Fundamentals of AI

Introduction and the most basic concepts

Part 3. The notion of **probability distribution, probability density function (PDF))**

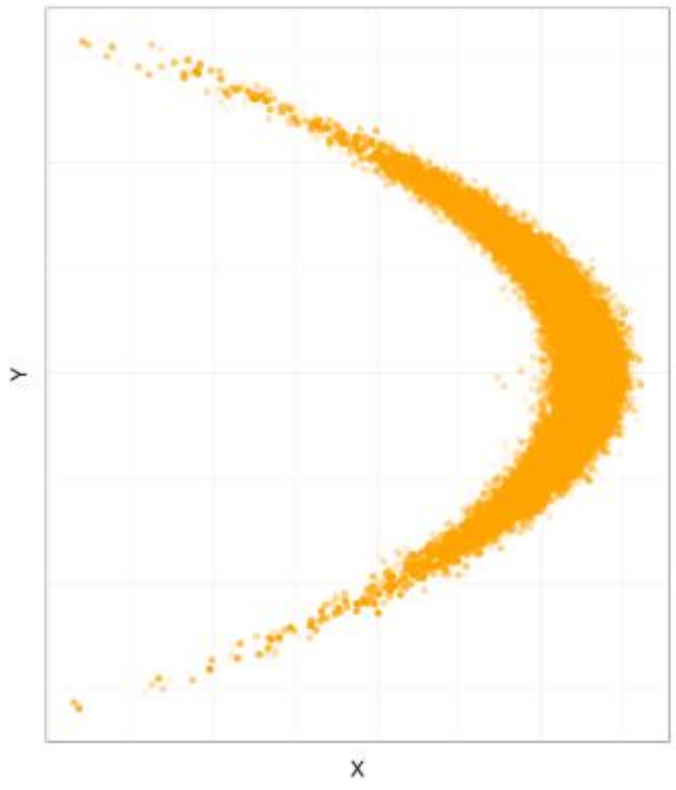
Disclaimer

- The amount of material related to probability distributions and probability densities is enormous
- This is (really!) the heart of the statistical learning theory
- Here we will just scratch the surface which will be necessary for introducing some of the machine learning methods later
- Words that should become familiar after the lecture: conditional independence, likelihood, probability density, naïve Bayes assumption, Bayesian networks, kernel density estimate, conditional distributions

Joint Probability Distribution

- Probability of any combination of features to happen
- Fundamental assumption: dataset is i.i.d. (Independent and identically distributed) sample following PDF
- If we know PDF underlying our dataset then we can predict everything (any dependence, together with uncertainties)!
- Moreover, knowing PDF we can generate infinite number of similar datasets with the same or different number of points
- *Really Platonian thing!*

'Banana-shaped probability distribution'



Probability density function (PDF)

$$f(x,y) = \exp\left(-\frac{x^2}{200} - \frac{1}{2}(y + Bx^2 - 100B)^2\right)$$

What is Likelihood?

Very generally likelihood is the probability that a given data point cloud is sampled from a given joint probability distribution.

Usually, it works with probability distributions defined by analytical functions with some parameters (statistical model)

Then it is the is the goodness of fit of a statistical model to a sample of data for given values of the unknown parameters

$$\mathcal{L}(\theta \mid x) = p_\theta(x) = P_\theta(X = x)$$

Calculations of Likelihood can deal with very small numbers, so it is convenient to work with log-likelihood: main tool in probabilistic approach to machine learning

Describing joint probability distribution

- Discrete variables : tabulations, histograms
- Continuous variables: Probability Density Function (PDF)
- Mixed type data : combining two representations

