Review on Math for AI

Daifeng Wang

daifeng.wang@wisc.edu

University of Wisconsin, Madison

Based on slides from Xiaojin Zhu (http://pages.cs.wisc.edu/~jerryzhu/cs540.html), modified by Daifeng Wang

Outline

- Probability and inference
  - Axioms of probability
  - Joint, Marginal, Conditional probability
  - Bayes rule
  - Independence, Conditional independence
  - Expected value
  - Maximum Likelihood Estimation (MLE)
  - Maximum a posteriori (MAP) estimation

Sample space

- A space of events that we assign probabilities to
- Events can be binary, multi-valued, or continuous
- Events are mutually exclusive

Random variable

- A variable, x, whose domain is the sample space, and whose value is somewhat uncertain

Probability for discrete events

- Probability P(x=a) or P(a) is the fraction of times x takes value a

Probability table

- Weather

Sunny	Cloudy	Rainy
200/365	100/365	65/365

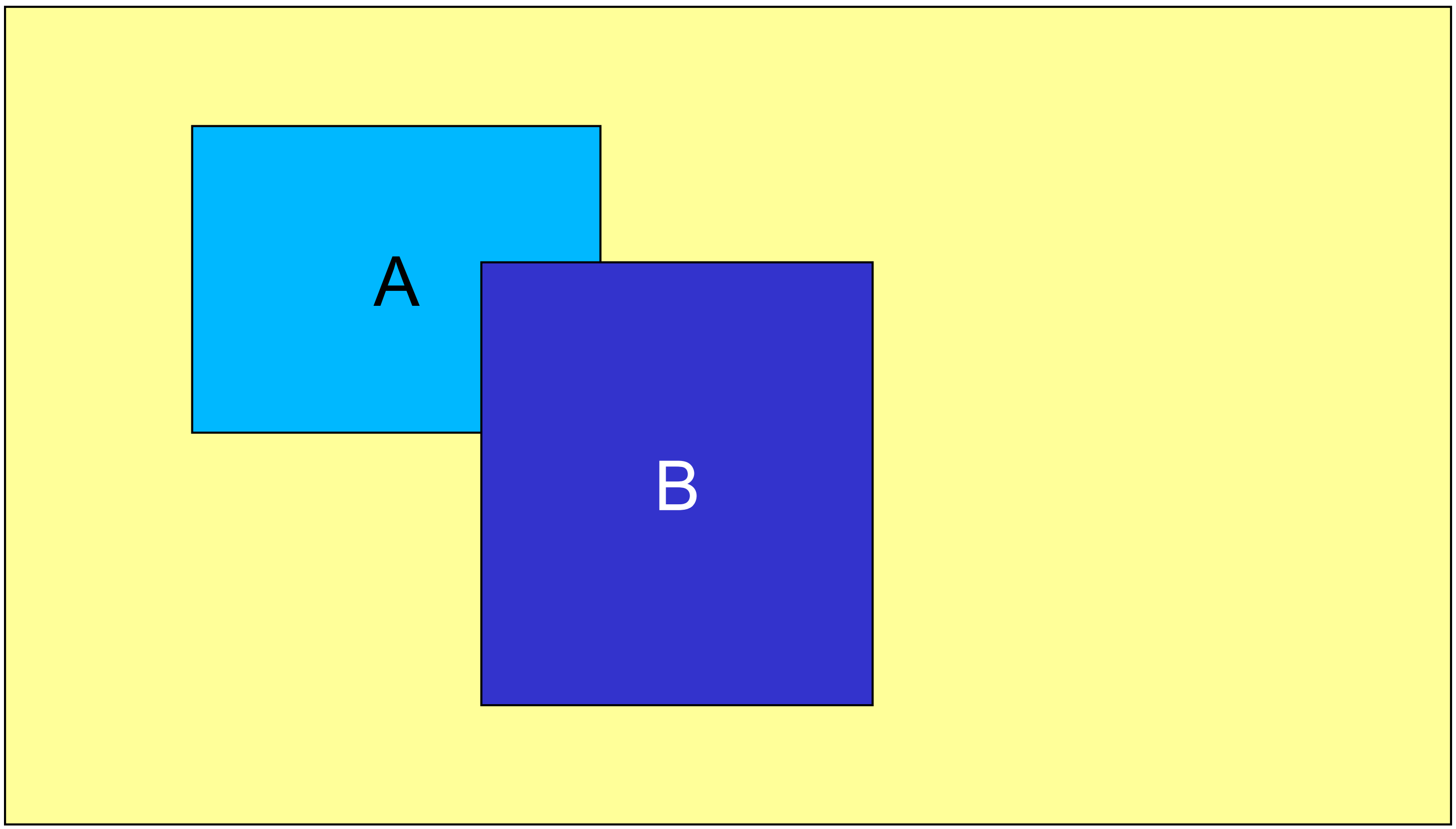
- P(Weather = sunny) = P(sunny) = 200/365

