L665 ML for NLP

Damir Cavar

Indiana University

January 2018

Probability Theory

- Introduction, take a Statistics class for a deeper intro

- Topics:
  - How likely is an event to occur?

Probability Space

- **Experiment** (or **trial**): Flipping three coins
- **Outcomes**: Possible results
  - e.g. coin 1 = tail, coin 2 = head, coin 3 = tail (THT)
- **Event**: a set of results
  - e.g. two tails and one head = {HTT, THT, TTH}
- **Sample space (Ω)**: set of possible outcomes
  - Discrete sample space: countable infinite outcomes {1, 2, 3, …}, or finite outcomes {H, T}
  - Continuous sample space: uncountable infinite outcomes {0.45546…, 8.5621…, …}, weight, height, etc.

Sample Space and Events

- Tossing a die:
  - Outcome unknown.
  - Set of possible outcomes is known.
- **Sample space (Ω)**: {1, 2, 3, 4, 5, 6}
- **Any subset of Ω is an event**: {2, 4, 6} (*even numbers*)
- The possible events of tossing a die once:
  - {{1}, {2}, {3}, {4}, {5}, {6}}
- Tossing it twice (6^2):
  - {{1, 1}, {1, 2}, {1, 3}, … {6, 6}}

Principles of Counting

- **Multiplication Principle**
  - For two independent events P and Q, if P can happen in p ways and Q in q ways, P and Q can happen in $p * q$ ways.

- **Addition Principle**
  - For two independent events P and Q, if P can happen in p ways and Q in q ways, P or Q can happen in $p + q$ ways.

Example

- If there are 3 roads from Bloomington to Indianapolis and 5 roads from Indianapolis to Chicago, how many ways are there to get from Bloomington to Chicago?

Example

- If there are 3 roads from Bloomington to Indianapolis and 5 roads from Indianapolis to Chicago, how many ways are there to get from Bloomington to Chicago?

  - Q (Bloomington to Indy) has 3 possibilities
  - P (Indy to Chicago) has 5 possibilities

  - Possibilities: 3 * 5 = 15

Example

- If there are 3 roads going north from Bloomington and 2 roads going south, how many roads are there going south or north?

Example

- If there are 3 roads going north from Bloomington and 2 roads going south, how many roads are there going south **or** north?

  - Q has 3 possibilities
  - P has 2 possibilities

  - Q or P = 3 + 2 = 5

Example

- How many different 7-place license plates are possible, if the first two places are for letters and the other 5 places for numbers?

- John, Jim, Jack, and Jay form a band consisting of 4 instruments.
  - If each of them can play all 4 instruments, how many different arrangements are possible?
  - If John and Jim can play all 4 instruments, but Jack and Jay each can only play piano and drums?

Probability Functions

- If *A* is an event, *P(A)* is the probability of *A* occurring:
  - $0 \leq P(A) \leq 1$

- A **Probability Function (Distribution)** distributes a probability mass of 1 over Ω (sample space):
  - $P : \mathcal{F} \to [0,1]$ with $P(\Omega) = 1$
  - $A_j \in \mathcal{F} : P(\bigcup_{j=1}^{\infty} A_j) = \sum_{j=1}^{\infty} P(A_j)$

Probability Functions

- A Probability Function (Distribution) distributes a probability mass of 1 over Ω (the sample space):
  - $P : \mathcal{F} \to [0,1]$ with $P(\Omega) = 1$
  - $A_j \in \mathcal{F} : P(\bigcup_{j=1}^{\infty} A_j) = \sum_{j=1}^{\infty} P(A_j)$
- The probability of any event A_j happening is the sum of the probabilities of any individual event happening.

Example

- Tossing a fair coin 3 times:
  - What is the probability of exactly 2 heads coming up?