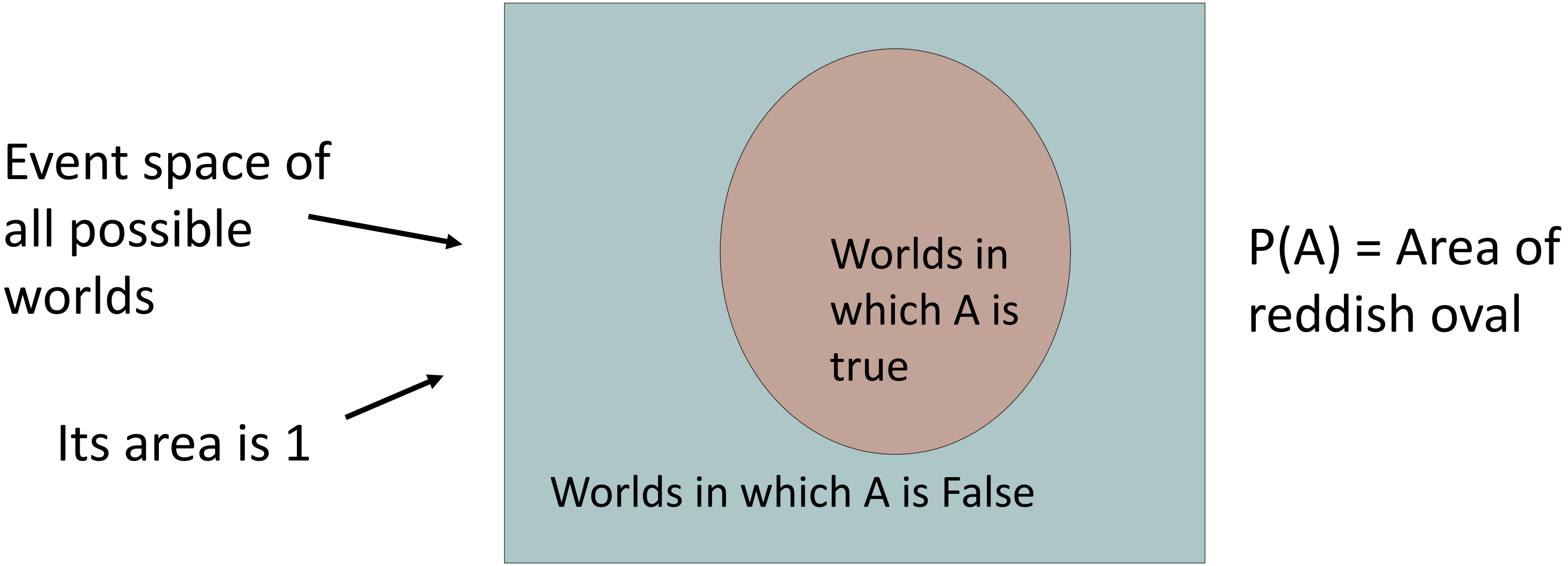
Short reminder on probability theory

- probability theory is simple in the case of discrete variables
- A is a Boolean-valued random variable if A denotes an event, and there is some degree of uncertainty as to whether A occurs.
  - Examples
    - A = The US president in 2023 will be male
    - A = You wake up tomorrow with a headache
    - A = You have COVID

Probabilities

- We write P(A) as “the fraction of possible worlds in which A is true”
- We could at this point spend 2 hours on the philosophy of this.
- But we won’t.

