Machine Learning

Basic Probabilities

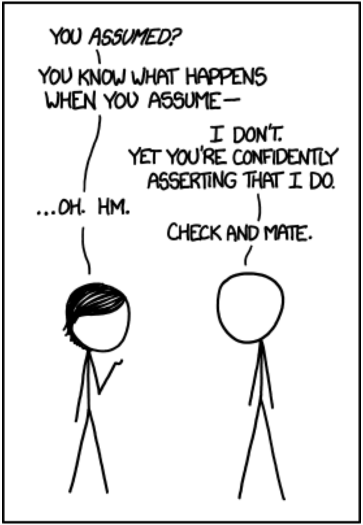
Carl Henrik Ek - carlhenrik.ek@bristol.ac.uk

October 1, 2019

http://www.carlhenrik.com

Introduction

Assumptions



Assumptions

- Observations cannot be argued with
- Interpretations of observations are relative to assumptions
- Good assumptions structures the world in a useful manner
- Wrong assumptions can be very scary

Learning?

Truth?

Belief?

Laplace Demon [1]

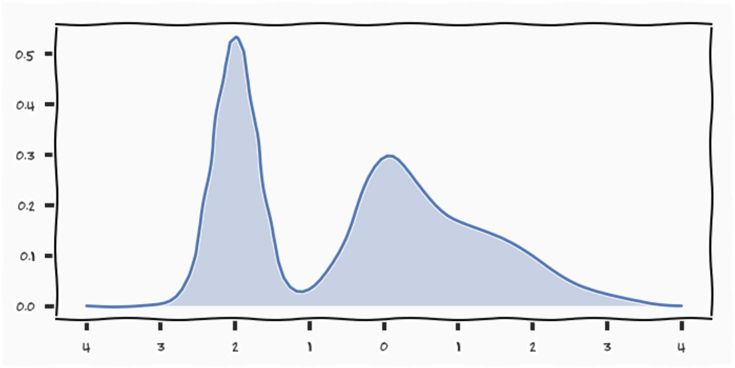
Laplace's Demon [1]

An intelligence which at a given instant knew all the forces acting in nature and the position of every object in the universe - if endowed with a brain sufficiently vast to make all necessary calculations - could describe with a single formula the motions of the largest astronomical bodies and those of the smallest atoms. To such an intelligence, nothing would be uncertain; the future, like the past, would be an open book.

Laplace Demon [1]

All these efforts in the search for truth tend to lead the mind continously towards the intelligence we have just mentioned, although it will always remain infinetly distant from this intelligence.

Uncertainty



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

Variables

Deterministic Variable

Code

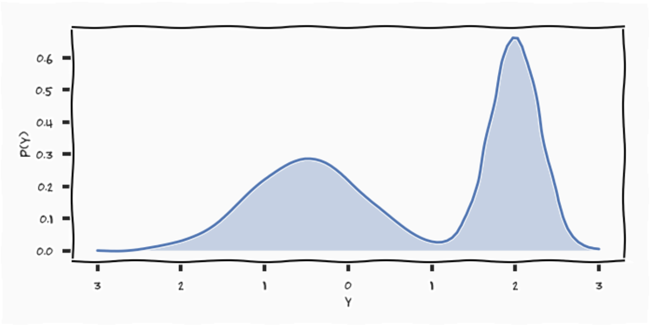
```
int x = 3
float y = 3.14
```

Stochastic Variable

$$x \sim p(x)$$
$$y \sim \mathcal{N}(0, I)$$

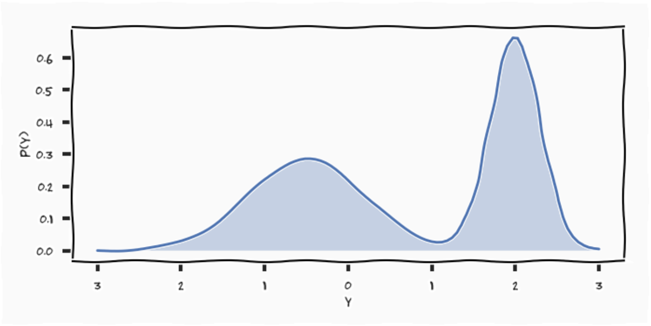
Probabilities

Probability Theory

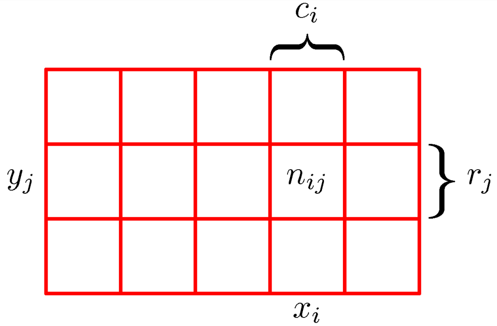


- Probability theory is a framework for manipulating uncertainty
- Random variable, is a stochastic variable that follows a distribution
- Random does not mean max entropy

