

Probability

PA154 Language Modeling (1.2)

Pavel Rychlý

pary@fi.muni.cz

February 16, 2023

Source: Introduction to Natural Language Processing (600.465)
Jan Hajič, CS Dept., Johns Hopkins Univ.
www.cs.jhu.edu/~hajic

Experiments & Sample Spaces

- Experiment, process, test, ...
- Set of possible basic outcomes: sample space Ω základní prostor obsahující možné výsledky
 - coin toss ($\Omega = \{\text{head, tail}\}$), die ($\Omega = \{1..6\}$)
 - yes/no opinion poll, quality test (bad/good) ($\Omega = \{0,1\}$)
 - lottery ($|\Omega| \cong 10^7 .. 10^{12}$)
 - # of traffic accidents somewhere per year ($\Omega = \mathbb{N}$)
 - spelling errors ($\Omega = Z^*$), where Z is an alphabet, and Z^* is set of possible strings over such alphabet
 - missing word ($|\Omega| \cong$ vocabulary size)

Events

- Event (jev) A is a set of basic outcomes
- Usually $A \subset \Omega$, and all $A \in 2^\Omega$ (the event space, jevové pole)
 - Ω is the certain event (jistý jev), \emptyset is the impossible event (nemožný jev)
- Example:
 - experiment: three times coin toss
 - $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$
 - count cases with exactly two tails: then
 - $A = \{HTT, THT, TTH\}$
 - all heads:
 - $A = \{HHH\}$

Probability

- Repeat experiment many times, record how many times a given event A occurred (“count” c_1).
- Do this whole series many times; remember all c_i s.
- Observation: if repeated really many times, the ratios of $\frac{c_i}{T_i}$ (where T_i is the number of experiments run in the i -th series) are close to some (unknown but) **constant** value.
- Call this constant a **probability of A**. Notation: **p(A)**

Estimating Probability

- Remember: ...close to an *unknown* constant.
- We can only estimate it:
 - from a single series (typical case, as mostly the outcome of a series is given to us we cannot repeat the experiment):

$$p(A) = \frac{c_1}{T_1}$$

- otherwise, take the weighted average of all $\frac{c_i}{T_i}$ (or, if the data allows, simply look at the set of series as if it is a single long series).
- This is the **best** estimate.

Example

- Recall our example:
 - experiment: three times coin toss
 - $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$
 - count cases with exactly two tails: $A = \{HTT, THT, TTH\}$
- Run an experiment 1000 times (i.e. 3000 tosses)
- Counted: 386 cases with two tails (**HTT, THT or TTH**)
- estimate: $p(A) = 386/1000 = .386$
- Run again: 373, 399, 382, 355, 372, 406, 359
 - $p(A) = .379$ (weighted average) or simply 3032/8000
- *Uniform* distribution assumption: $p(A) = 3/8 = .375$

Basic Properties

- Basic properties:
 - $p: 2^\Omega \rightarrow [0, 1]$
 - $p(\Omega) = 1$
 - Disjoint events: $p(\cup A_i) = \sum_i p(A_i)$
- NB: axiomatic definition of probability: take the above three conditions as axioms
- Immediate consequences:
 - $P(\emptyset) = 0$
 - $p(\bar{A}) = 1 - p(A)$
 - $A \subseteq B \Rightarrow p(A) \leq P(B)$
 - $\sum_{a \in \Omega} p(a) = 1$

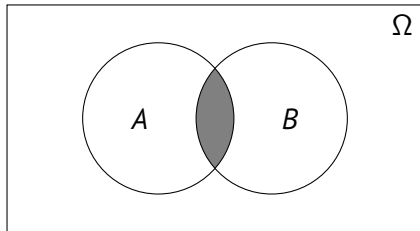
Joint and Conditional Probability

- $p(A, B) = p(A \cap B)$

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

- Estimating from counts:

- $p(A|B) = \frac{p(A, B)}{p(B)} = \frac{\frac{c(A \cap B)}{T}}{\frac{c(B)}{T}} = \frac{c(A \cap B)}{c(B)}$

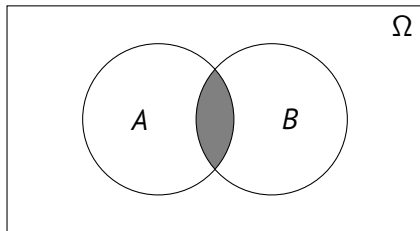


Bayes Rule

- $p(A,B) = p(B,A)$ since $p(A \cap B) = p(B \cap A)$
 - therefore $p(A|B)p(B) = p(B|A)p(A)$, and therefore:

