Natural Language Processing

Introduction to Probability

Joakim Nivre

Uppsala University
Department of Linguistics and Philology
joakim.nivre@lingfil.uu.se

Probability and Statistics

"Once upon a time, there was a . . ."

Probability and Statistics

"Once upon a time, there was a . . ."

- Can you guess the next word?
- Hard in general, because language is not deterministic
- But some words are more likely than others

Probability and Statistics

"Once upon a time, there was a . . ."

- Can you guess the next word?
- Hard in general, because language is not deterministic
- But some words are more likely than others

- We can model uncertainty using probability theory
- We can use statistics to ground our models in empirical data

The Mathematical Notion of Probability

- The probability of A, $P(A)$, is a real number between 0 and 1:
  1. If $P(A) = 0$, then A is impossible (never happens)
  2. If $P(A) = 1$, then A is necessary (always happens)
  3. If $0 < P(A) < 1$, then A is possible (may happen)

The Mathematical Notion of Probability

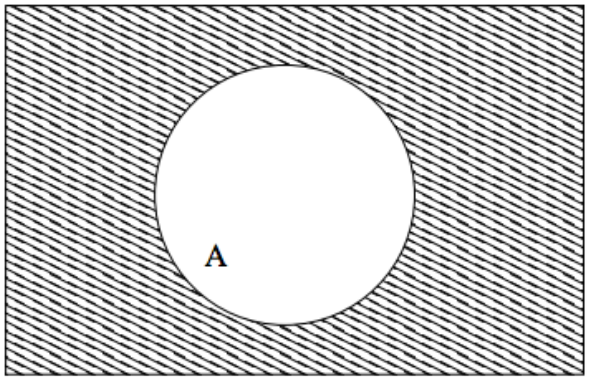
- The probability of A, $P(A)$, is a real number between 0 and 1:
  1. If $P(A) = 0$, then A is impossible (never happens)
  2. If $P(A) = 1$, then A is necessary (always happens)
  3. If $0 < P(A) < 1$, then A is possible (may happen)

- A is an event in a sample space Ω
  - Sample space = all possible outcomes of an "experiment"
  - Event = a subset of the sample space
  - Events can be described as a variable taking a certain value
    1. $\{w \in \Omega \mid w \text{ is a noun}\} \Leftrightarrow \text{PoS} = \text{noun}$
    2. $\{s \in \Omega \mid s \text{ consists of 8 words}\} \Leftrightarrow \#\text{Words} = 8$

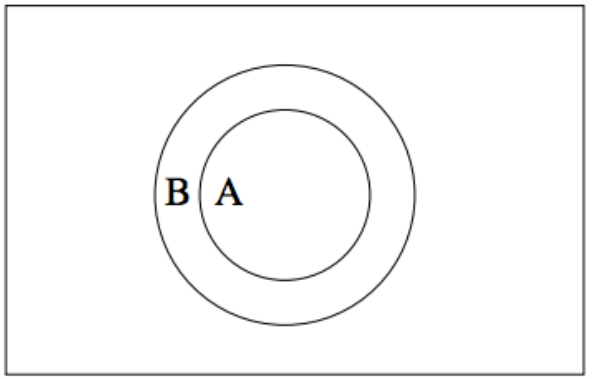
Logical Operations on Events

- Often we are interested in combinations of two or more events
- This can be represented using set theoretic operations
- Assume a sample space Ω and two events A and B:
  1. Complement $\overline{A}$ (also A') = all elements of Ω that are not in A
  2. Subset $A \subseteq B$ = all elements of A are also elements of B
  3. Union $A \cup B$ = all elements of Ω that are in A or B
  4. Intersection $A \cap B$ = all elements of Ω that are in A and B