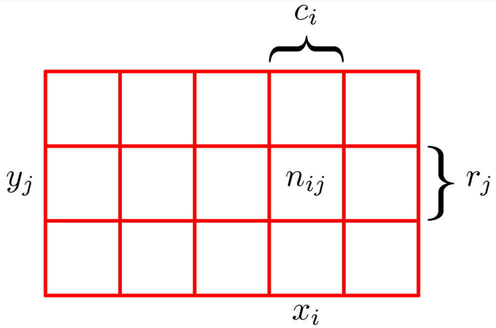
Probability Theory



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

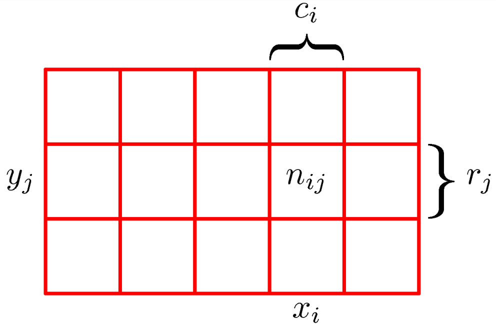


$$\{X = x_i, Y = y_j\} = n_{ij}$$



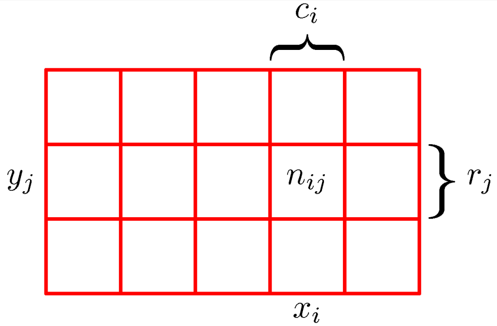
Joint Probability

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



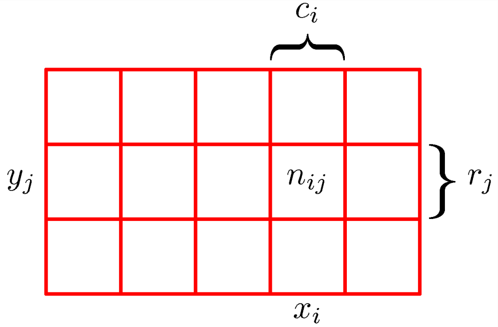
Marginal Probability

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



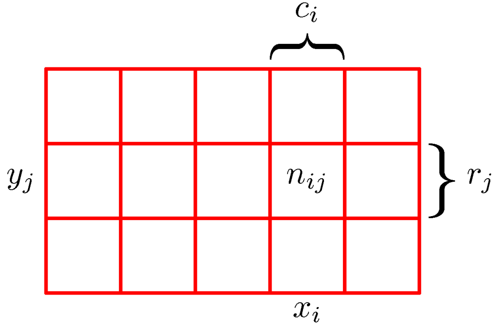
Sum rule

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



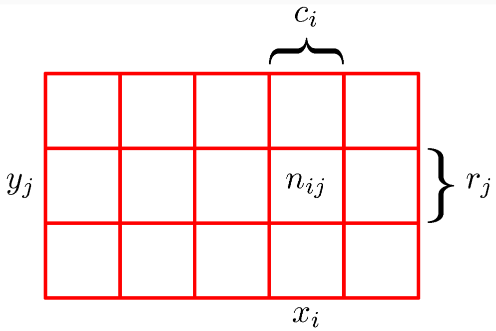
Sum rule

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



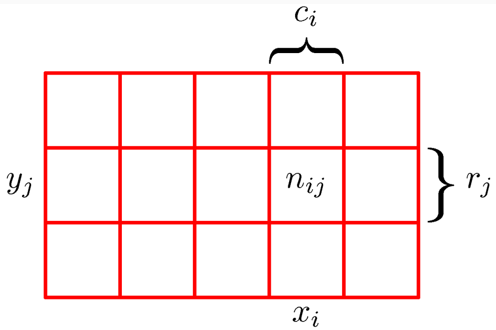
Sum rule

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



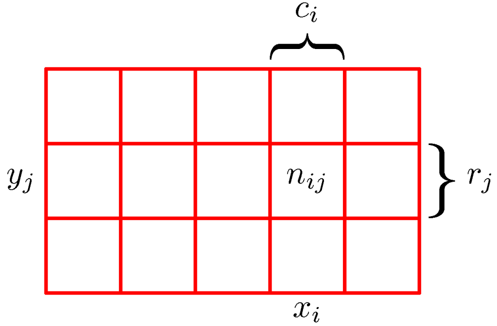
Conditional

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



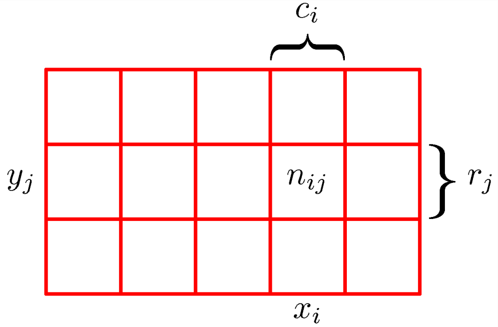
