NLP

Introduction to NLP

Probabilities

Probabilistic Reasoning

- Very important for language processing
- Example in speech recognition:
  - "recognize speech" vs "wreck a nice beach"
- Example in machine translation:
  - "l'avocat general": "the attorney general" vs. "the general avocado"
- Probabilities make it possible to combine evidence from multiple sources in a systematic way.

Probabilities

- Probability theory
  - predicting how likely it is that something will happen
- Experiment (trial)
  - e.g., throwing a coin
- Possible outcomes
  - heads or tails
- Sample spaces
  - discrete or continuous
- Events
  - Ω is the certain event
  - $\emptyset$ is the impossible event
  - event space – all possible events

Probabilities

- Probabilities
  - numbers between 0 and 1
- Probability distribution
  - distributes a probability mass of 1 throughout the sample space Ω.
- Example:
  - A fair coin is tossed three times.
  - What is the probability of 3 heads?
  - What is the probability of 2 heads?

Meaning of Probabilities

- Frequentist
  - I threw the coin 10 times and it turned up heads 5 times
- Subjective
  - I am willing to bet 50 cents on heads

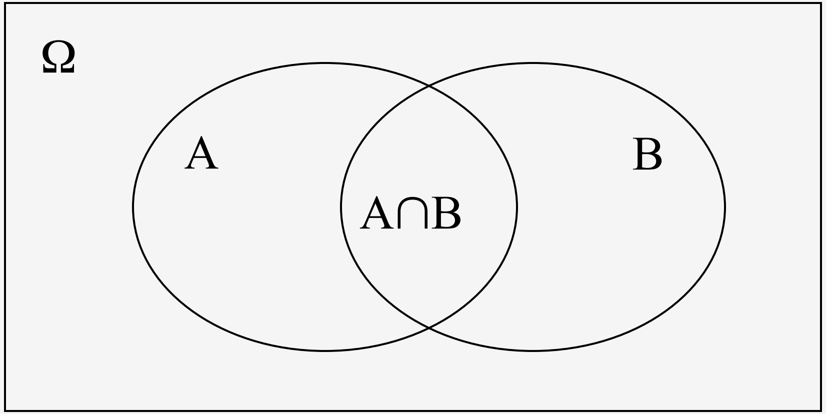
Properties of Probabilities

- $p(\emptyset) = 0$
- $P(\text{certain event}) = 1$
- $p(X) \leq p(Y)$, if $X \subseteq Y$
- $p(X \cup Y) = p(X) + p(Y)$, if $X \cap Y = \emptyset$

Conditional Probability

- Prior and posterior probability
- Conditional probability

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



Conditional Probability

- Six-sided fair die
  - P(D even)=?
  - P(D>=4)=?
  - P(D even|D>=4)=?
  - P(D odd|D>=4)=?
- Multiple conditions
  - P(D odd|D>=4, D<=5)=?

Conditional Probability

- Six-sided fair die
  - P(D even)=3/6=1/2
  - P(D>=4)=3/6=1/2
  - P(D even|D>=4)=2/3
  - P(D odd|D>=4)=1/3
- Multiple conditions
  - P(D odd|D>=4, D<=5)=1/2

The Chain Rule

- $P(w_1,w_2,w_3\ldots w_n) = ?$
- Using the chain rule:
  - $P(w_1,w_2,w_3\ldots w_n) = P(w_1)\ P(w_2|w_1)\ P(w_3|w_1,w_2)\ldots P(w_n|w_1,w_2\ldots w_{n-1})$
- This rule is used in many ways in statistical NLP, more specifically in Markov Models

Independence

- Two events are independent when
  - $P(A \cap B) = P(A)P(B)$
- Unless $P(B)=0$ this is equivalent to saying that $P(A) = P(A|B)$
- If two events are not independent, they are considered dependent

Adding vs. Removing Constraints

- Adding constraints
  - P(walk=yes|weather=nice)
  - P(walk=yes|weather=nice,freetime=yes,crowded=yes)
  - More accurate
  - But more difficult to estimate
- Removing constraints (Backoff)
  - P(walk=yes|weather=nice,freetime=yes,crowded=yes)
  - P(walk=yes|weather=nice,freetime=yes)
  - P(walk=yes|weather=nice)
  - Note that it is *not* possible to do backoff on the left hand side of the conditional

Random Variables

- Simply a function: $\boxed{X: \Omega \rightarrow R^n}$
- The numbers are generated by a *stochastic process* with a certain probability distribution
- Example
  - the discrete random variable X that is the sum of the faces of two randomly thrown fair dice
- Probability mass function (pmf) which gives the probability that the random variable has different numeric values: $\boxed{P(x) = P(X = x) = P(A_x) \text{ where } A_x = \{\omega \in \Omega : X(\omega) = x\}}$

Random Variables

- If a random variable X is distributed according to the pmf p(*x*), then we write X ~ p(*x*)
- For a discrete random variable, we have

$$\sum p(x_i) = P(\Omega) = 1$$

Example

- $p(1) = 1/6$
- $p(2) = 1/6$
- etc.
- $\mathbf{P(D)}=?$
- $\mathbf{P(D)} = \{1/6, 1/6, 1/6, 1/6, 1/6, 1/6\}$
- $\mathbf{P(D|}\text{odd}) = \{1/3, 0, 1/3, 0, 1/3, 0\}$

NLP