University of BRISTOL

Machine Learning

Basic Probabilities

Carl Henrik Ek - carlhenrik.ek@bristol.ac.uk

September 26, 2017

http://www.carlhenrik.com

Introduction

Assumptions

- Observations cannot be argued with
- Interpretations of observations are relative to assumptions
- Good assumptions structures the world in a useful way (reduce energy)

Flat Earth Society[1]

[1] Flat Earth Society Wiki

Flat Earth Society[2]

"There is no Flat Earth Conspiracy. NASA is not hiding the shape of the earth from anyone. The purpose of NASA is not to 'hide the shape of the earth' or 'trick people into thinking it's round' or anything of the sort. There is a Space Travel Conspiracy. The purpose of NASA is to fake the concept of space travel to further America's militaristic dominance of space"

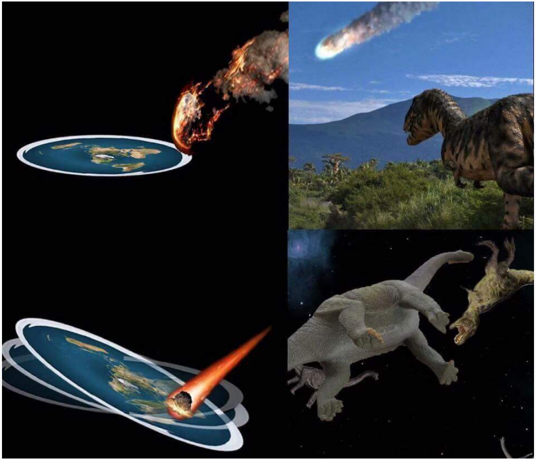
[2] `Globe Conspiracy`

Flat Earth Society[2]

"NASA's early rocket research is well documented to have been a complete failure, plagued by one disaster after another. At some point, perhaps after the Apollo 1 disaster, it was decided to fake the space program outright and use rockets which only needed to fly into the air until they disappeared from sight."

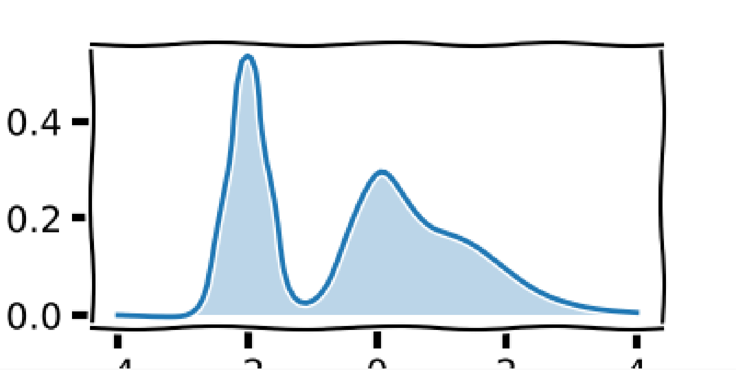
[2] `Globe Conspiracy`

Finally an explanation that makes sense



Uncertainty

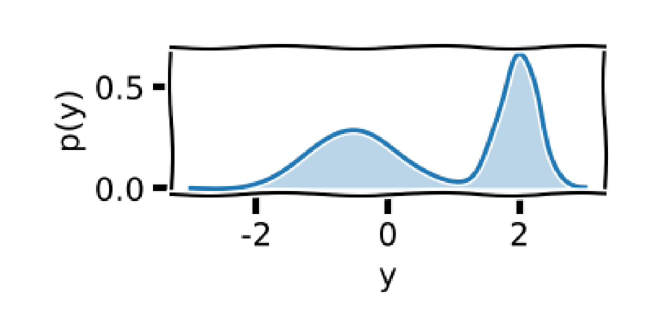
Uncertainty



- Uncertainty is a "realisation" of an assumption
- Probabilities are a quantification of uncertainty

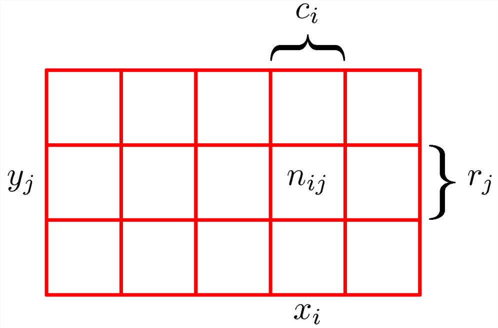
Probabilities

Probability Theory



- Probability theory is a framework for manipulating uncertainty
- Random variable, is a stochastic variable that follows a distribution
- Random does not mean max entropy

