NLP: Probability

Dan Garrette
dhg@cs.utexas.edu

September 10, 2013

1 Basics

- $\mathcal{E} \neq \emptyset$: event space (sample space)
- We will be dealing with sets of discrete events.

Example 1: Coin

- Trial: flipping a coin
- Two possible outcomes: heads or tails, $\mathcal{E} = \{\text{H, T}\}$
- $p(H)$ is the probability of heads
- if $p(H) = 0.8$, we would expect that flipping 100 times would yield 80 heads

Example 2: 3 coins

- Trial: flipping three coins
- Still two possible outcomes: heads or tails
- e.g. first=H, second=T, third=T (HTT)
- $\mathcal{E} = \{\text{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT}\}$ (2^3)
- event: set of results
  - e.g. two tails and one head (A = HTT, THT, TTH)

Example 3: Die

- Trial: rolling a die
- outcomes: 1, 2, 3, 4, 5, or 6
- $\mathcal{E} = \{1, 2, 3, 4, 5, 6\}$

- event: set of results
  - e.g. 1 or 2 (1,2)
  - e.g. even (2,4,6)
  - $2^6 = 64$ distinct events

2 Probability functions

- 0 (impossible) to 1 (certain)
- p distributions probability mass 1 over the sample space
- $p(X)$ for $X \subseteq \mathcal{E}$: function mapping sets of events to $[0, 1]$, the probability X
  - How likely is an event to occur
  - e.g. Fair coin: $p(A) = \frac{|A|}{|\mathcal{E}|}$
  - e.g. Die: Event B = divisible by 3: $P(B) = P(\{3, 6\}) = \frac{2}{6} = \frac{1}{3}$
- $p(\mathcal{E}) = 1$, $p(\emptyset) = 0$

Properties

- If A and B are disjoint events (sets of outcomes), i.e. $A \cap B = \emptyset$, then $p(A \cup B) = p(A) + p(B)$
  - e.g. A=roll a 3, B=roll a 6: p(3 OR 6) = p($\{3,6\}$) = p($\{3\} \cup \{6\}$) = p($\{3\}$) + p($\{6\}$)
  - e.g. A=raining, B=snowing: p(raining OR snowing) = p(raining) + p(snowing)
- $p(A \cup B) = p(A) + p(B) - p(A \cap B)$
  - notice that this subsumes the previous because we assumed $A \cap B = \emptyset$, so $p(A \cap B) = 0$

3 Conditional Probability

- probability of event A given event B: $p(A \mid B)$
- prior probabiliy of A: $p(A)$
- posterior probability of A given B: $p(A \mid B)$
- $p(A \mid B) = \frac{p(A \cap B)}{p(B)}$
  - $p(4 \mid \text{even}) = \frac{|\{4\} \cap \{2,4,6\}|}{|\{2,4,6\}|} = \frac{|\{4\}|}{|\{2,4,6\}|} = \frac{1}{3}$
  - $p(\text{even} \mid 4) = \frac{|\{2,4,6\} \cap \{4\}|}{|\{4\}|} = \frac{|\{4\}|}{|\{4\}|} = \frac{1}{1}$
  - B becomes the sample space
  - $\sum_x p(x \mid B) = 1$

The chain rule

- $p(A \cap B) = p(A \mid B) \cdot p(B) = p(B \mid A) \cdot p(A)$
- $p(A_1 \cap ... \cap A_n) = p(A_1) \cdot p(A_2 \mid A_1) \cdot p(A_3 \mid A_1, A_2) \cdot ... \cdot p(A_n \mid \bigcap_{i=1}^{n-1} A_i)$

4 Independence

- Outcomes of two events do not affect each other
- $p(A) = p(A \mid B)$
- $p(A \cap B) = p(A) \cdot p(B)$
- $p(H \mid HHHHH) = \frac{1}{2}$

5 Random Variable Notation

- $p(X = a)$
- X is a random variable. It selects an event from the sample space
- e.g. coin flip: $p(X = \text{H}) = \frac{1}{2}$ and $p(X = \text{T}) = \frac{1}{2}$
- e.g. tennis
  - $p(tennis = \text{yes})$
  - $p(tennis = \text{yes} \mid outlook = \text{rain})$
  - $p(tennis = \text{yes} \mid outlook = \text{sunny})$

6 Joint Probability

- Probability of both x and y happening: $p(x, y) = p(X = x, Y = y)$
- Probability of x happening is the sum of probabilities across all ys: $p_X(x) = \sum_y p(x, y)$
- If X and Y are independent, then $p(x, y) = p_X(x) \cdot p_Y(y)$
  - Rolling two sixes: $p(X = 6, Y = 6) = p(X = 6) \cdot p(Y = 6) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$

7 Maximum Likelihood Estimate (MLE)

Coin

- Data: H H T H T
- How do we determine probabilities from the data? What should $p(\text{H})$ be? And $p(\text{T})$?