Product rule

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



Product rule

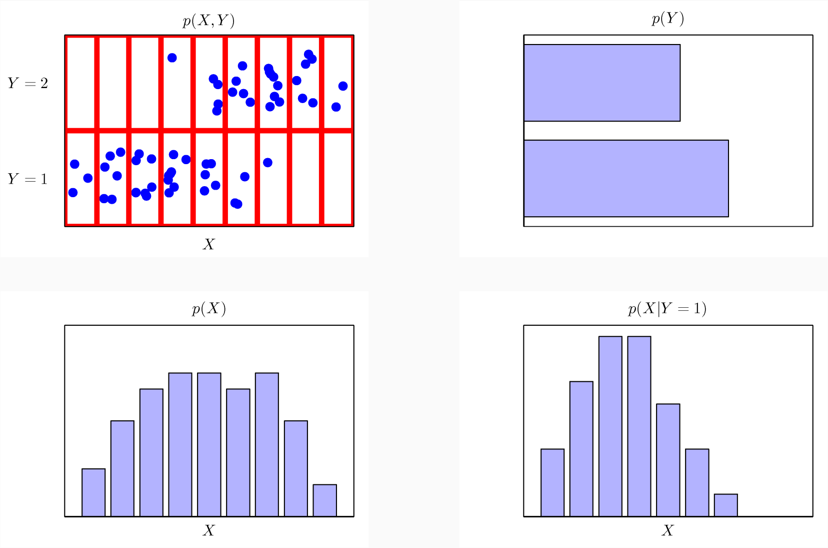
$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N}$$



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 [2] 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

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

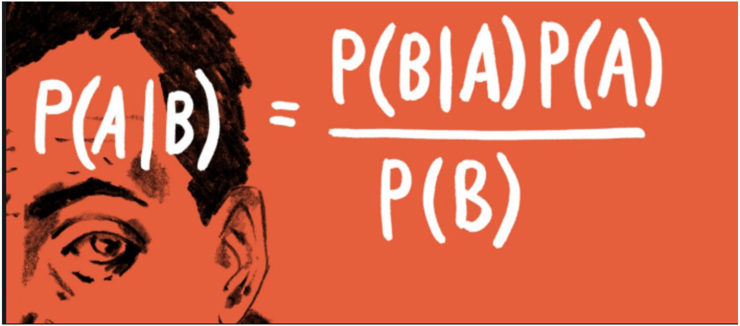
Baye's Rule

$$p(X, Y) = p(Y|X)p(X)$$
$$p(X, Y) = p(X|Y)p(Y)$$
$$p(X|Y)p(Y) = p(Y|X)p(X)$$
$$p(X|Y) = \frac{p(Y|X)p(X)}{p(Y)}$$

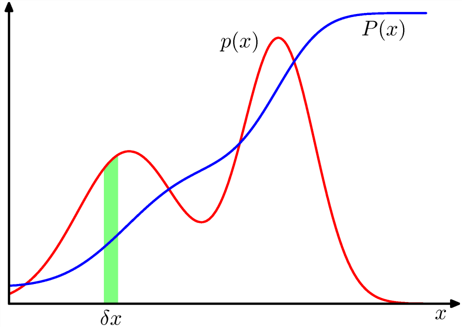
Baye's Rule

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

Baye's Rule



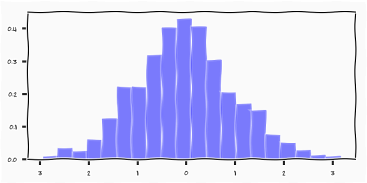
Probability Densities [2] ch 1.2.1

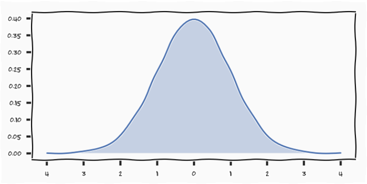


$$\lim_{\delta x \to 0} p(x \in (x, x+\delta x)) = \lim_{\delta x \to 0} \int_x^{x+\delta x} p(x) \mathrm{d}x = p(x) \cdot \delta x$$