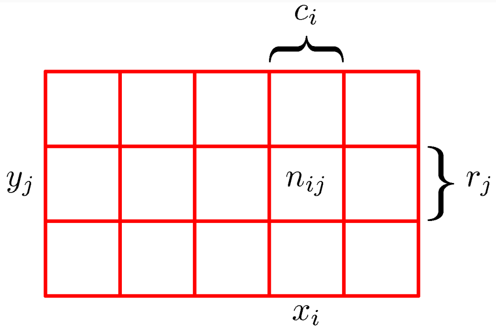
Probability Theory [1] ch 1.2



$$\{X = x_i, Y = y_j\} = n_{ij} \quad \{X = x_i\} = c_i \quad \{Y = y_j\} = r_j$$

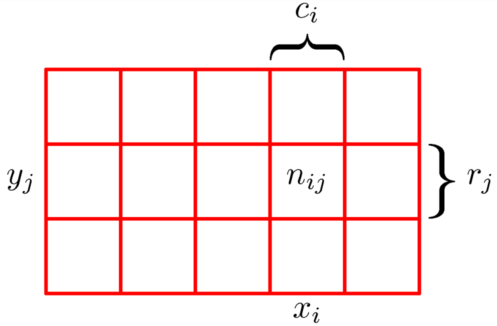
$$X = \{x_i\}_{i=1}^{M} \quad Y = \{y_j\}_{j=1}^{N}$$

Probability Theory [1] ch 1.2



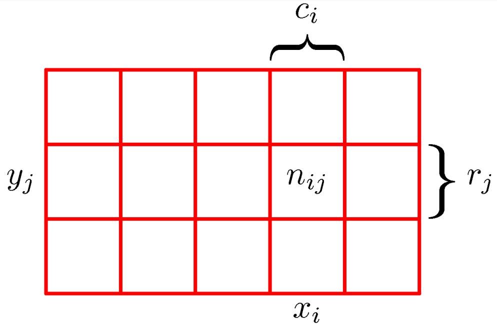
Joint Probability

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{\sum_{kl} n_{kl}} = \frac{n_{ij}}{N}$$



Marginal Probability

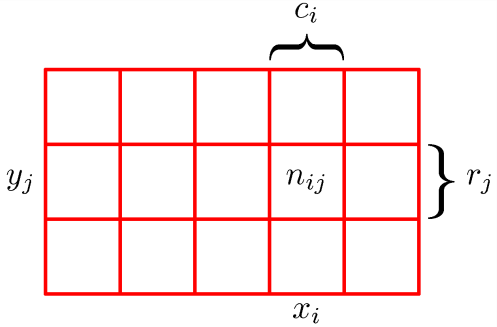
$$p(X = x_i) = \frac{\sum_j n_{ij}}{N} = \frac{c_i}{N}$$



Sum rule

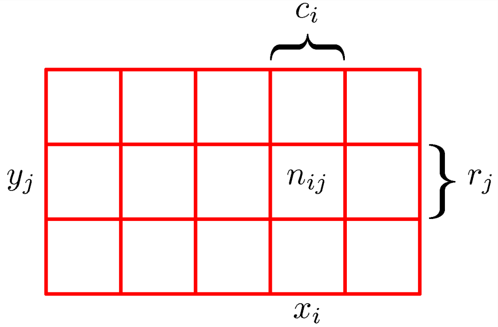
$$p(X = x_i) = \frac{\sum_j n_{ij}}{N}$$