- Choose $p(\text{H})$ and $p(\text{T})$ to maximize the probability of the data
- Remember that flips are independent: $p(data) = p(\text{H,H,T,H,T}) = p(\text{H}) \cdot p(\text{H}) \cdot p(\text{T}) \cdot p(\text{H}) \cdot p(\text{T})$
- Want maximum $p(data)$
  - $p(\text{H}) = 1.0$ ($p(\text{T}) = 0.0$): $p(data) = 1.0 \cdot 1.0 \cdot 0.0 \cdot 1.0 \cdot 0.0 = 0.0$
  - $p(\text{H}) = 0.8$ ($p(\text{T}) = 0.2$): $p(data) = 0.8 \cdot 0.8 \cdot 0.2 \cdot 0.8 \cdot 0.2 = 0.02048$
  - $p(\text{H}) = 0.6$ ($p(\text{T}) = 0.4$): $p(data) = 0.6 \cdot 0.6 \cdot 0.4 \cdot 0.6 \cdot 0.4 = \mathbf{0.03456}$
  - $p(\text{H}) = 0.5$ ($p(\text{T}) = 0.5$): $p(data) = 0.5 \cdot 0.5 \cdot 0.5 \cdot 0.5 \cdot 0.5 = 0.03125$
- $p(data) = p(\text{H})^3 \cdot p(\text{T})^2$
- We can get the MLE by taking counts from the data:
  - $p(\text{H}) = \frac{C(H)}{C(H)+C(T)} = \frac{3}{5}$, $p(\text{T}) = \frac{C(T)}{C(H)+C(T)} = \frac{2}{5}$

Named Entity Recognition

- Data:
  endsWith=ia,location
  endsWith=er,person
  endsWith=ia,person
  endsWith=er,location
  endsWith=ia,location
  endsWith=nd,location
  endsWith=ia,location
- $p(type = \text{location} \mid endswith = \text{ia}) = \frac{C(type=location \text{ AND } endswith=ia)}{C(endswith=ia)} = \frac{3}{4} = 0.75$
- $p(type = \text{person} \mid endswith = \text{ia}) = \frac{C(type=person \text{ AND } endswith=ia)}{C(endswith=ia)} = \frac{1}{4} = 0.25$
- $p(type = \text{location} \mid endswith = \text{nd}) = \frac{C(type=location \text{ AND } endswith=nd)}{C(endswith=nd)} = \frac{1}{1} = 1.0$
- $p(type = \text{person} \mid endswith = \text{nd}) = \frac{C(type=person \text{ AND } endswith=nd)}{C(endswith=nd)} = \frac{0}{1} = 0.0$

8 Bayes Theorem

- $p(A \mid B) = \frac{p(B|A) \cdot p(A)}{p(B)}$
- proof:
  $p(X \mid Y) = \frac{p(X \cap Y)}{p(Y)}$,
  so $p(A \cap B) = p(A \mid B) \cdot p(B) = p(B \mid A) \cdot p(A)$,
  so $p(A \mid B) = \frac{p(B|A) \cdot p(A)}{p(B)}$
- $p(A \mid B)$: the posterior, what we are trying to figure out
- $p(A)$: the "prior", useful for encoding prior knowledge about the likelihood of A

- $p(B \mid A)$: the likelihood of the evidence
- Example:
  - Setup:
    1% of women age forty who are screened have breast cancer.
    80% of women with breast cancer will get positive mammographies.
    9.6% of women get false positive mammographies.
  - If a woman has a positive test result, what is the probability that she has breast cancer?
  - 85% of doctors get this wrong. They usually say 80%.
  - Solution:

$$\begin{aligned} p(cancer) &= 0.01 & p(\overline{cancer}) &= 0.99 \\ p(pos \mid cancer) &= 0.8 & p(pos \mid \overline{cancer}) &= 0.096 \\ \frac{p(pos \cap cancer)}{p(cancer)} &= 0.8 & \frac{p(pos \cap \overline{cancer})}{p(\overline{cancer})} &= 0.096 \\ p(pos \cap cancer) &= 0.8 \cdot 0.01 & p(pos \cap \overline{cancer}) &= 0.096 \cdot 0.99 \end{aligned}$$

$$\begin{aligned} p(pos) &= p(pos \mid cancer) + p(pos \mid \overline{cancer}) \\ &= p(pos \mid cancer) \cdot p(cancer) + p(pos \mid \overline{cancer}) \cdot p(\overline{cancer}) \\ &= 0.8 \cdot 0.01 + 0.096 \cdot 0.99 = 0.10304 \\ p(cancer \mid pos) &= \frac{p(pos \mid cancer) \cdot p(cancer)}{p(pos)} \\ &= \frac{0.8 \cdot 0.01}{0.10304} \approx 7.76\% \end{aligned}$$

  - But 7.76% is much less than 80%
  - So 92.24% of positive results are false alarms, meaning that a huge number of women undergo unnecessary procedures. Since all procedures themselves carry risks, many women are put at risk unncessarily, which is why many health organizations are now against blanket screenings.