AAI - Probabilities

Nikolas Huhnstock

What are Probabilities

It’s all about uncertainties.

Put values on the possibilities of possible worlds.

Applications of Probabilities

- Probabilistic algorithms
  - Testing of prime numbers, Monte-Carlo evaluation of integrals
- Probabilistic analysis of algorithms (computational complexity)
  - Average complexity instead of worst
- Information Theory
  - Data compression, Error correction
- Computer Graphics
  - Random object generation
- Machine Learning
  - Classifiers, Density Estimators, Topic Models, ...

Notation and Axioms

Ω := sample space, all possible worlds ω:= one such world

P(ω) := numerical probability value for world ω

$$0 \leq P(\omega) \leq 1 \text{ for every } \omega$$

$$\sum_{\omega \in \Omega} P(\omega) = 1$$

$$P(True) = 1$$

$$P(False) = 0$$

$$P(\neg A) = 1 - P(A)$$

Propositions

A statement that either holds or not in a subset of all considered worlds.

A = It will rain tomorrow
B = two rolled dice sum up to 11

Discrete Random Variables
A random variable denoting if corresponding event occurs or not

$$P(\mathrm{A}) = \sum_{\omega \in A} P(\omega)$$

Inverse Probability

$$
\begin{aligned}
P(\neg A) &= \sum_{\omega \in \neg A} P(\omega) \\
&= \sum_{\omega \in \neg A} P(\omega) + \sum_{\omega \in A} P(\omega) - \sum_{\omega \in \neg A} P(\omega) \\
&= \sum_{\omega \in \Omega} P(\omega) - \sum_{\omega \in A} P(\omega) \\
&= 1 - P(A)
\end{aligned}
$$

Conditional Probabilities

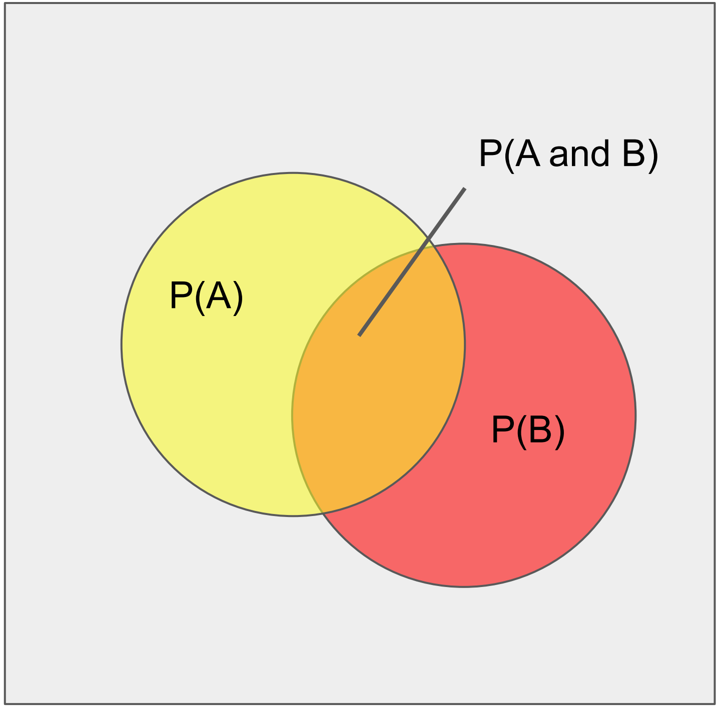
P(A) given that we know B is true.

$$P(\mathrm{A} \mid B) = \frac{P(A \wedge B)}{P(B)}$$

Example: S = ‘sick’; H = ‘headache’