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

Discrete vs. Continous

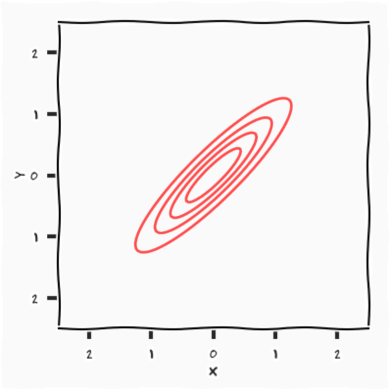




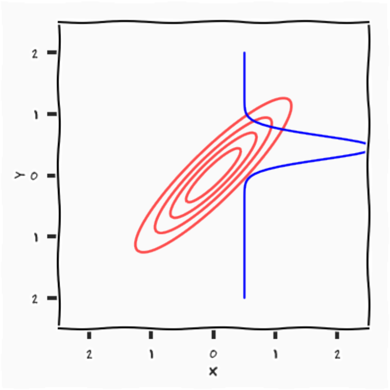
$$\sum$$

$$\int$$

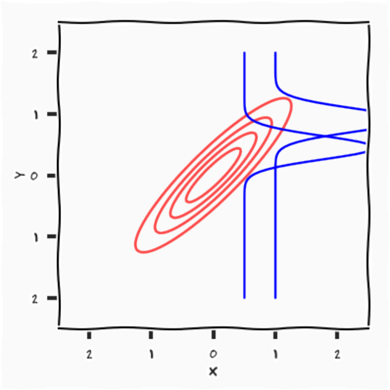
Continous: Conditional



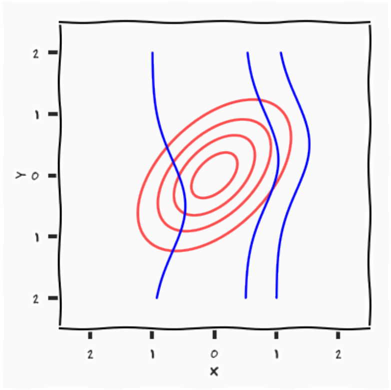
Continous: Conditional



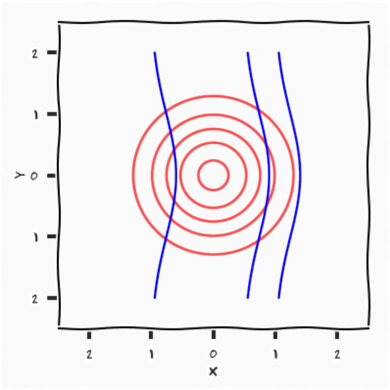
Continous: Conditional



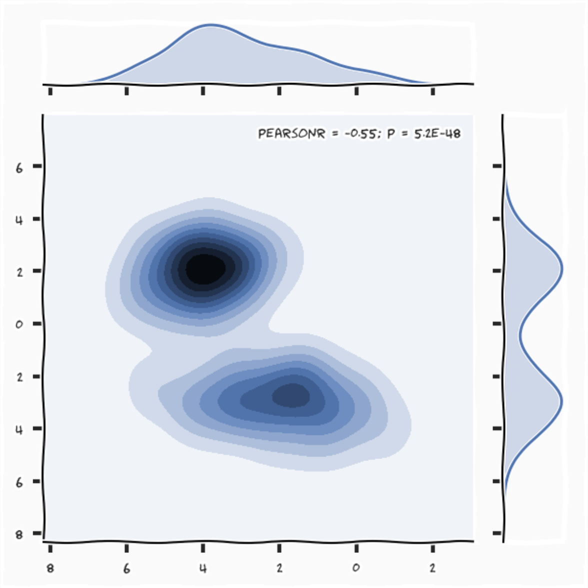
Continous: Conditional



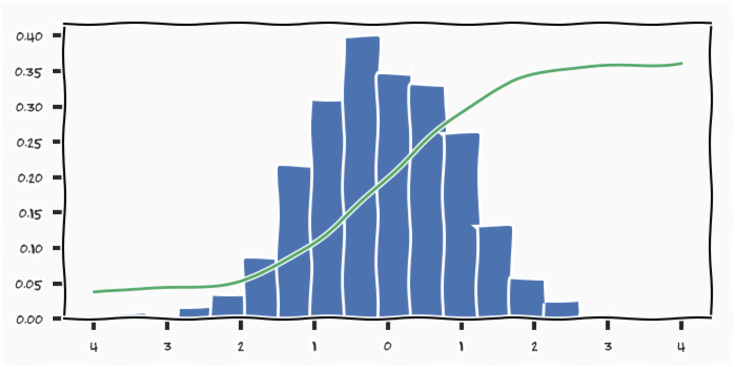
Continous: Conditional



Continous: Marginal



Expectations



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

Expectations

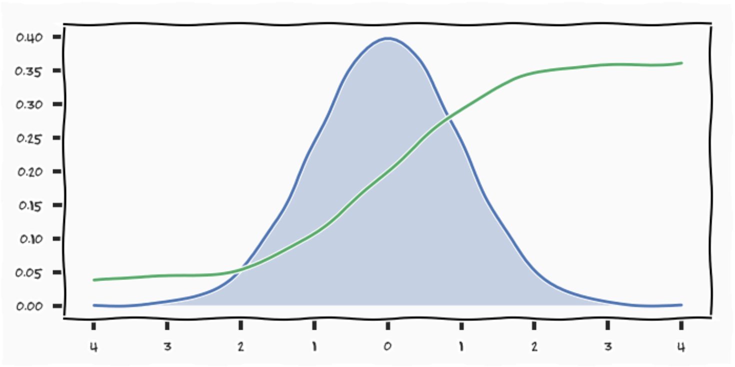
Code

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