Probability Theory [1] ch 1.2



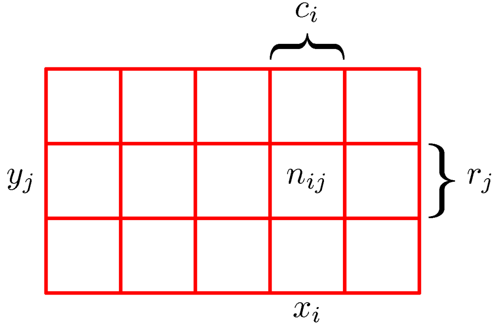
Sum rule

$$p(X = x_i) = \frac{\sum_j n_{ij}}{N} = \sum_j \frac{n_{ij}}{N}$$



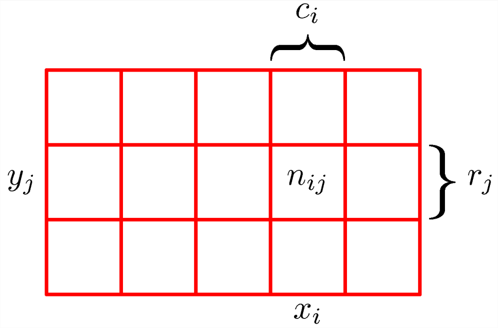
Sum rule

$$p(X = x_i) = \frac{\sum_j n_{ij}}{N} = \sum_j \frac{n_{ij}}{N} = \sum_j p(X = x_i, Y = y_j)$$



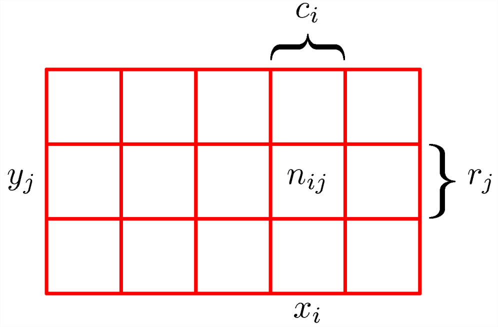
Conditional

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$



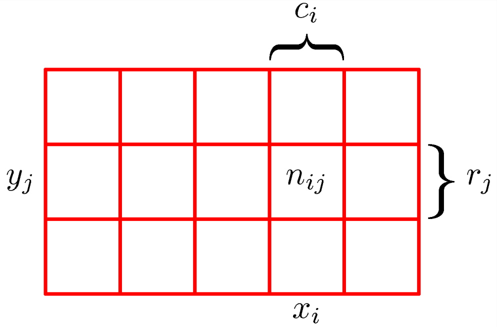
Product rule

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$



Product rule

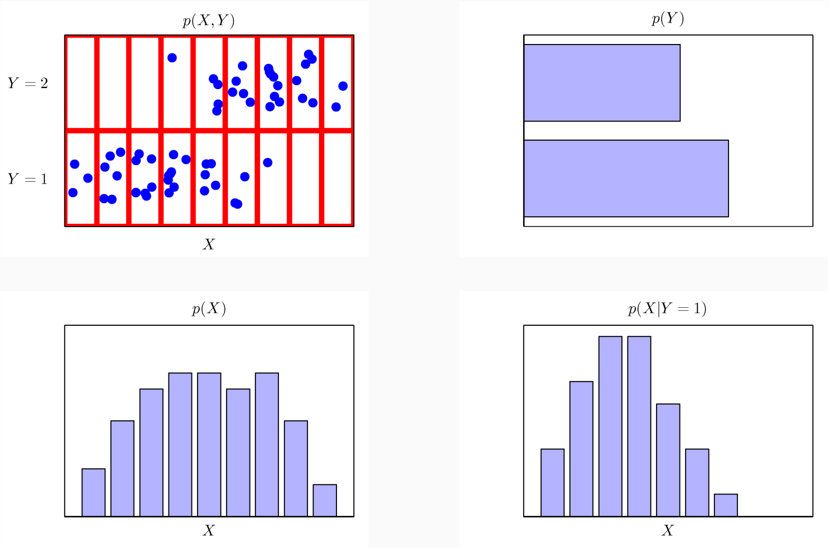
$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N}$$



Product rule

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N} = p(Y = y_j | X = x_i) p(X = x_i)$$

Probability Theory [1] ch 1.2



Probability Theory

Notation

- The probability distribution over the random variable X

$$p(X) = p(X = x_i)$$

Notation

- The probability distribution over the random variable X

$$p(X) = p(X = x_i)$$

- The probability distribution over X evaluated at x_i

$$p(x_i)$$

The Rules of Probability

Sum Rule

$$p(X) = \sum_{Y} p(X, Y)$$

Product Rule

$$p(X, Y) = p(Y|X)p(X)$$

Baye's Rule

$$p(X, Y) = p(Y|X)p(X)$$

Baye's Rule

$$p(X, Y) = p(Y|X)p(X)$$
$$p(X, Y) = p(X|Y)p(Y)$$

Baye's Rule

$$
\begin{aligned}
p(X, Y) &= p(Y|X)p(X) \\
p(X, Y) &= p(X|Y)p(Y) \\
p(X|Y)p(Y) &= p(Y|X)p(X)
\end{aligned}
$$