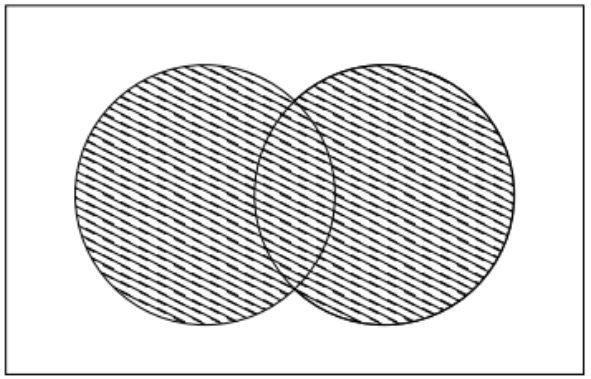
Venn Diagrams



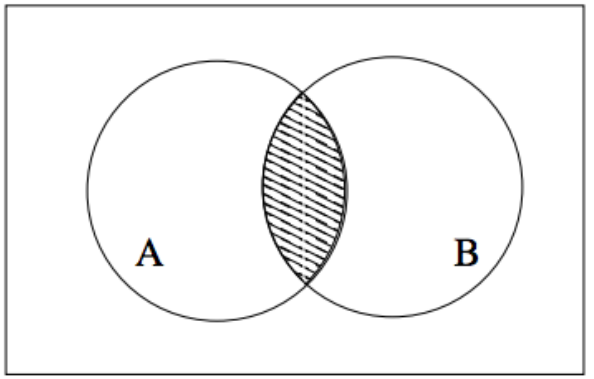
$\bar{A}$



$A \subseteq B$



$A \cup B$



$A \cap B$

Axioms of Probability

- $P(A)$ = The probability of event A
- Axioms:
  1. $P(A) \geq 0$
  2. $P(\Omega) = 1$
  3. If A and B are disjoint, then $P(A \cup B) = P(A) + P(B)$

Probability of an Event

- If A is an event and $\{x_1, \ldots, x_n\}$ its individual outcomes, then

$$P(A) = \sum_{i=1}^{n} P(x_i)$$

- Assume all 3-letter strings are equally probable
- What is the probability of a string of three vowels?

Probability of an Event

- If A is an event and $\{x_1, \ldots, x_n\}$ its individual outcomes, then

$$P(A) = \sum_{i=1}^{n} P(x_i)$$

- Assume all 3-letter strings are equally probable
- What is the probability of a string of three vowels?
  1. There are 26 letters, of which 6 are vowels
  2. There are $N = 26^3$ 3-letter strings
  3. There are $n = 6^3$ 3-vowel strings
  4. Each outcome (string) is equally with $P(x_i) = \frac{1}{N}$
  5. So, a string of three vowels has probability $P(A) = \frac{n}{N} = \frac{6^3}{26^3} \approx 0.012$

Rules of Probability

- Theorems:
  1. If A and $\overline{A}$ are complementary events, then $P(\overline{A}) = 1 - P(A)$
  2. $P(\emptyset) = 0$ for any sample space Ω
  3. If $A \subseteq B$, then $P(A) \leq P(B)$
  4. For any event A, $0 \leq P(A) \leq 1$

Addition Rule

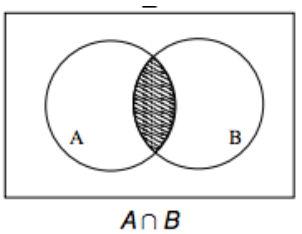
- Axiom 3 allows us to add probabilities of disjoint events
- What about events that are not disjoint?

Addition Rule

- Axiom 3 allows us to add probabilities of disjoint events
- What about events that are not disjoint?

- Theorem: If A and B are events in Ω, then

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



  - $A =$ "has glasses", $B =$ "is blond"
  - $P(A) + P(B)$ counts blondes with glasses twice

Quiz 1

- Assume that the probability of winning in a lottery is 0.01
- What is the probability of not winning?
  1. 0.01
  2. 0.99
  3. Impossible to tell

Quiz 2

- Assume that A and B are events in a sample space Ω
- Which of the following could possibly hold:
  1. $P(A \cup B) < P(A \cap B)$
  2. $P(A \cup B) = P(A \cap B)$
  3. $P(A \cup B) > P(A \cap B)$

Natural Language Processing

Joint, Conditional and Marginal Probability

Joakim Nivre

Uppsala University
Department of Linguistics and Philology
joakim.nivre@lingfil.uu.se