- P(Weather) = {200/365, 100/365, 65/365}

- For now we'll be satisfied with obtaining the probabilities by counting frequency from data…

The axioms of probability

- $P(A) \in [0,1]$
- $P(true)=1$, $P(false)=0$
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$



Sample space

Some theorems derived from the axioms

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

- If A can take k different values $a_1 \ldots a_k$:

  $$P(A=a_1) + \ldots P(A=a_k) = 1$$

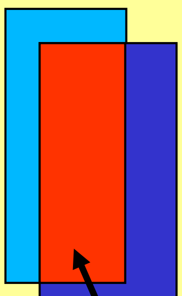
- $P(B) = P(B \wedge \neg A) + P(B \wedge A)$, if A is a binary event

- $P(B) = \Sigma_{i=1\ldots k} P(B \wedge A=a_i)$, if A can take *k* values

Joint probability

- The joint probability P(A=a, B=b) is a shorthand for P(A=a $\wedge$ B=b), the probability of both A=a and B=b happen

P(A=a), e.g. P(1st word on a random page = “San”) = 0.001
(possibly: San Francisco, San Diego, …)



P(B=b), e.g. P(2nd word = “Francisco”) = 0.0008
(possibly: San Francisco, Don Francisco, Pablo Francisco …)

P(A=a,B=b), e.g. P(1st =“San”,2nd =“Francisco”)=0.0007

Marginal probability

- Sum over other variables

weather

temp		Sunny	Cloudy	Rainy
	hot	150/365	40/365	5/365
	cold	50/365	60/365	60/365
	Σ	200/365	100/365	65/365

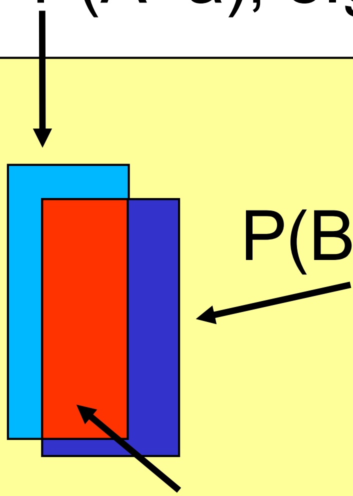
P(Weather)={200/365, 100/365, 65/365}

- The name comes from the old days when the sums are written on the margin of a page

Conditional probability

- The conditional probability P(A=a | B=b) is the fraction of times A=a, within the region that B=b

P(A=a), e.g. P(1^{st} word on a random page = "San") = 0.001



P(B=b), e.g. P(2^{nd} word = "Francisco") = 0.0008

P(A=a | B=b), e.g. P(1^{st}="San" | 2^{nd} ="Francisco")=**0.875**

(possibly: San, Don, Pablo …)

Although "San" is rare and "Francisco" is rare, given "Francisco" then "San" is quite likely!