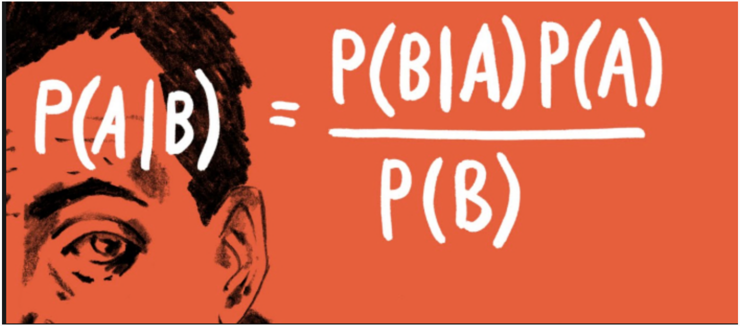
Baye's Rule

$$
\begin{aligned}
p(X, Y) &= p(Y|X)p(X) \\
p(X, Y) &= p(X|Y)p(Y) \\
p(X|Y)p(Y) &= p(Y|X)p(X) \\
p(X|Y) &= \frac{p(Y|X)p(X)}{p(Y)}
\end{aligned}
$$

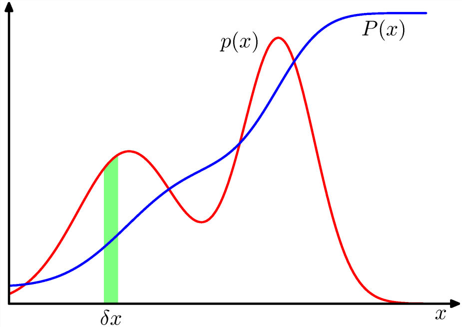
Baye's Rule

$$\begin{aligned}
p(X,Y) &= p(Y|X)p(X) \\
p(X,Y) &= p(X|Y)p(Y) \\
p(X|Y)p(Y) &= p(Y|X)p(X) \\
p(X|Y) &= \frac{p(Y|X)p(X)}{p(Y)} \\
&= \frac{p(Y|X)p(X)}{\sum_X p(Y|X)p(X)}
\end{aligned}$$

Baye's Rule



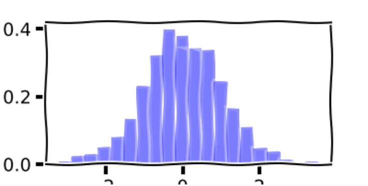
Probability Densities [1] ch 1.2.1

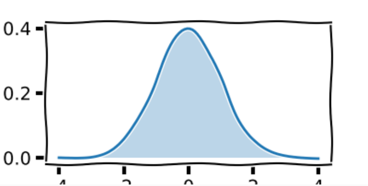


$$\lim_{\delta x \to 0} p(x \in (x, x + \delta x)) = \lim_{\delta x \to 0} \int_{x}^{x+\delta x} p(x)\mathrm{d}x = p(x) \cdot \delta x$$