return e
```

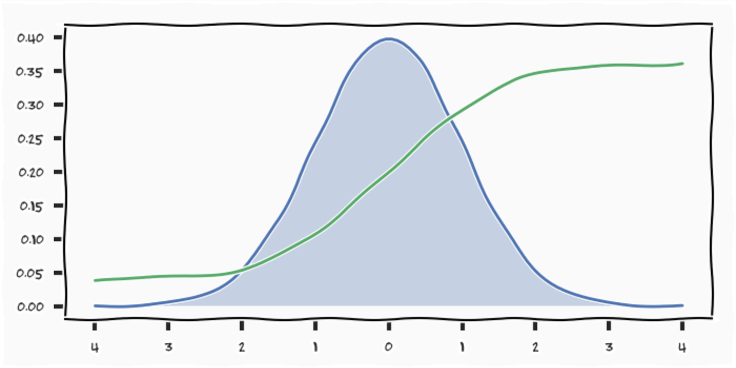
- simple to write
- can be infeasible to compute when domain is high dimensional

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

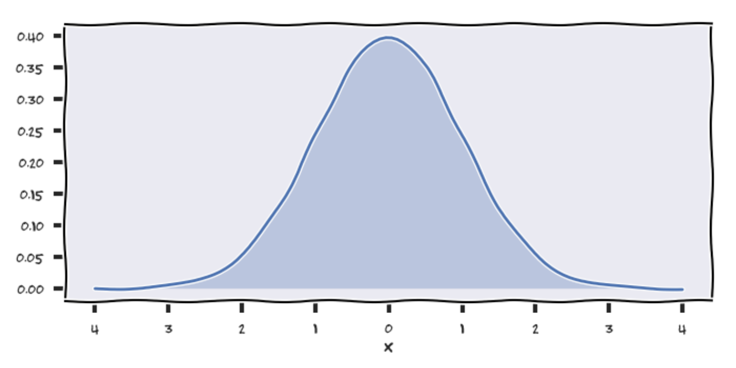
Code

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

return e/N
```

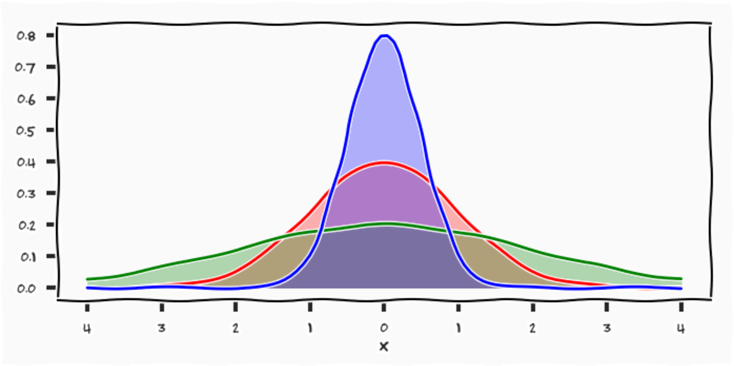
- drawing samples might be tricky
- can be infeasible when entropy of $p(x)$ is large, i.e. many samples

Expectations



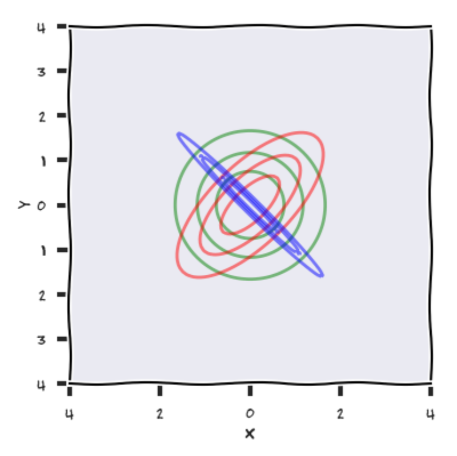
$$\mathbb{E}[x] = \int x p(x) \mathrm{d}x$$

Variance



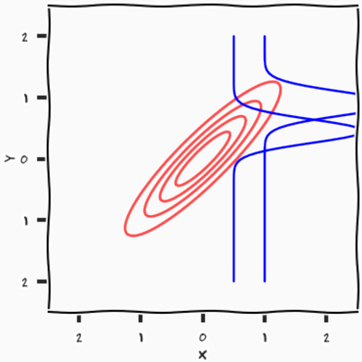
$$\mathrm{var}[x] = \mathbb{E}\left[(x - \mathbb{E}[x])^2\right]$$