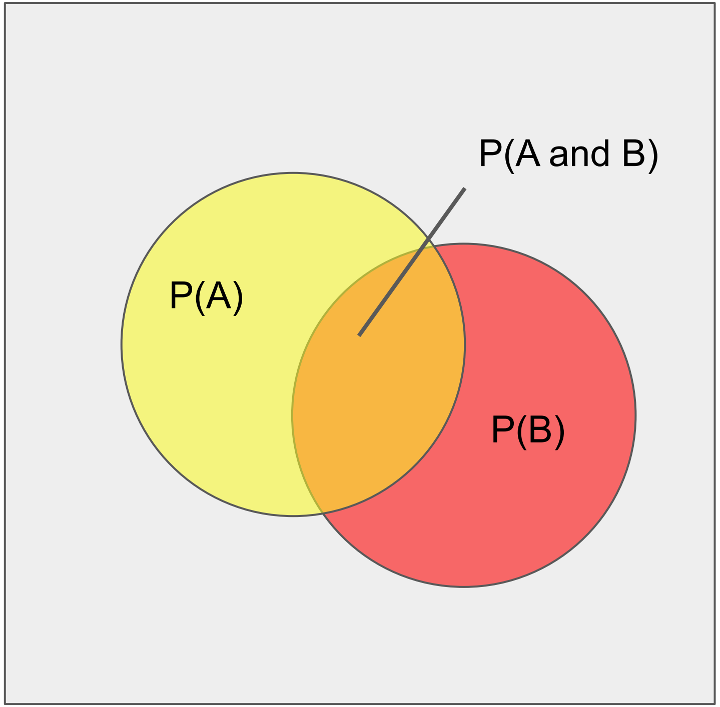
P(S) = 1/40 = 0,025
P(H) = 1/10 = 0,1

P(H|S) =



Conditional Probabilities

$$P(\mathrm{A} \mid B) = \frac{P(A \wedge B)}{P(B)}$$



Product Rule

$$P(A \wedge B) = P(A \mid B) P(B)$$

Conditional Probabilities

P(A|B) := P(A) given that we know P(B) is true.

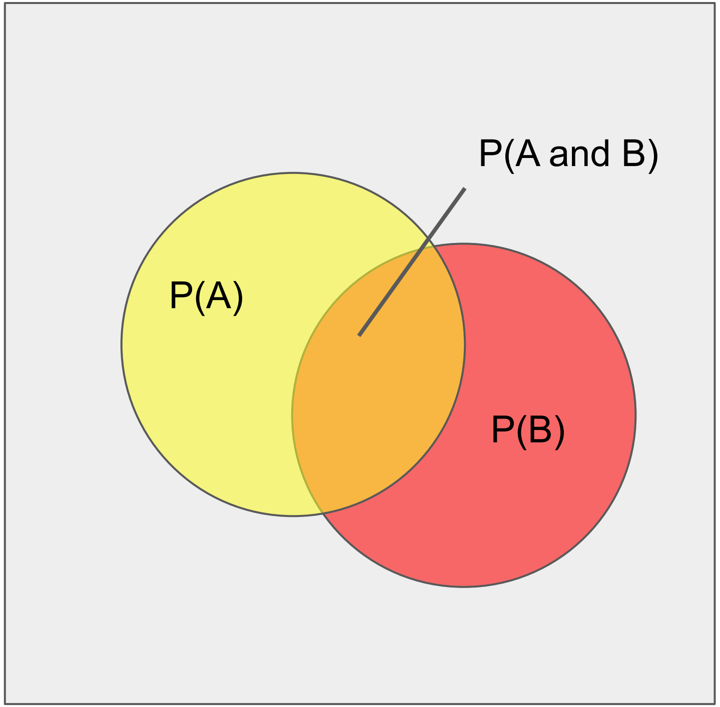
=P(A ^ B)/P(B)

Example:

S = ‘sick’
H = ‘headache’