$$p(x) \geq 0, \quad \int_{-\infty}^{\infty} p(x)\mathrm{d}x = 1$$

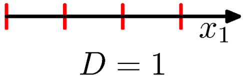
Discrete vs. Continous

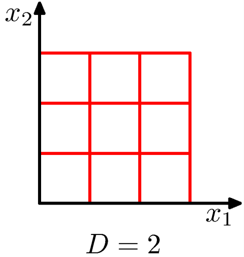


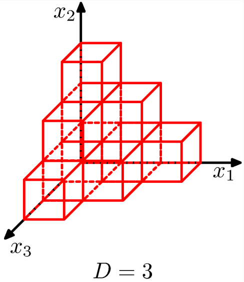


$$\sum \qquad\qquad \int$$

Curse of Dimensionality [1] ch 1.4



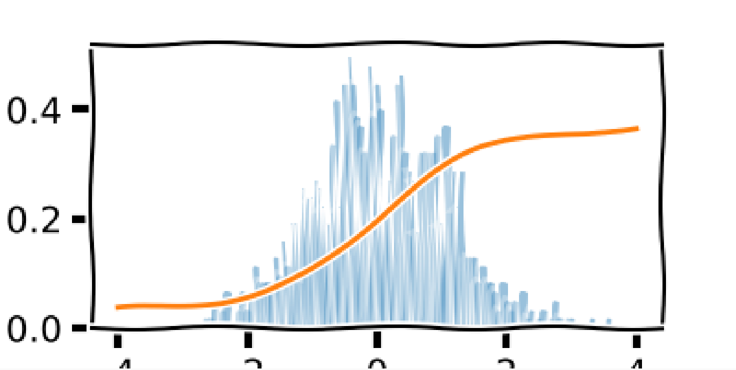




Curse of Dimensionality

```
for i in range(0,N1):
    for j in range(0,N2):
        for k in range(0,N3):
            ...
            ...
            ...
```

Expectations



$$\mathbb{E}[f] = \sum_x p(x) f(x)$$

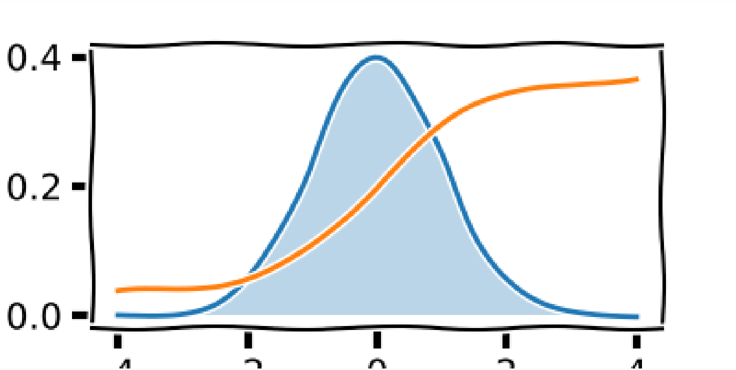
Expectations

```
e = 0.0
for x in range(Xmin,Xmax):
    e += f(x)*p(x)

return e
```

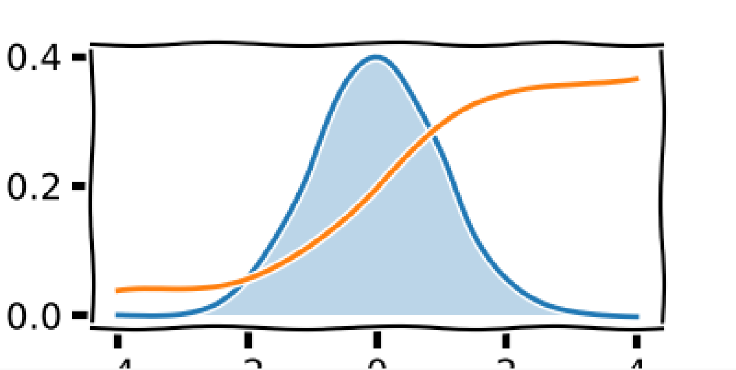
- simple to write
- can be infeasible to compute when domain is high dimensional

Expectations



$$\mathbb{E}[f] = \int p(x) f(x) \mathrm{d}x$$

Expectations



$$\mathbb{E}[f] = \int p(x) f(x) \mathrm{d}x \approx \frac{1}{N} \sum_{i}^{N} f(x_i)$$

$$x_i \sim p(x)$$

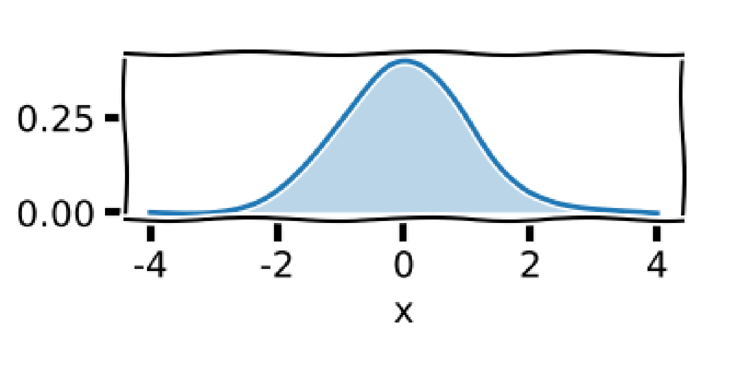
Expectations

```
e = 0.0
for i in range(0,N):
    x = 0.0 + 1.0*np.random.randn(1)
    e += f(x)

return e/N
```