- Sample space:
  - Ω = { HHH, HHT, HTH, HTT, THH, THT, TTH, TTT }
  - Favorable events: A = { HHT, HTH, THH }

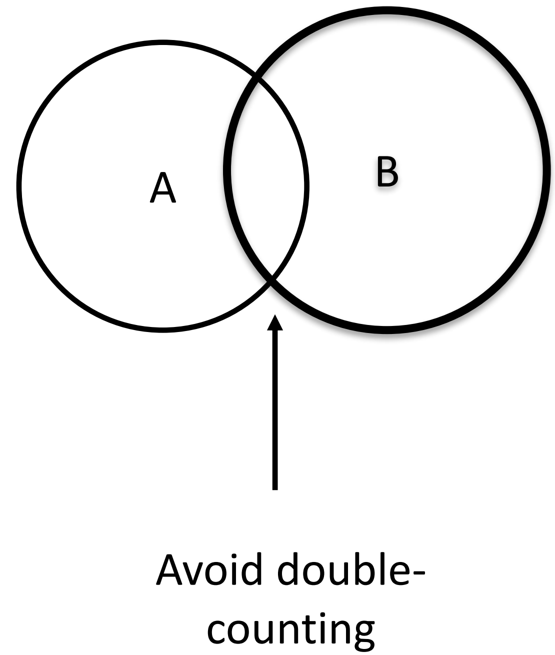
- Uniform distribution of fair coin, i.e. each outcome is equally likely:

  - 8 possible outcomes, each: $P(x) = \frac{1}{8}$

  - 3 favorable outcomes: $\mathrm{P(A)} = \frac{favorable\ outcomes}{possible\ outcomes} = \frac{3}{8}$

Non-disjoint Set Additivity

- The probability of the union of A and B:
  - Adding the probability of A and B
  - Subtracting the intersection to avoid double counting.



$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Conditional Probability

- The conditional probability of an event A occurring if an event B already occurred is: $P(A|B)$
  - Prior probability of A: P(A)
  - Posterior probability of A (after additional information B): $P(A|B)$

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

- Joint probability: $P(A, B)$ (or $P(AB)$)
  - The intersection of two events A and B
  - Note: P(AB) = P(BA)

Conditional Probability

- Tossing a coin twice
- All four events in the sample space are equally likely: $\Omega = \{ HH, HT, TH, TT \}$

- What is the conditional probability that both flips result in heads, given that the first flip does?

Conditional Probability

- Drawing cards
- Given that we draw a red card, what is the probability that it is a four?

- $P(four|red) = ?$
- $P(four) = \frac{4}{52}$
- $P(red) = \frac{26}{52}$

Conditional Probability

- Drawing cards
- Given that we draw a red card, what is the probability that it is a four?

- $P(four|red) = \frac{\frac{2}{52}}{\frac{26}{52}} = \frac{\frac{1}{26}}{\frac{1}{2}} = 2\frac{1}{26} = \frac{1}{13}$
- $P(four) = \frac{4}{52}$
- $P(red) = \frac{26}{52}$

Chain Rule

- Conditional Probability
  - $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B \cap A)}{P(A)}$
- Multiplication:
  - $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A) = P(B \cap A)$
- Chain Rule (used later with Markov Chains)
  - $P(A_1 \cap A_2 \cap ... \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2)...P(A_n| \cap_{i=1}^{n-1} A_i)$
  - Multiply:
    - Probability of first event
    - Probability of second event given, first event
    - …
    - Probability of nth event, given all previous events

Independent Events

- Two events are independent, if one does not affect the probability of the other.
- Events A and B are independent, if:
  - $P(A) = P(A|B)$
  - or $P(A \cap B) = P(A)P(B)$
- Two events are independent, if:
  - The probability of seeing them together is the same as the product of their individual probabilities.
  - Example: tossing two fair coins

Bayes' Theorem

- Calculating P(B|A) in terms of P(A|B)

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

- Note: P(A) is a normalizing constant, the same for every event. If we want to find the value of B that maximizes the equation, we can ignore the denominator:

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

Example

- Maximizing the probability of an event given two probabilistic distributions.

  - Texts on topic 1
  - Texts on topic 2

  - Text on unknown topic

- How can we calculate the topic of the third text?