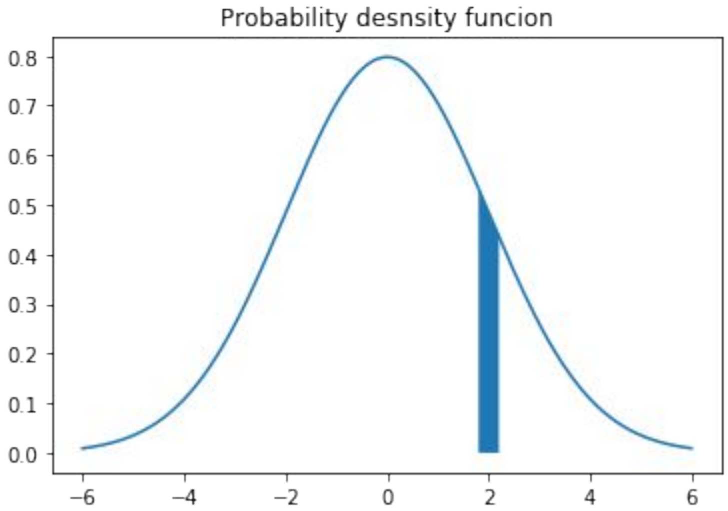
P(S) = 1/40 = 0,025
P(H) = 1/10 = 0,1

P(H|S) =



Probability Density Function (pdfs)

Used to describe continuous variables.



$$P(x) = \lim_{dx \to 0} P(x \leq X \leq +dx)/dx$$

Probabilistic Inference

Tomorrow, you wake up with a headache.
You remember that 50% percent of flus are associated with headaches so you think that you have a 50-50 chance of coming down with the flu.

Is that reasonable? Do you agree?

Probabilistic Inference

$$P(\text{cause} \mid \text{effect}) = \frac{P(\text{cause} \wedge \text{effect})}{P(\text{effect})}$$

$$P(\text{cause} \mid \text{effect}) = \frac{P(\text{effect} \mid \text{cause})P(\text{cause})}{P(\text{effect})}$$

Probabilistic Inference

What we know:

P(S) = 1/40
P(H) = 1/10
P(H|S) = ½

What are we looking for?

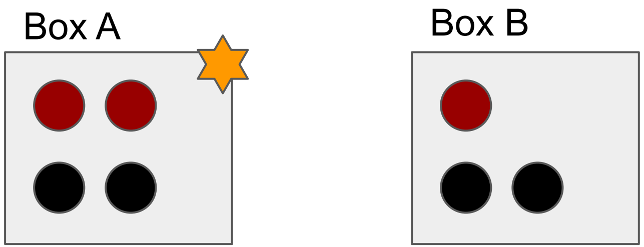
$$P(S \mid H) = \frac{P(S \wedge H)}{P(H)}$$

$$P(S \mid H) = \frac{P(S \wedge H)}{P(H)} = \frac{P(H|S)P(S)}{P(H)}$$

Bayes Rule

$$P(A \mid B) = \frac{P(B|A)P(A)}{P(B)}$$

Incorporating Information



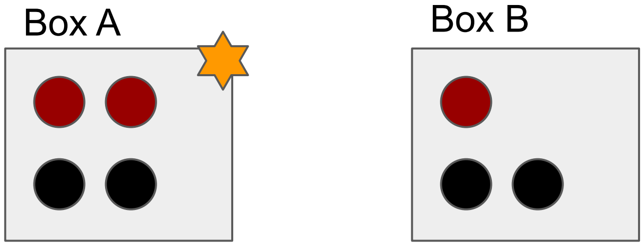
What are the probabilities of winning?

Setting 1: Blindly choose one box

Setting 2: You are allowed to draw one ball before you choose.

How does that change your chances to win? Depending if the ball is black or red?

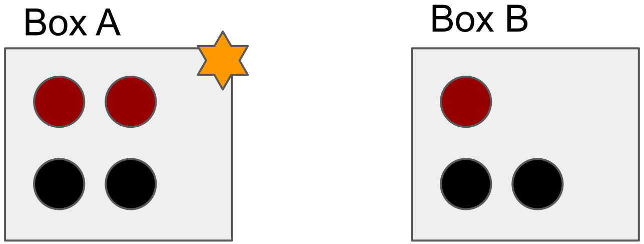
Incorporating Information



$$p(r) = p(r \mid win)p(win) + p(r \mid loose)p(loose)$$
$$= \frac{1}{2}\frac{1}{2} + \frac{1}{3}\frac{1}{2} = \frac{5}{12}$$

$$p(win \mid r) = \frac{p(r \mid win)p(win)}{p(r)} = \frac{\frac{1}{2}\frac{1}{2}}{\frac{5}{12}} = \frac{3}{5}$$

Incorporating Information



What are the probabilities of winning?