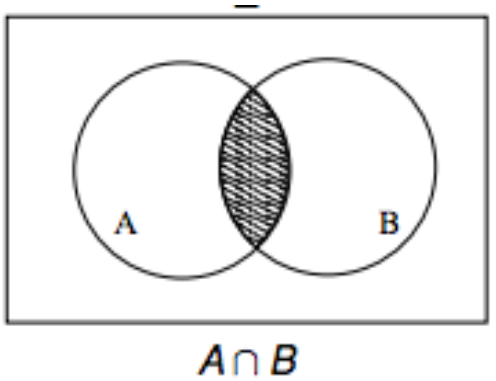
Conditional Probability

- Given events A and B in Ω, with $P(B) > 0$, the conditional probability of A given B is:

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

- $P(A \cap B)$ or $P(A, B)$ is the joint probability of A and B.
  - The prob that a person is rich and famous – joint
  - The prob that a person is rich if they are famous – conditional
  - The prob that a person is famous if they are rich – conditional

Conditional Probability



$$
\begin{aligned}
P(A) &= \text{size of } A \text{ relative to } \Omega \\
P(A, B) &= \text{size of } A \cap B \text{ relative to } \Omega \\
P(A|B) &= \text{size of } A \cap B \text{ relative to } B
\end{aligned}
$$

Example

- We sample word bigrams (pairs) from a large text T
- Sample space and events:
  - $\Omega = \{(w_1, w_2) \in T\}$ = the set of bigrams in T
  - $A = \{(w_1, w_2) \in T \mid w_1 = \text{run}\}$ = bigrams starting with run
  - $B = \{(w_1, w_2) \in T \mid w_2 = \text{amok}\}$ = bigrams ending with amok
- Probabilities:
  - $P(\text{run}_1) = P(A) = 10^{-3}$
  - $P(\text{amok}_2) = P(B) = 10^{-6}$
  - $P(\text{run}_1, \text{amok}_2) = (A, B) = 10^{-7}$
- Probability of amok following run? Of run preceding amok?

Example

- We sample word bigrams (pairs) from a large text T
- Sample space and events:
  - $\Omega = \{(w_1, w_2) \in T\}$ = the set of bigrams in T
  - $A = \{(w_1, w_2) \in T \mid w_1 = \text{run}\}$ = bigrams starting with run
  - $B = \{(w_1, w_2) \in T \mid w_2 = \text{amok}\}$ = bigrams ending with amok
- Probabilities:
  - $P(\text{run}_1) = P(A) = 10^{-3}$
  - $P(\text{amok}_2) = P(B) = 10^{-6}$
  - $P(\text{run}_1, \text{amok}_2) = (A, B) = 10^{-7}$
- Probability of amok following run? Of run preceding amok?
  - $P(\text{run before amok}) = P(A|B) = \frac{10^{-7}}{10^{-6}} = 0.1$
  - $P(\text{amok after run}) = P(B|A) = \frac{10^{-7}}{10^{-3}} = 0.0001$

Multiplication Rule for Joint Probability

- Given events A and B in Ω, with $P(B) > 0$:

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

- Since $A \cap B = B \cap A$, we also have:

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

- The multiplication rule is also known as the chain rule

Quiz 1

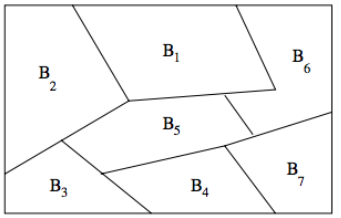
- The probability of winning the Nobel Prize if you have a PhD in Physics is 1 in a million $[P(A|B) = 0.000001]$
- Only 1 in 10,000 people have a PhD in Physics $[P(B) = 0.0001]$
- What is the probability of a person both having a PhD in Physics and winning the Nobel Prize? $[P(A, B) = ?]$
  1. Smaller than 1 in a million
  2. Greater than 1 in a million
  3. Impossible to tell

Marginal Probability

- Marginalization, or the law of total probability
- If events $B_1, \ldots, B_k$ constitute a partition of the sample space Ω (and $P(B_i) > 0$ for all i), then for any event A in Ω:

$$P(A) = \sum_{i=1}^{k} P(A, B_i) = \sum_{i=1}^{k} P(A|B_i)P(B_i)$$

- Partition = pairwise disjoint and $B_1 \cup \cdots \cup B_k = \Omega$



Joint, Marginal and Conditional

- Joint probabilities for rain and wind:

	no wind	some wind	strong wind	storm
no rain	0.1	0.2	0.05	0.01
light rain	0.05	0.1	0.15	0.04
heavy rain	0.05	0.1	0.1	0.05