Visualizing A

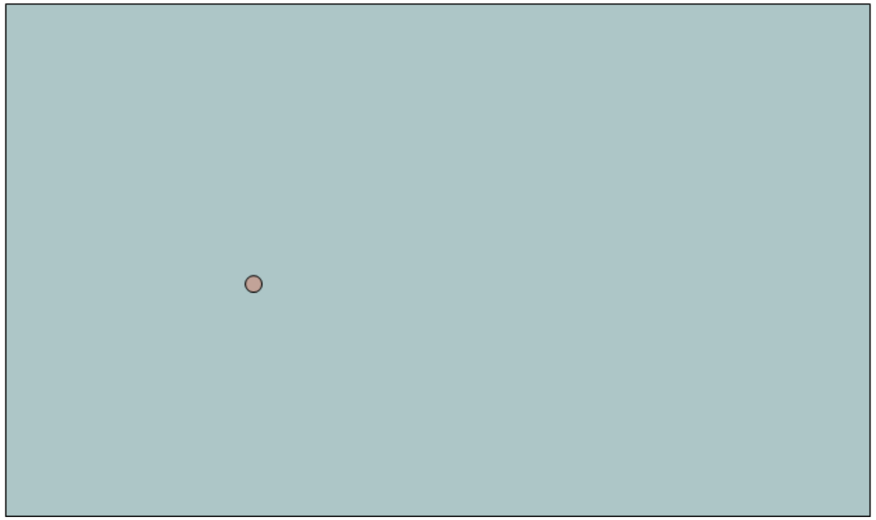


The Axioms of Probability

- $0 <= P(A) <= 1$
- $P(True) = 1$
- $P(False) = 0$
- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

Interpreting the axioms

- 0 <= P(A) <= 1
- P(True) = 1
- P(False) = 0
- P(A or B) = P(A) + P(B) - P(A and B)

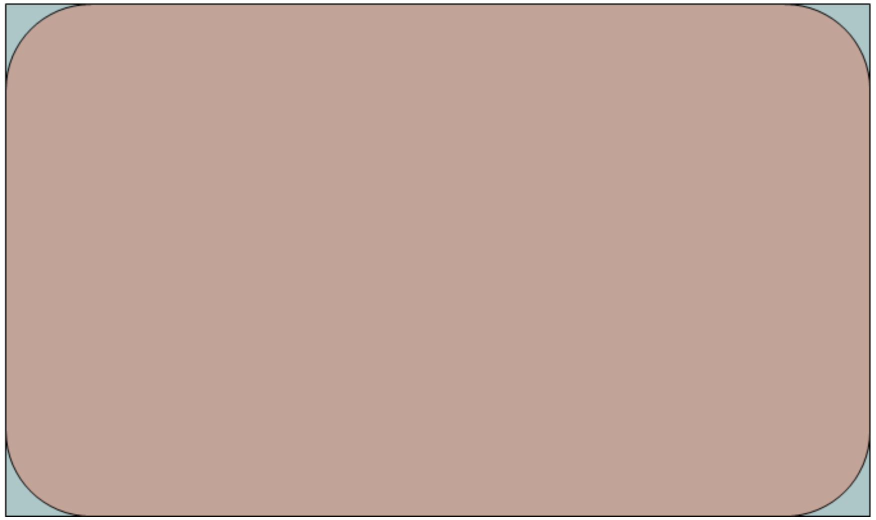


The area of A can't get any smaller than 0

And a zero area would mean no world could ever have A true

Interpreting the axioms

- 0 <= P(A) <= 1
- P(True) = 1
- P(False) = 0
- P(A or B) = P(A) + P(B) - P(A and B)

