which question do you ask?

index	cloudy (c)	rain forecasted (f)	rain
1	yes	yes	yes
2	no	no	no
3	yes	no	no
4	no	yes	no
5	yes	no	yes

Simple Example

is rain forcasted ?

index	cloudy (c)	rain forecasted (f)	rain
1	yes	yes	yes
2	no	no	no
3	yes	no	no
4	no	yes	no
5	yes	no	yes

$$H\left(rain \mid f\right) = \frac{2}{5} H\left(rain \mid f = 1\right) + \frac{3}{5} H\left(rain \mid f = 0\right)$$

Simple Example

is rain forcasted ?

index	cloudy (c)	rain forecasted (f)	rain
1	yes	yes	yes
2	no	no	no
3	yes	no	no
4	no	yes	no
5	yes	no	yes

$$H\left(rain \mid f = 1\right) = -0.5 \log\left(0.5\right) - 0.5 \log\left(0.5\right) = 1$$

Simple Example

is rain forcasted ?

index	cloudy (c)	rain forecasted (f)	rain
1	yes	yes	yes
2	no	no	no
3	yes	no	no
4	no	yes	no
5	yes	no	yes

$$H\left(rain \mid f=0\right) = -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} = 0.9183$$

Simple Example

is rain forcasted ?

index	cloudy (c)	rain forecasted (f)	rain
1	yes	yes	yes
2	no	no	no
3	yes	no	no
4	no	yes	no
5	yes	no	yes

$$H\left(rain \mid f\right) = \frac{2}{5} \cdot 1 + \frac{3}{5} \cdot 0.9183 = 0.951$$

Simple Example

is it cloudy?

index	cloudy (c)	rain forecasted (f)	rain
1	yes	yes	yes
2	no	no	no
3	yes	no	no
4	no	yes	no
5	yes	no	yes

$$H\left(rain \mid c\right) = \frac{3}{5} H\left(rain \mid c = 1\right) + \frac{2}{5} H\left(rain \mid c = 0\right)$$

Simple Example

is it cloudy?

index	cloudy (c)	rain forecasted (f)	rain
1	yes	yes	yes
2	no	no	no
3	yes	no	no
4	no	yes	no
5	yes	no	yes

$$H\left(rain \mid c\right) = \frac{3}{5} H\left(rain \mid c = 1\right)$$

Simple Example

is it cloudy?

index	cloudy (c)	rain forecasted (f)	rain
1	yes	yes	yes
2	no	no	no
3	yes	no	no
4	no	yes	no
5	yes	no	yes

$$H\left(rain \mid c\right) = \frac{3}{5}\left[-\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3}\right]$$

Simple Example

is it cloudy?

index	cloudy (c)	rain forecasted (f)	rain
1	yes	yes	yes
2	no	no	no
3	yes	no	no
4	no	yes	no
5	yes	no	yes

$$H\left(rain \mid c\right) = \frac{3}{5} \cdot 0.9183 = 0.551$$

Simple Example

index	cloudy (c)	rain forecasted (f)	rain
1	yes	yes	yes
2	no	no	no
3	yes	no	no
4	no	yes	no
5	yes	no	yes

$$H\left(rain \mid c\right) = 0.551 < 0.951 = H\left(rain \mid f\right)$$

Simple Example

index	cloudy (c)	rain forecasted (f)	rain
1	yes	yes	yes
2	no	no	no
3	yes	no	no
4	no	yes	no
5	yes	no	yes

$$H\left(rain \mid c\right) = 0.551 < 0.951 = H\left(rain \mid f\right)$$

$$H\left(rain\right) = 0.971$$

Simple Example

index	cloudy (c)	rain forecasted (f)	rain
1	yes	yes	yes
2	no	no	no
3	yes	no	no
4	no	yes	no
5	yes	no	yes

$$I\left(rain \mid question\right) = H\left(rain\right) - H\left(rain \mid question\right)$$

$$I\left(rain \mid c\right) = 0.42 < 0.02 = I\left(rain \mid f\right)$$

Information Gain

- question choosing criterion for data D'

$$q = \arg\max_{q} \; I\left(D'; q\right)$$

Information Gain

- question choosing criterion for data D'

$$q = \arg\max_{q} \; I\left(D'; q\right)$$

- optimal?

Information Gain

- question choosing criterion for data D'

$$q = \arg\max_{q} \; I\left(D'; q\right)$$

- optimal?
- no - useful heuristic

Decision Trees

- important class of classifiers
- extensions used in:
  - computer vision
  - halo - matching

Information Theory

- wide range of applications
  - communication
- concepts
  - information (entropy)
  - mutual information
  - side information
  - relative entropy

Credits

figures

- maze image - upload.wikimedia.org/wikipedia/commons/b/bf/Maze_01.svg
- borat - www.quickmeme.com/meme/3opuzs
- parking fail - www.flickr.com/photos/thienzieyung/5318972121
- fry - creative naming - i can haz falafel
- who cares - www.livememe.com/enys03l