Joint, Marginal and Conditional

- Joint probabilities for rain and wind:

	no wind	some wind	strong wind	storm
no rain	0.1	0.2	0.05	0.01
light rain	0.05	0.1	0.15	0.04
heavy rain	0.05	0.1	0.1	0.05

- Marginalize to get simple probabilities:
  - $P(\text{no wind}) = 0.1 + 0.05 + 0.05 = 0.2$
  - $P(\text{light rain}) = 0.05 + 0.1 + 0.15 + 0.04 = 0.34$

Joint, Marginal and Conditional

- Joint probabilities for rain and wind:

	no wind	some wind	strong wind	storm
no rain	0.1	0.2	0.05	0.01
light rain	0.05	0.1	0.15	0.04
heavy rain	0.05	0.1	0.1	0.05

- Marginalize to get simple probabilities:
  - $P(\text{no wind}) = 0.1 + 0.05 + 0.05 = 0.2$
  - $P(\text{light rain}) = 0.05 + 0.1 + 0.15 + 0.04 = 0.34$
- Combine to get conditional probabilities:
  - $P(\text{no wind}|\text{light rain}) = \frac{0.05}{0.34} = 0.147$
  - $P(\text{light rain}|\text{no wind}) = \frac{0.05}{0.2} = 0.25$

Bayes Law

- Given events A and B in sample space Ω:

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

  - Follows from definition using chain rule
  - Allows us to "invert" conditional probabilities

- Denominator can be computed using marginalization:

$$P(B) = \sum_{i=1}^{k} P(B, A_i) = \sum_{i=1}^{k} P(B|A_i)P(A_i)$$

  - Special case of partition: $P(A)$, $P(\overline{A})$

Independence

- Two events A and B are independent if and only if:

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

- Equivalently:

$$P(A) = P(A|B)$$
$$P(B) = P(B|A)$$

Independence

- Two events A and B are independent if and only if:

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

- Equivalently:

$$P(A) = P(A|B)$$
$$P(B) = P(B|A)$$

- Example:
  - $P(\text{run}_1) = P(A) = 10^{-3}$
  - $P(\text{amok}_2) = P(B) = 10^{-6}$
  - $P(\text{run}_1, \text{amok}_2) = (A, B) = 10^{-7}$
- A and B are not independent

Quiz 2

- Research has shown that people with disease D exhibit symptom S with 0.9 probability
- A doctor finds that a patient has symptom S
- What can we conclude about the probability that the patient has disease D
  1. The probability is 0.1
  2. The probability is 0.9
  3. Nothing

UPPSALA
UNIVERSITET

Natural Language Processing

Statistical Inference

Joakim Nivre

Uppsala University
Department of Linguistics and Philology
joakim.nivre@lingfil.uu.se

Statistical Inference

- Inference from a finite set of observations (a sample) to a larger set of unobserved instances (a population or model)
- Two main kinds of statistical inference:
  1. Estimation
  2. Hypothesis testing
- In natural language processing:
  - Estimation – learn model parameters (probability distributions)
  - Hypothesis tests – assess statistical significance of test results

Random Variables

- A random variable is a function X that partitions the sample space Ω by mapping outcomes to a value space Ω_X
- The probability function can be extended to variables:

$$P(X = x) = P(\{\omega \in \Omega \mid X(\omega) = x\})$$

- Examples:
  1. The part-of-speech of a word $X : \Omega \rightarrow \{\text{noun}, \text{verb}, \text{adj}, \ldots\}$
  2. The number of words in a sentence $Y : \Omega \rightarrow \{1, 2, 3, \ldots\}$.
- When we are not interested in particular values, we write $P(X)$