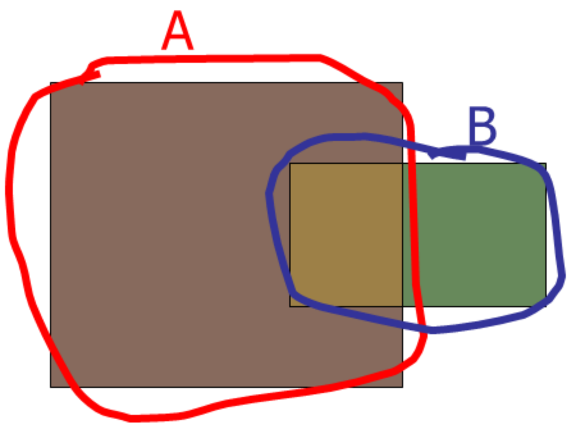


The area of A can't get any bigger than 1

And an area of 1 would mean all worlds will have A true

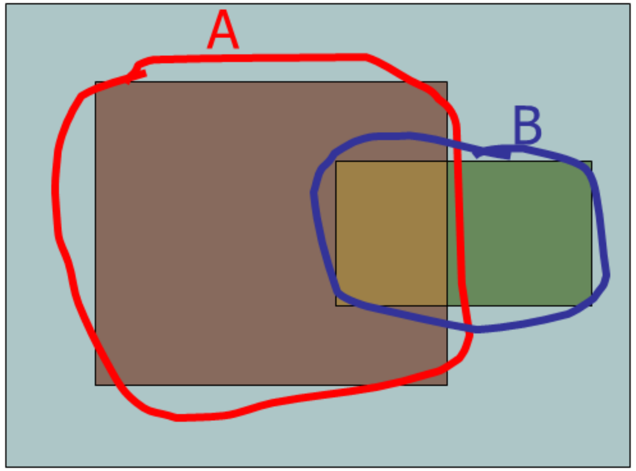
Interpreting the axioms

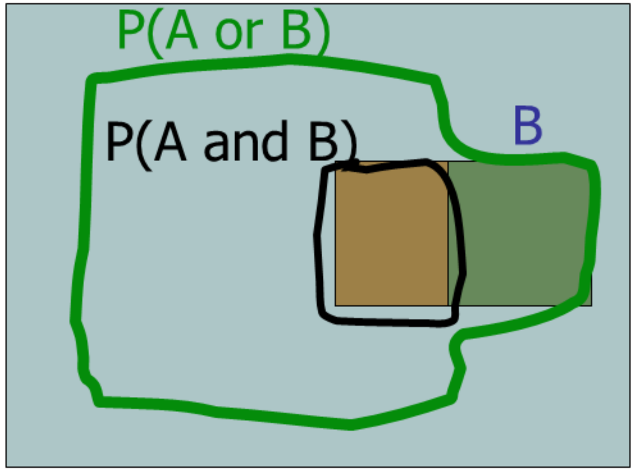
- 0 <= P(A) <= 1
- P(True) = 1
- P(False) = 0
- P(A or B) = P(A) + P(B) - P(A and B)



Interpreting the axioms

- 0 <= P(A) <= 1
- P(True) = 1
- P(False) = 0
- P(A or B) = P(A) + P(B) - P(A and B)





Simple addition and subtraction

These Axioms are Not to be Trifled With

- There have been attempts to do different methodologies for uncertainty
    - Fuzzy Logic
    - Three-valued logic
    - Dempster-Shafer
    - Non-monotonic reasoning

- But the axioms of probability are the only system with this property:

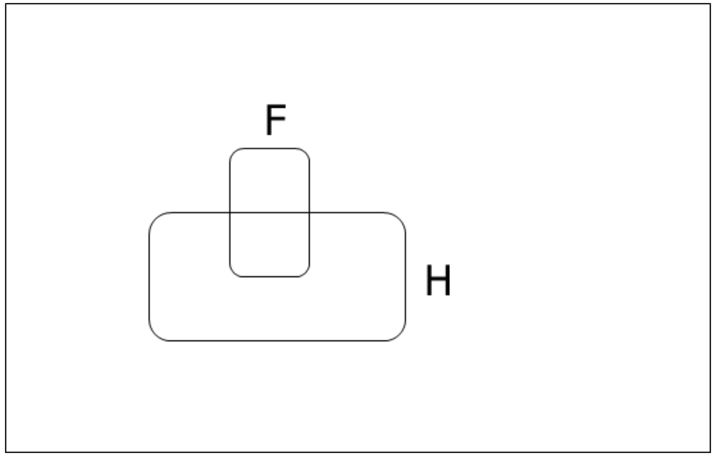
  If you gamble using them you can't be unfairly exploited by an opponent using some other system [di Finetti 1931]

- All this was just elementary applications of basic set theory

- The actual probability theory starts from the notion of *conditional probability and conditional independence!*