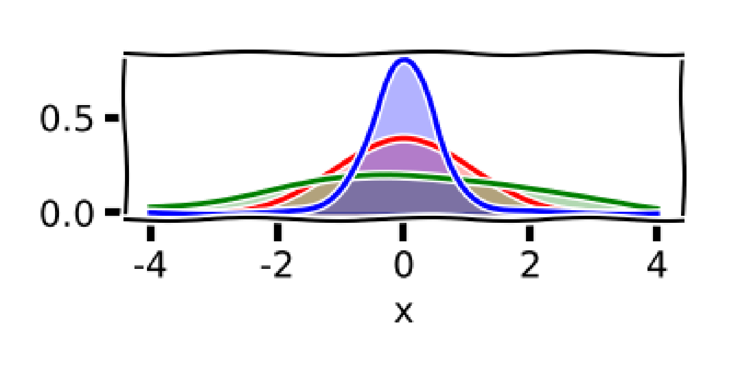
- drawing samples might be tricky
- can be infeasible when entropy of $p(x)$ is large, i.e. many samples

Expectations



$$\mathbb{E}[x] = \int x p(x) \mathrm{d}x$$

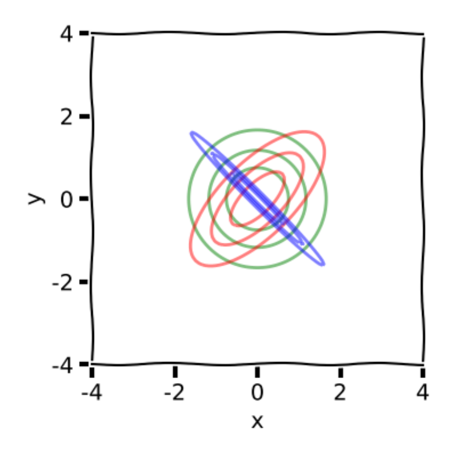
Variance



$$\mathrm{var}[f] = \mathbb{E}\left[\left(f(x) - \mathbb{E}[f(x)]\right)^2\right]$$

$$\mathrm{var}[x] = \mathbb{E}\left[\left(x - \mathbb{E}[x]\right)^2\right]$$

Covariance



$$\mathrm{cov}[x, y] = \mathbb{E}\left[(x - \mathbb{E}[x])(y - \mathbb{E}[y])\right]$$

Summary

Distributions

Joint $p(x, y)$

Rules

Summary

Distributions

Joint	$p(x, y)$
Marginal	$p(x)$, $p(y)$

Rules

Summary

Distributions

Joint	$p(x, y)$
Marginal	$p(x)$, $p(y)$
Conditional	$p(y\|x)$, $p(x\|y)$

Rules

Summary

Distributions

Joint	$p(x, y)$
Marginal	$p(x)$, $p(y)$
Conditional	$p(y\|x)$, $p(x\|y)$

Rules

Sum $p(x) = \sum_y p(y, x)$

Summary

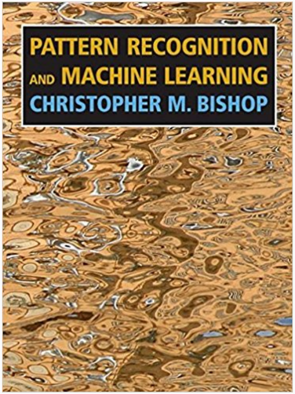
Distributions

Joint	$p(x, y)$
Marginal	$p(x)$, $p(y)$
Conditional	$p(y\|x)$, $p(x\|y)$

Rules

Sum	$p(x) = \sum_y p(y, x)$
Product	$p(x, y) = p(y\|x)p(x)$

Book [1]



Ch 1.0, 1.2.1-1.2.2

Bayesian Probabilities

Interpretations of probabilities [1] ch 1.2.3

Frequentist

- a probability is a frequency of a repeatable random event

Bayesian

Interpretations of probabilities [1] ch 1.2.3

Frequentist

- a probability is a frequency of a repeatable random event
- have no beliefs

Bayesian

Interpretations of probabilities [1] ch 1.2.3

Frequentist

- a probability is a frequency of a repeatable random event
- have no beliefs

Bayesian

- a probability is a quantification of a belief

Interpretations of probabilities [1] ch 1.2.3

Frequentist

- a probability is a frequency of a repeatable random event
- have no beliefs

Bayesian

- a probability is a quantification of a belief
- probabilities are usually attributed to random/stochastic variables