Setting 1: Blindly choose one box -> $p(win) = 0.5$

Setting 2: You are allowed to draw one ball before you choose.

$p(win|Ball = red) = \frac{3}{5} = 0.6$

$p(win|Ball = black) = 3/7 = 0.43$

How do we interpret this result?

Coffee Break

Next up Density Estimators

Joint Distributions

Student	Assignment 1	Assignment 2	Assignment 3
1	pass	pass	pass
2	pass	pass	pass
3	fail	fail	fail
. . .	. . .	. . .	. . .
n	fail	pass	fail
n+1	pass	pass	fail
n+2	pass	fail	fail

Joint Distributions

A	B	C	P(row)
0	0	0	0.05
1	0	0	0.1
0	1	0	0.025
1	1	0	0.05
0	0	1	0.025
1	0	1	0.1
0	1	1	0.05
1	1	1	0.6

Why Joint Distributions?

Inference!

$$P(A) = \sum_A P(row)$$

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)} = \frac{\sum_{A \wedge B} P(row)}{\sum_B P(row)}$$

A	B	C	P(row)
0	0	0	0.05
1	0	0	0.1
0	1	0	0.025
1	1	0	0.05
0	0	1	0.025
1	0	1	0.1
0	1	1	0.05
1	1	1	0.6

Inference

I got some evidence. What’s the chance that this hypothesis is true?

- I have a headache -> how likely is it that I have the flu?
- I passed the first assignment -> what are the chances to pass the remaining?

Applications

- Decision Making: Medicine, Help Desk Support

Density Estimator

A	B	C	P(row)
0	0	0	0.05
1	0	0	0.1
0	1	0	0.025
1	1	0	0.05
0	0	1	0.025
1	0	1	0.1
0	1	1	0.05
1	1	1	0.6

Excourse: Evaluating Density Estimators

Given a record **x**, a density estimator M can tell you how likely the record is:

$$P(\mathbf{x} \mid M)$$

Excourse: Evaluating Density Estimators

Given a record **x**, a density estimator M can tell you how likely the record is:

$$P(\mathbf{x} \mid M)$$

Given a dataset D with n records, a density estimator can tell us how likely *D* is:
(assuming all records were independently generated)

$$\begin{aligned} P(\mathcal{D} \mid M) &= P(\mathbf{x}_1 \wedge \mathbf{x}_2 \wedge \ldots \wedge \mathbf{x}_n \mid M) \\ &= \prod_{k=1}^{n} P(\mathbf{x}_k \mid M) \end{aligned}$$

Log-likelihood

Leads to log-probabilities

$$
\begin{aligned}
logP(\mathcal{D} \mid M) &= log \prod_{k=1}^{n} P(\mathbf{x}_k \mid M) \\
&= \sum_{k=1}^{n} logP(\mathbf{x}_k \mid M)
\end{aligned}
$$

What we did

Create a density estimator.

Perform inference with it.

Problem: Overfitting!

Overfitting Joint

$$P(A = 1 \mid D = 0) = ?$$

What is the problem?

A	B	C	D	P(row)
0	0	0	1	0.05
1	0	0	1	0.1
0	1	0	1	0.025
1	1	0	1	0.05
0	0	1	1	0.025
1	0	1	1	0.1
0	1	1	1	0.05
1	1	1	1	0.6

Overfitting Joint

$$P(A=1 \mid D=0) = \frac{\sum_{A \wedge \neg D} P(row)}{\sum_{\neg D} P(row)}$$

A	B	C	D	P(row)
0	0	0	1	0.05
1	0	0	1	0.1
0	1	0	1	0.025
1	1	0	1	0.05
0	0	1	1	0.025
1	0	1	1	0.1
0	1	1	1	0.05
1	1	1	1	0.6

Naive Density estimators

Joint density estimator just mirrored the data -> we need something more general.

So, now we assume that each attribute is **distributed independently** of all the others.

What does that mean?

Excourse: Independence

Combining information of multiple variables:

$$p(A, B) = P(A)P(B)$$

Under the assumption that A and B are independent.