Conditional Probability

- P(A|B) = Fraction of worlds in which B is true that also have A true

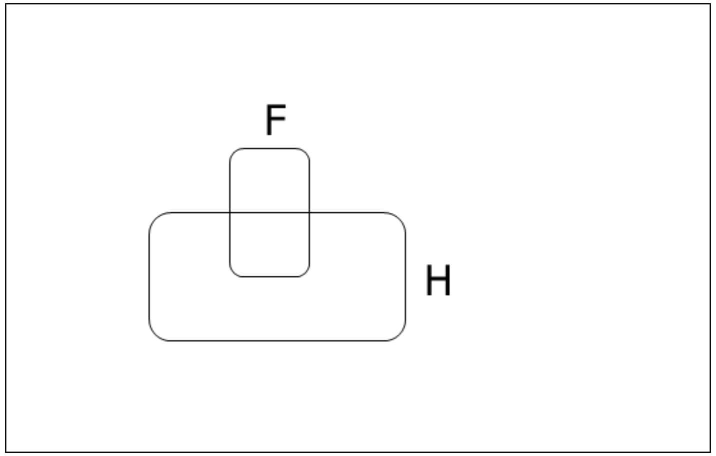


H = “Have a headache”
F = “Coming down with Flu”

P(H) = 1/10
P(F) = 1/40
P(H|F) = 1/2

“Headaches are rare and flu is rarer, but if you’re coming down with ‘flu there’s a 50-50 chance you’ll have a headache.”

Conditional Probability



H = “Have a headache”
F = “Coming down with Flu”

P(H) = 1/10
P(F) = 1/40
P(H|F) = 1/2

P(H|F) = Fraction of flu-inflicted worlds in which you have a headache

$$= \frac{\#\text{worlds with flu and headache}}{\#\text{worlds with flu}}$$

$$= \frac{\text{Area of “H and F” region}}{\text{Area of “F” region}}$$

$$= \frac{P(H \wedge F)}{P(F)}$$

Definition of Conditional Probability

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

Corollary: The Chain Rule

$$P(A \wedge B) = P(A|B)\,P(B)$$

Bayes rule

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

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London,* **53:370-418**

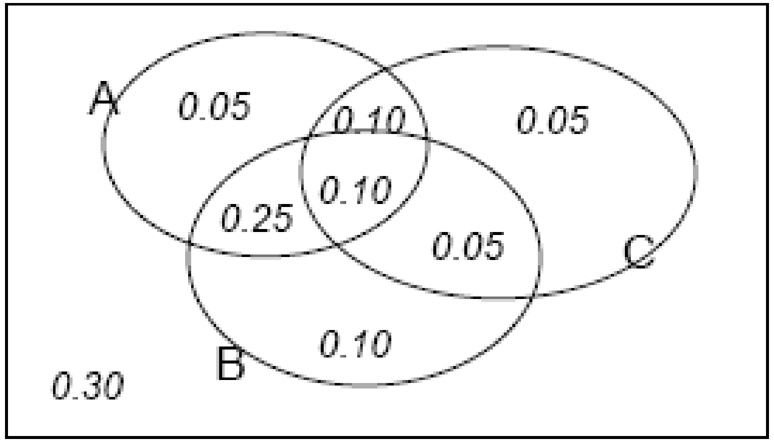
The Joint Distribution

Example: Boolean variables A, B, C

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all combinations of values of your variables (if there are M Boolean variables then the table will have 2^M rows).
2. For each combination of values, say how probable it is.
3. If you subscribe to the axioms of probability, those numbers must sum to 1.

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10



Using the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

P(Poor Male) = 0.4654

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

Inference with the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

$$P(E_1 \mid E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum\limits_{\text{rows matching } E_1 \text{ and } E_2} P(\text{row})}{\sum\limits_{\text{rows matching } E_2} P(\text{row})}$$

Inference with the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

$$P(E_1 \mid E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{\text{rows matching } E_1 \text{ and } E_2} P(\text{row})}{\sum_{\text{rows matching } E_2} P(\text{row})}$$

P(Male | Poor) = 0.4654 / 0.7604 = 0.612

Joint distributions

- Good news

  Once you have a joint distribution, you can ask important questions about stuff that involves a lot of uncertainty

- Bad news

  Impossible to create for more than about ten attributes because there are so many numbers needed when you build the damn thing.