The chain rule

- From the definition of conditional probability we have the chain rule

$$P(A, B) = P(B) * P(A \mid B)$$

- It works the other way around

$$P(A, B) = P(A) * P(B \mid A)$$

- It works with more than 2 events too

$P(A_1, A_2, \ldots, A_n) =$

$P(A_1) * P(A_2 \mid A_1) * P(A_3 \mid A_1, A_2) * \ldots * P(A_n \mid A_1, A_2 \ldots A_{n-1})$

Inference with Bayes' rule

- P(A|B) = P(B|A)P(A) / P(B)   Bayes' rule
- Why do we make things this complicated?
  - Often P(B|A), P(A), P(B) are easier to get
  - Some names:
    - **Prior P(A)**: probability before any evidence
    - **Likelihood P(B|A)**: assuming A, how likely is the evidence
    - **Posterior P(A|B)**: conditional prob. after knowing evidence
    - **Inference**: deriving unknown probability from known ones
- In general, if we have the full joint probability table, we can simply do P(A|B)=P(A, B) / P(B) – more on this later…

Independence

- Two events A, B are independent, if (the following are equivalent)
  - $P(A, B) = P(A) * P(B)$
  - $P(A \mid B) = P(A)$
  - $P(B \mid A) = P(B)$

Conditional independence

- In general A, B are conditionally independent given C
  - if $P(A \mid B, C) = P(A \mid C)$, or
  - $P(B \mid A, C) = P(B \mid C)$, or
  - $P(A, B \mid C) = P(A \mid C) * P(B \mid C)$

Expected values

- The *expected value* of a random variable that takes on numerical values is defined as:

$$\mathbf{E}[X] = \sum_x x P(x)$$

This is the same thing as the *mean*

- We can also talk about the expected value of a function of a random variable

$$\mathbf{E}[g(X)] = \sum_x g(x) P(x)$$

Suppose we have data $\mathcal{D} = \{x^{(i)}\}_{i=1}^{N}$

Maximum Likelihood Estimation (MLE)

Find optimal θ^* to maximize the likelihood given the data

$$\theta^*_{MLE} = argmax_\theta P(D|\theta)$$

Maximum a posteriori (MAP) estimation

$$\theta^*_{MAP} = argmax_\theta \; P(\theta|\mathcal{D})$$ Posterior

$$= argmax_\theta \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})}$$ Bayes rule

$$= argmax_\theta P(\mathcal{D}|\theta)P(\theta)$$

Play outside or not?

- If weather is sunny, would you like to play outside?

Posterior probability P(Yes|Sunny) vs P(No|Sunny)

- Weather = {Sunny, Rainy, Overcast}
- Play = {Yes, No}
- Observed data {Weather, Play on Day m}, m=1,2,…,N

How can we calculate posterior probabilities?

$$P(Play|Weather) = \frac{P(Weather|Play)P(Play)}{P(Weather)}$$

Bayes rule

Play outside or not?

- **Step 1**: Convert the data to a frequency table of Weather and Play

- **Step 2**: Based on the frequency table, calculate likelihoods $P(Weather|Play)$ and priors $P(Play)$

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

Likelihood table				
Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		

$$e.g., P(Play = Yes) = \frac{9}{14} = 0.64$$

$$P(Sunny|Yes) = \frac{3}{9} = 0.33$$

https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

Play outside or not?

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

Likelihood table				
Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		

- **Step 3**: Based on the likelihoods and priors, calculate posteriors $P(\text{Play} \mid \text{Weather})$

- P(No|Sunny)

=P(Sunny|No)*P(No)/P(Sunny)

=0.4*0.36/0.36

=0.4

- P(Yes|Sunny)

=P(Sunny|Yes)*P(Yes)/P(Sunny)

=0.33*0.64/0.36

=0.6

- P(Yes|Sunny)>P(No|Sunny), you should go outside and play!