This mean p(A | B) is independent of the value of B:

$$p(A \mid B) = P(A)$$

Excourse: Independently distributed data

For $x = (x_1, .., x_i, .., x_M) \in D$

independently distributed means that for any $i, u_1, .., u_m$

$P(x_i = v \mid x_1 = u_1, .. x_{i-1} = u_{i-1}, x_{i+1} = u_{i+1}, .., x_M = u_M)$

$= P(x_i = u_i)$

alternate formulation:

$x_i \perp x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_M$

Naive Density estimators

What is $P(A \wedge \neg B \wedge C)$?

Naive Density estimators

What is $P(A \wedge \neg B \wedge C\neg)$?

$= P(A \mid \neg B \wedge C)P(\neg B \wedge C)$

$= P(A)P(\neg B \mid C)P(C)$

$= P(A)P(\neg B)P(C)$

in general:

$P(x_1 = u_1, \ldots, x_M = u_M) = \prod_{i=1}^{M} P(x_i = u_i)$

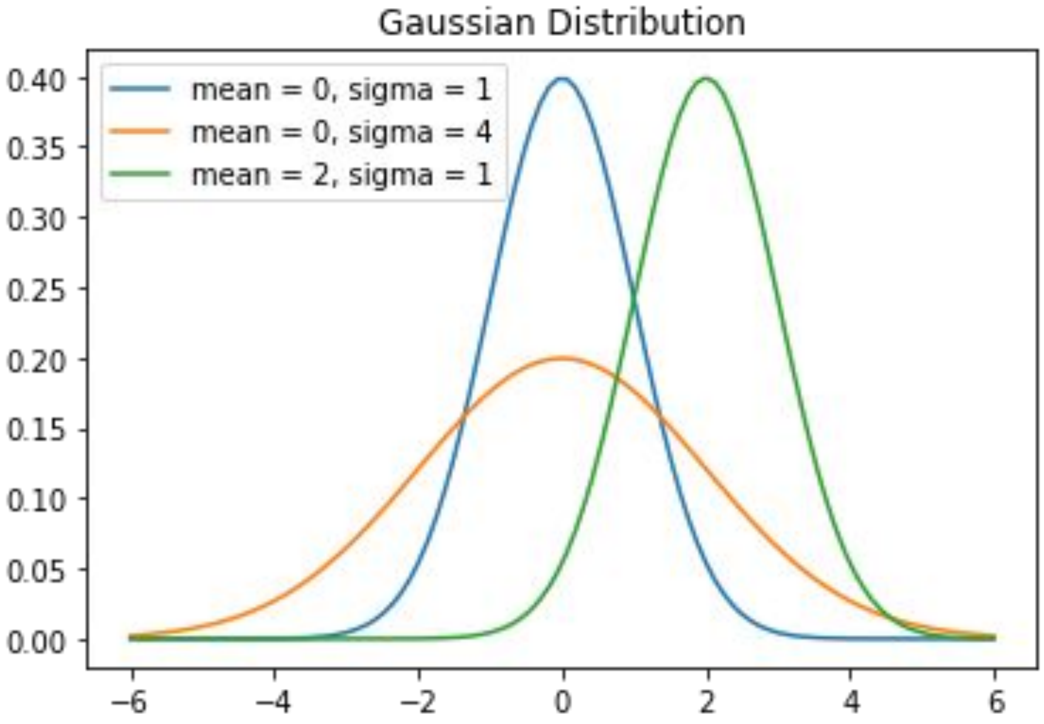
Naive Density estimators

$$P(x_i = u) = \frac{\#\text{records in which } x_i = u}{\text{total number of records}}$$