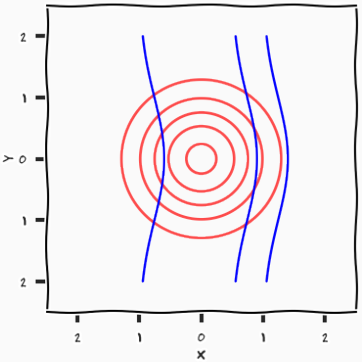
Covariance



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

Covariance



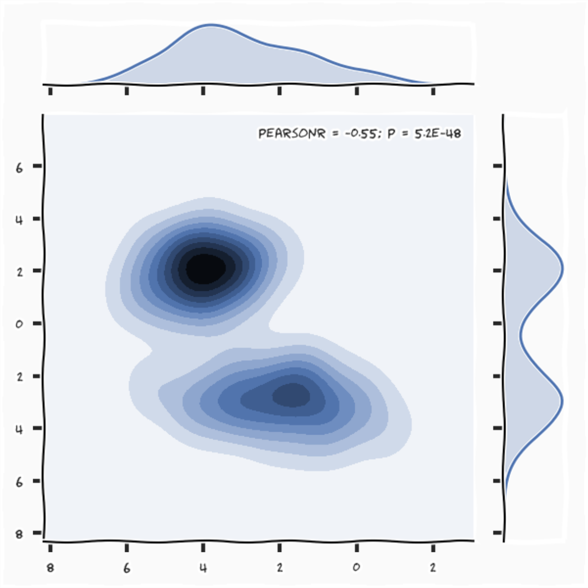


Marginalisation

$$p(y) = \int p(x, y)\mathrm{d}x = \int p(y|x)p(x)\mathrm{d}x$$

- Marginalisation accounts for all your belief in a variable
- Importantly it does not "remove" the effect of the variable
- Marginalisation is an expectation over a conditional distribution

Continous: Marginal



What does Expectations Mean?

What does Expectations Mean?

1000 GBP

$p(x = \text{dad}) = 0.2$

5000 GBP

$p(x = \text{mum}) = 0.5$

500 GBP

$p(x = \text{sister}) = 0.3$

What does Expectations Mean?

1000 GBP

$p(x = \text{dad}) = 0.2$

5000 GBP

$p(x = \text{mum}) = 0.5$

500 GBP

$p(x = \text{sister}) = 0.3$

$$1000 \cdot 0.2 + 5000 \cdot 0.5 + 500 \cdot 0.3 = 2850$$

What does Expectations Mean?

Next time you want to give your friends a compliment, tell them that you have completely marginalised them from your life

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	$p(x, y) = p(y\|x)p(x)$

Summary

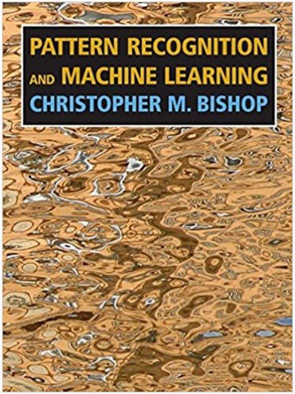
Probability Mass/Density Functions

it is important to note, these are just like any function, and you can deal with them in the same way, the difference is just that they have the additional constraints

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

When we call them $p(y|x)$ is just a semantic added to the function

Book [2]



Ch 1.0, 1.2.1-1.2.2

Laplace [1]

"On voit, par cet Essai, que la théorie des probabilités n'est, au fond, que le bon sens réduit au calcul; elle fait apprécier avec exactitude ce que les esprits justes sentent par une sorte d'instinct, sans qu'ils puissent souvent s'en rendre compte."

– Simon Laplace

Laplace [1]

"One sees, from this Essay, that the theory of probabilities is basically just common sense reduced to calculus; it makes one appreciate with exactness that which accurate minds feel with a sort of instinct, often without being able to account for it."

– Simon Laplace

Bayesian Probabilities

Interpretations of probabilities [2] ch 1.2.3

Frequentist

- a probability is a frequency of a repeatable random event

Bayesian

[1] https://en.wikipedia.org/wiki/Cox%27s_theorem

Interpretations of probabilities [2] ch 1.2.3

Frequentist

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

Bayesian

[1] https://en.wikipedia.org/wiki/Cox%27s_theorem

Interpretations of probabilities [2] 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

[1] https://en.wikipedia.org/wiki/Cox%27s_theorem

Interpretations of probabilities [2] 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

[1] https://en.wikipedia.org/wiki/Cox%27s_theorem

Interpretations of probabilities [2] 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[1]

[1] `https://en.wikipedia.org/wiki/Cox%27s_theorem`

Interpretations of probabilities [2] ch 1.2.3

Frequentist

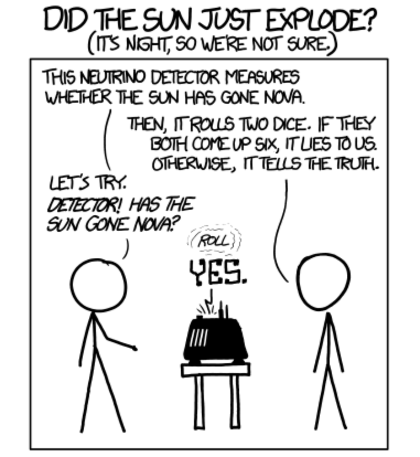
- a probability is a frequency of a repeatable random event
- have no beliefs

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[1]
- requires beliefs

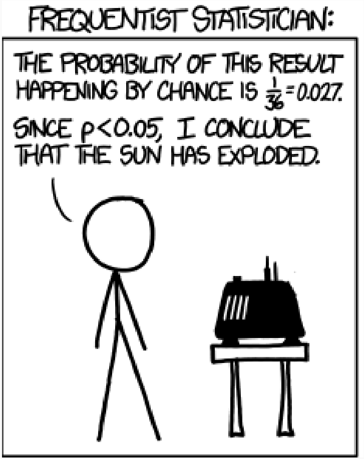
[1] https://en.wikipedia.org/wiki/Cox%27s_theorem

XKCD[2]


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.