Expectation

- The expectation $E[X]$ of a (discrete) numerical variable X is:

$$E[X] = \sum_{x \in \Omega_X} x \cdot P(X = x)$$

- Example: The expectation of the sum Y of two dice:

$$E[Y] = \sum_{y=2}^{12} y \cdot P(Y = y) = \frac{252}{36} = 7$$

Entropy

- The entropy $H[X]$ of a discrete random variable X is:

$$H[X] = E[-\log_2 P(X)] = -\sum_{x \in \Omega_X} P(X = x) \log_2 P(X = x)$$

- Entropy can be seen as the expected amount of information (in bits), or as the difficulty of predicting the variable
  - Sum of two dice: $-\sum_{y=2}^{12} P(Y = y) \log_2 P(Y = y) \approx 3.27$
  - 11-sided die (2–12): $-\sum_{z=2}^{12} \frac{1}{11} \log_2 \frac{1}{11} \approx 3.46$

Quiz 1

- Let X be a random variable that map (English) words to the number of characters they concern
  - For example, $X(\text{run}) = 3$ and $X(\text{amok}) = 4$
- Which of the following statements do you think are true:
  1. $P(X = 0) = 0$
  2. $P(X = 5) < P(X = 50)$
  3. $E[X] < 50$

Statistical Samples

- A random sample of a variable X is a vector $(X_1, \ldots, X_N)$ of independent variables X_i with the same distribution as X
  - It is said to be i.i.d. = independent and identically distributed
  - In practice, it is often hard to guarantee this
  - Observations may not be independent (not i.)
  - Distribution may be biased (not i.d.)
- What is the intended population?
  - A Harry Potter novel is a good sample of J.K. Rowling, or fantasy fiction, but not of scientific prose
  - This is relevant for domain adaptation in NLP

Estimation

- Given a random sample of X, we can define sample variables, such as the sample mean:

$$\overline{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$$

- Sample variables can be used to estimate model parameters (population variables)
  1. Point estimation: use variable X to estimate parameter ϕ
  2. Interval estimation: use variables X_{min} and X_{max} to construct an interval such that $P(X_{\text{min}} < \phi < X_{\text{max}}) = p$, where p is the confidence level adopted

Maximum Likelihood Estimation (MLE)

- Likelihood of parameters θ given sample $x_1, \ldots, x_N$:

$$\mathcal{L}(\theta|x_1, \ldots, x_N) = P(x_1, \ldots, x_N|\theta) = \prod_{i=1}^{N} P(x_i|\theta)$$

- Maximum likelihood estimation – choose θ to maximize $\mathcal{L}$:

$$\max_{\theta} \mathcal{L}(\theta|x_1, \ldots, x_N)$$

- Basic idea:
  - A good sample should have a high probability of occurring
  - Thus, choose the estimate that maximizes sample probability

Examples

- Sample mean is an MLE of expectation:

$$\hat{E}[X] = \overline{X}$$

  - For example, estimate expected sentence length in a certain type of text by mean sentence length in a representative sample

- Relative frequency is an MLE of probability:

$$\hat{P}(X = x) = \frac{f(x)}{N}$$

  - For example, estimate the probability of a word being a noun by the relative frequency of nouns in a suitable corpus

MLE for Different Distributions

- Joint distribution of X and Y:

$$\hat{P}_{\text{MLE}}(X = x, Y = y) = \frac{f(x, y)}{N}$$

- Marginal distribution of X:

$$\hat{P}_{\text{MLE}}(X = x) = \sum_{y \in \Omega_Y} \hat{P}_{\text{MLE}}(X = x, Y = y)$$

- Conditional distribution of X given Y:

$$\begin{aligned} \hat{P}_{\text{MLE}}(X = x | Y = y) &= \frac{\hat{P}_{\text{MLE}}(X{=}x, Y{=}y)}{\hat{P}_{\text{MLE}}(Y{=}y)} \\ &= \frac{\hat{P}_{\text{MLE}}(X{=}x, Y{=}y)}{\sum_{x \in \Omega_X} \hat{P}_{\text{MLE}}(X{=}x, Y{=}y)} \end{aligned}$$

Quiz 2

- Consider the following sample of English words:

$$\{\textit{once}, \textit{upon}, \textit{a}, \textit{time}, \textit{there}, \textit{was}, \textit{a}, \textit{frog}\}$$

- What is the MLE of word length (number of characters) based on this sample?
  1. 4
  2. 8
  3. 3.25