Interpretations of probabilities [1] ch 1.2.3

Frequentist

- a probability is a frequency of a repeatable random event
- have no beliefs

Bayesian

- a probability is a quantification of a belief
- probabilities are usually attributed to random/stochastic variables
- can be seen as an extension to Boolean logic for uncertain events

Interpretations of probabilities [1] ch 1.2.3

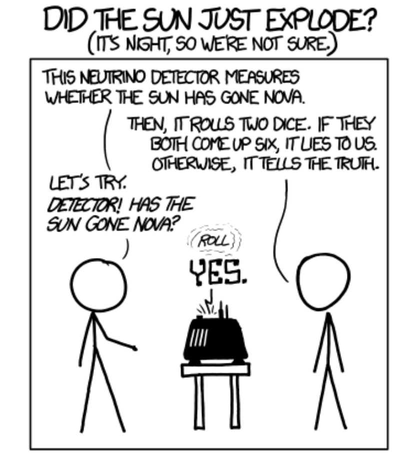
Frequentist

- a probability is a frequency of a repeatable random event
- have no beliefs

Bayesian

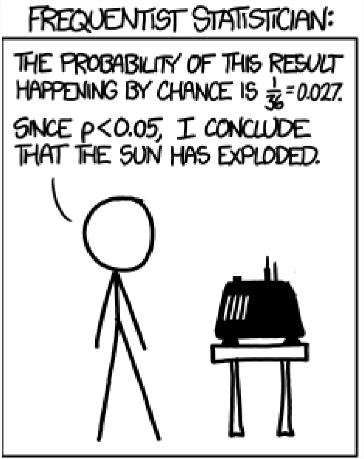
- a probability is a quantification of a belief
- probabilities are usually attributed to random/stochastic variables
- can be seen as an extension to Boolean logic for uncertain events
- requires beliefs

XKCD[3]


DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY.
DETECTOR! HAS THE SUN GONE NOVA?
ROLL
YES.

[3] XKCD

XKCD[3]

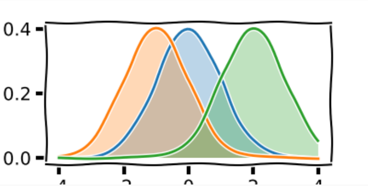

FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS 1/36 = 0.027.
SINCE p<0.05, I CONCLUDE THAT THE SUN HAS EXPLODED.

[3] XKCD

XKCD[3]

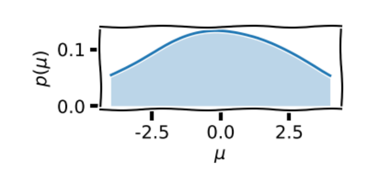
[3] XKCD

Interpretations



$$p(y|\mu) = \mathcal{N}(\mu, 1.0)$$

Likelihood



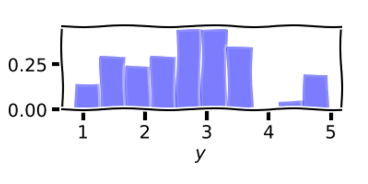
$$p(\mu)$$

Prior

Bayes Rule

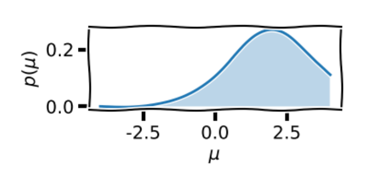
$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