[2] XKCD

XKCD[2]

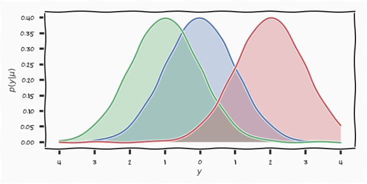

FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS $\frac{1}{36}$ = 0.027.
SINCE $p < 0.05$, I CONCLUDE THAT THE SUN HAS EXPLODED.

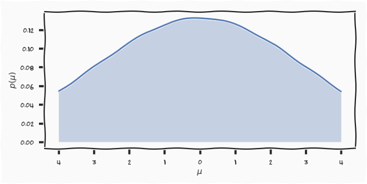
[2] XKCD

XKCD[2]

[2] XKCD

Interpretations





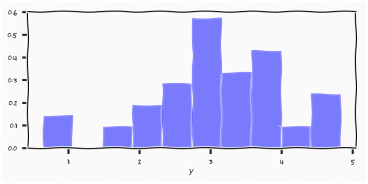
$$p(y|\mu) = \mathcal{N}(\mu, 1.0)$$

Likelihood

$$p(\mu)$$

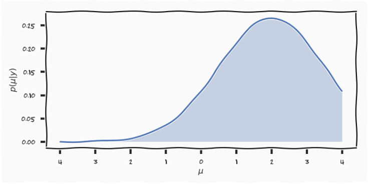
Prior

$$p(\mu|y) = \frac{p(y|\mu)p(\mu)}{\int p(y|\mu)p(\mu)}$$



y

Data



$p(\mu|y)$

Posterior

Are beliefs objective?

NO of course not and therefore we all learn different things from the same data

Are beliefs objective?

NO of course not and therefore we all learn different things from the same data

YES if two exact copies of the same "person" have different beliefs they cannot be the same person, therefore beliefs are objective, its only different amount of data/knowledge that generates differences

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." [3]

---

[3] Wikipedia

Interesting Reads

- Davey, S., Gordon, N., Holland, I., Rutten, M., & Williams, J., Bayesian methods in the search for mh370 (2016), : Springer Singapore. [3]
- 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). [4]

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 (if I'm right)

eof

References

Pierre Simon Laplace.
A philosophical essay on probabilities, 1814.

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.

Appendix

Decisions

[4]

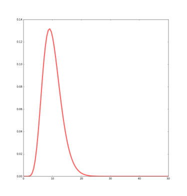
[4] Reservoir Dogs Tipping Scene YouTube

Tip

$$p(y)$$

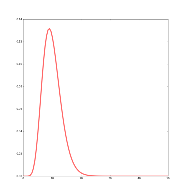
- what do I believe about tip before I see data?
- what is a sensible tip?

Tipping



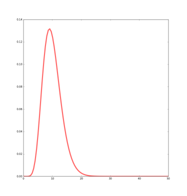
- I believe that 10£ is a sensible tip
- You cannot tip negative
- There is potentially an upper bound

Tipping



- I believe that 10£ is a sensible tip
- You cannot tip negative
- There is potentially an upper bound
- This is not a model, its just a belief in a variable

Tipping



- I believe that 10£ is a sensible tip
- You cannot tip negative
- There is potentially an upper bound
- This is not a model, its just a belief in a variable
- *a model relates new phenomenon to knowledge*

Tipping

- it is quite hard to say something about tip without any other knowledge
- Assumption the value of tip is related to the quality of the food

Likelihoods

$$p(y|x)$$

- how likely do I think the observed data y is to come from this specific x.

Tipping if I know the quality of the food what do I believe the tip should be

What is the tip that I should expect to get?

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

- What should I expect to get in tip
- I have an idea of the general distribution of quality of food
- *Understanding is when we can relate knowledge to new phenomenon*

Tipping

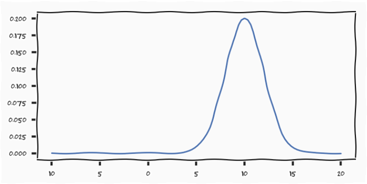
$$p(x|c)$$

Hierarchical distribution

- Its quite hard to think of a prior over quality of food
- Can we parametrise the quality?