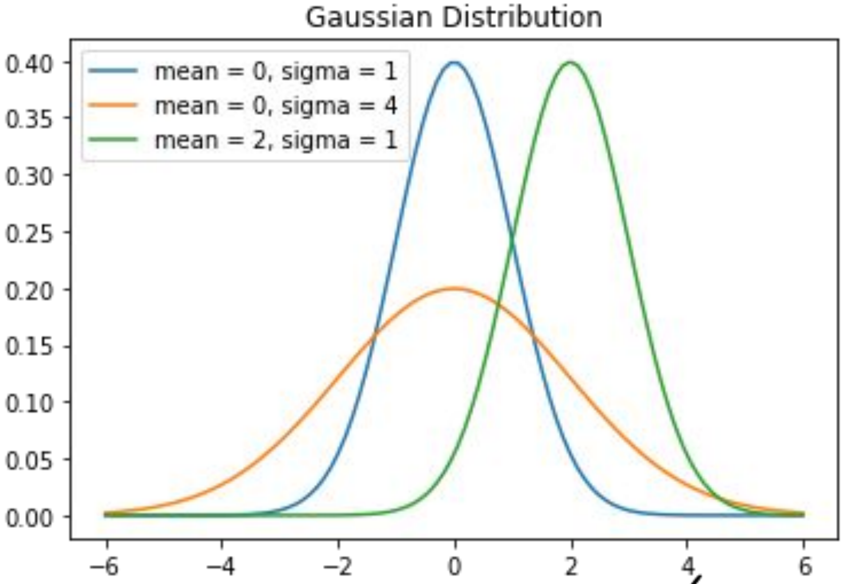
Next Learning with Maximum Likelihood

But first Gaussians.

Gaussian Distributions



Gaussian Distribution



$$\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right)$$

Gaussian Distribution (Unique Properties)

- Affine transformations (adding constants and multiplying by scalars) are Gaussians:

$$X \sim \mathcal{N}(\mu, \sigma^2)$$
$$Y = aX + b \rightarrow Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$$

- Sum Gaussians is Gaussian

$$X \sim \mathcal{N}(\mu_X, \sigma_X^2)$$
$$Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$$
$$Z = X + Y \rightarrow Z \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

Maximum Likelihood Estimation - Example

$$
\begin{aligned}
x_1, x_2, \ldots x_N &\sim \mathcal{N}(\mu, \sigma^2) \\
\mu_{MLE} &= \arg\max_{\mu} p(x_1, \ldots, x_N \mid \mu, \sigma^2) \\
&= \arg\max_{\mu} \prod_{i=1}^{N} p(x_i \mid \mu, \sigma^2) \\
&= \arg\max_{\mu} \sum_{i=1}^{N} \log p(x_i \mid \mu, \sigma^2)
\end{aligned}
$$

Maximum Likelihood Estimation - Example

$$\begin{aligned}
\mu_{MLE} &= \arg\max_{\mu} \sum_{i=1}^{N} \log p(x_i \mid \mu, \sigma^2) \\
&= \arg\max_{\mu} \frac{1}{\sqrt{2\pi\sigma}} \sum_{i=1}^{N} -\frac{(x_i - \mu)^2}{2\sigma^2} \\
&= \arg\min_{\mu} \sum_{i=1}^{N} (x_i - \mu)^2
\end{aligned}$$

Maximum Likelihood Estimation - Example

$$0 = \frac{\partial LL}{\partial \mu} = \frac{\partial}{\partial \mu} \sum_{i=1}^{N} (x_i - \mu)^2$$

$$= -\sum_{i=1}^{N} 2(x_i - \mu)$$

$$\mu_{MLE} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Maximum Likelihood Estimation - general

1. Write LL = log P(Data | θ, parameters)
2. Work out partial derivative $\partial LL/\partial\theta$
3. Set $\partial LL/\partial\theta = 0$ and solve it for θ
4. Optionally: ensure you fund a maximum and not a minimum

MLE for univariate Gaussian

1. Suppose $\boldsymbol{\theta} = (\theta_1, \ldots \theta_N)$
2. Write LL = log P(Data | θ, parameters)
3. Work out partial derivative $\partial LL / \partial \boldsymbol{\theta}$
4. Set $\partial LL / \partial \boldsymbol{\theta} = 0$ and all of them simultaneously

$$\frac{\partial LL}{\partial \Theta_1} = 0, \ldots, \frac{\partial LL}{\partial \Theta_N} = 0$$

5. Optionally: ensure you fund a maximum and not a minimum