Bayes Rule

$$p(A|B) = \frac{p(B|A) \times p(A)}{p(B)}$$



Independence

- Can we compute $p(A,B)$ from $p(A)$ and $p(B)$?
- Recall from previous foil:

$$p(A|B) = \frac{p(B|A) \times p(A)}{p(B)}$$

$$p(A|B) \times p(B) = p(B|A) \times p(A)$$

$$p(A, B) = p(B|A) \times p(A)$$

...we're almost there: how $p(B|A)$ relates to $p(B)$?

- $p(B|A) = p(B)$ iff A and B are **independent**
- Example: two coin tosses, weather today and weather on March 4th 1789;
- Any two events for which $p(B|A) = P(B)$!

Chain Rule

$$\begin{aligned} p(A_1, A_2, A_3, A_4, \dots, A_n) = & \\ p(A_1|A_2, A_3, A_4, \dots, A_n) \times & p(A_2|A_3, A_4, \dots, A_n) \times \\ \times p(A_3|A_4, \dots, A_n) \times \dots \times & p(A_{n-1}|A_n) \times p(A_n) \end{aligned}$$

- this is a direct consequence of the Bayes rule.

The Golden Rule of Classic Statistical NLP

- Interested in an event A given B (where it is not easy or practical or desirable) to estimate $p(A|B)$:
- take Bayes rule, max over all Bs:
- $\operatorname{argmax}_A p(A|B) = \operatorname{argmax}_A \frac{p(B|A) \times p(A)}{p(B)} =$
 $\boxed{\operatorname{argmax}_A (p(B|A) \times p(A))}$
- ...as $p(B)$ is constant when changing As

Random Variables

- is a function $X : \Omega \rightarrow Q$
 - in general $Q = R^n$, typically R
 - easier to handle real numbers than real-world events
- random variable is *discrete* if Q is countable (i.e. also if finite)
- Example: *die*: natural “numbering” $[1,6]$, *coin*: $\{0,1\}$
- Probability distribution:
 - $p_X(x) = p(X = x) =_{df} p(A_x)$ where $A_x = \{a \in \Omega : X(a) = x\}$
 - often just $p(x)$ if it is clear from context what X is

Expectation

Joint and Conditional Distributions

- is a mean of a random variable (weighted average)
 - $E(X) = \sum_{x \in X(\Omega)} x \cdot p_X(x)$
- Example: one six-sided die: 3.5, two dice (sum): 7
- Joint and Conditional distribution rules:
 - analogous to probability of events
- Bayes: $p_{X|Y}(x, y) =_{notation} p_{XY}(x|y) =_{even\ simpler\ notation}$

$$p(x|y) = \frac{p(y|x) \cdot p(x)}{p(y)}$$

- Chain rule: $p(w, x, y, z) = p(z) \cdot p(y|z) \cdot p(x|y, z) \cdot p(w|x, y, z)$

Standard Distributions

- Binomial (discrete)
 - outcome: 0 or 1 (thus *binomial*)
 - make n trials
 - interested in the (probability of) numbers of successes r
- Must be careful: it's not uniform!
- $p_b(r|n) = \frac{\binom{n}{r}}{2^n}$ (for equally likely outcome)
- $\binom{n}{r}$ counts how many possibilities there are for choosing r objects out of n ;
- $\binom{n}{r} = \frac{n!}{(n-r)!r!}$

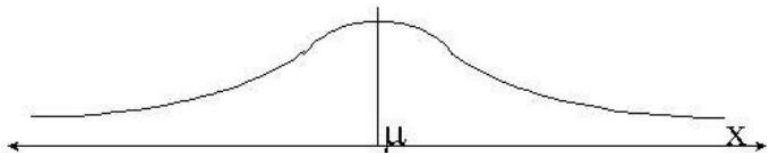
Continuous Distributions

- The normal distribution (“Gaussian”)

- $p_{norm}(x|\mu, \sigma) = \exp \left[\frac{-(x - \mu)^2}{2\sigma^2} \right] \frac{1}{\sigma\sqrt{2\pi}}$

- where:

- μ is the mean (x-coordinate of the peak) (0)
- σ is the standard deviation (1)



- other: hyperbolic, t