Tipping

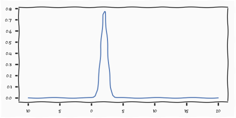
$$p(x|c) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_c)(x-\mu_c)}{2\sigma^2}}$$



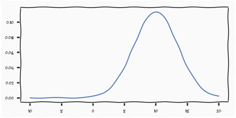
Hierarchical distribution

- Its quite hard to think of a prior over quality of food
- Can we parametrise the quality?
- Lets assume that if we know the cusine we have an idea

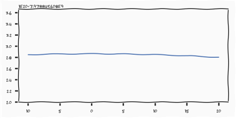
Tipping



Swedish



Italian



Uzbeki

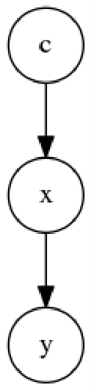
Hierarchical distribution

- Its quite hard to think of a prior over quality of food
- Can we parametrise the quality?
- Lets assume that if we know the cusine we have an idea
- *Relating to knowledge!*

Tipping model

$$p(y, x, c) = p(y|x)p(x|c)p(c)$$

- Graphical Model shows dependency structure
- Shows "minimal" factorisation of joint distribution (model)



Tipping model

$p(y|x)$ Lets assume that tip is linearly related to quality of food

- $y = Wx + m$

$p(x|c)$ We saw them before

$p(c)$ What do I believe the proportions of resturants to be?

- $p(c = \text{"}swedish\text{"}) = 0.04$
- $p(c = \text{"}italian\text{"}) = 0.90$
- $p(c = \text{"}uzbeki\text{"}) = 0.06$

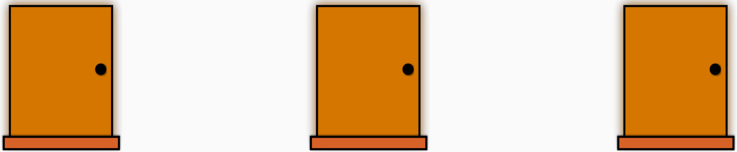
Decisions

Knowing the tip

- Which cusine did they eat if?
  - $p(c|y)$
- What was the quality of the food?
  - $p(x|y)$

Monty Hall

Monty Hall



Monty Hall

- The doors are exchangable
- You choose door 1 and Monty chooses door 3
- *should you switch door?*
- Lets call C the door with the car behind and D the door Monty chooses

Monty Hall

- What is the probability that the car is behind any specific door?

Monty Hall

- What is the probability that the car is behind any specific door?
- $p(C = \{1, 2, 3\}) = \frac{1}{3}$

Monty Hall

- What is the probability that the car is behind any specific door?
- $p(C = \{1, 2, 3\}) = \frac{1}{3}$
- What is the probability that Monty opens door 3 if the car is behind door 1?

Monty Hall

- What is the probability that the car is behind any specific door?
- $p(C = \{1, 2, 3\}) = \frac{1}{3}$
- What is the probability that Monty opens door 3 if the car is behind door 1?
- $p(D = 3 | C = 1) = \frac{1}{2}$

Monty Hall

- What is the probability that the car is behind any specific door?
- $p(C = \{1, 2, 3\}) = \frac{1}{3}$
- What is the probability that Monty opens door 3 if the car is behind door 1?
- $p(D = 3 | C = 1) = \frac{1}{2}$
- What is the probability that Monty opens door 3 if the car is behind door 2?

Monty Hall

- What is the probability that the car is behind any specific door?
- $p(C = \{1, 2, 3\}) = \frac{1}{3}$
- What is the probability that Monty opens door 3 if the car is behind door 1?
- $p(D = 3|C = 1) = \frac{1}{2}$
- What is the probability that Monty opens door 3 if the car is behind door 2?
- $p(D = 3|C = 2) = 1$

Monty Hall

- What is the probability that the car is behind any specific door?
- $p(C = \{1, 2, 3\}) = \frac{1}{3}$
- What is the probability that Monty opens door 3 if the car is behind door 1?
- $p(D = 3|C = 1) = \frac{1}{2}$
- What is the probability that Monty opens door 3 if the car is behind door 2?
- $p(D = 3|C = 2) = 1$
- Marginal of $D = 3$

Monty Hall

- What is the probability that the car is behind any specific door?
- $p(C = \{1, 2, 3\}) = \frac{1}{3}$
- What is the probability that Monty opens door 3 if the car is behind door 1?
- $p(D = 3|C = 1) = \frac{1}{2}$
- What is the probability that Monty opens door 3 if the car is behind door 2?
- $p(D = 3|C = 2) = 1$
- Marginal of $D = 3$
- $p(D = 3) = p(D = 3|C = 1)p(C = 1) + p(D = 3|C = 2)p(C = 2) + p(D = 3|C = 3)p(C = 3) = \frac{1}{2}\cdot\frac{1}{3} + 0\cdot\frac{1}{3} + 1\cdot\frac{1}{3} = \frac{1}{2}$

Bayes Rule

- Should I change from door 1 to 2?

$$p(C=1|D=3) = \frac{p(D=3|C=1)p(C=1)}{p(D=3)} = \frac{1/2 \cdot 1/3}{1/2} = 1/3$$

$$p(C=2|D=3) = \frac{p(D=3|C=2)p(C=2)}{p(D=3)} = \frac{1 \cdot 1/3}{1/2} = 2/3$$