Bayes Rule



y

Data



$p(\mu|y)$

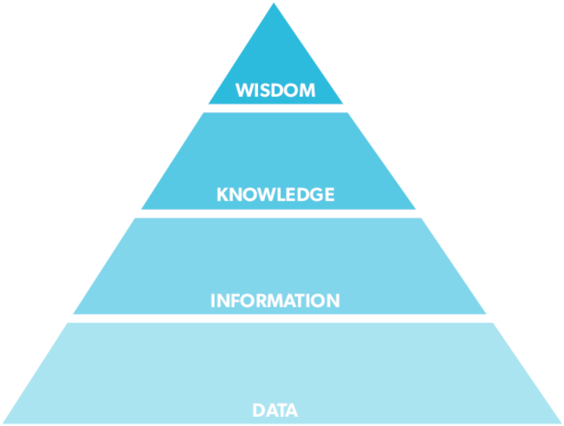
Posterior

Scientific Modelling

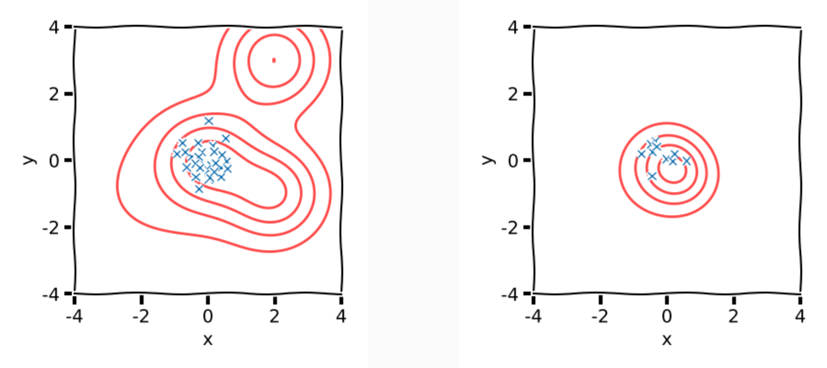
"Scientific modelling is a scientific activity, the aim of which is to make a particular part or feature of the world easier to understand, define, quantify, visualize, or simulate by referencing it to existing and usually commonly accepted knowledge." [4]

[4] Wikipedia

The importance of data



Assumptions



Interesting Reads

- Davey, S., Gordon, N., Holland, I., Rutten, M., & Williams, J., Bayesian methods in the search for mh370 (2016), : Springer Singapore. [2]
- Stone, L. D., Keller, C. M., Kratzke, T. M., & Strumpfer, J. P., Search for the wreckage of air france flight af 447, Statistical Science, 29(1), 69–80 (2014). [3]

Summary

Summary

- Learning can only be made by making assumptions

Summary

- Learning can only be made by making assumptions
- You don't learn something that is true, you accept it

Summary

- Learning can only be made by making assumptions
- You don't learn something that is true, you accept it
- Uncertainty is a "realisation" of an assumption

Summary

- Learning can only be made by making assumptions
- You don't learn something that is true, you accept it
- Uncertainty is a "realisation" of an assumption
- Probabilities are a quantification of uncertainty

Summary

- Learning can only be made by making assumptions
- You don't learn something that is true, you accept it
- Uncertainty is a "realisation" of an assumption
- Probabilities are a quantification of uncertainty
- Probabilities does not need to be frequencies of events

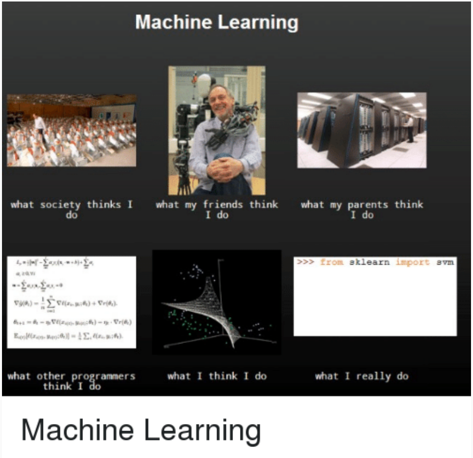
Summary

- Learning can only be made by making assumptions
- You don't learn something that is true, you accept it
- Uncertainty is a "realisation" of an assumption
- Probabilities are a quantification of uncertainty
- Probabilities does not need to be frequencies of events
- More assumptions means less data

eof

The Reddit


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS


Machine Learning
what society thinks I do
what my friends think I do
what my parents think I do
what other programmers think I do
what I think I do
>>> from sklearn import svm
what I really do
Machine Learning

References

Christopher M. Bishop.
***Pattern Recognition and Machine Learning (Information Science and Statistics)*.**
Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

Sam Davey, Neil Gordon, Ian Holland, Mark Rutten, and Jason Williams.
***Bayesian Methods in the Search for MH370*.**
SpringerBriefs in Electrical and Computer Engineering. Springer Singapore, 2016.

Lawrence D. Stone, Colleen M. Keller, Thomas M. Kratzke, and Johan P. Strumpfer.
Search for the wreckage of air france flight af 447.
Statistical Science, 29(1):69–80, 2014.