CS4705

Probability Review and
Naïve Bayes

Slides from Dragomir Radev

Classification using a Generative Approach

- Previously on *NLP*…
  - discriminative models $P(C|D)$
  - “here is a line with all the social media posts on one side and the scientific articles on the other side; which side is this example on?”
- Now…
  - generative models $P(C, D)$
  - “here are some characteristics of social media posts, and here are some characteristics of scientific articles; which is this example more like?”

Classification using a Generative Approach

- We’ll look in detail at the Naïve Bayes classifier and Maximum Likelihood Expectation

- But we need some background in probability first…

Probabilities in NLP

- Speech recognition:
  - “recognize speech” vs “wreck a nice beach”
- Machine translation:
  - “l’avocat general”: “the attorney general” vs. “the general avocado”
- Information retrieval:
  - If a document includes three occurrences of “stir” and one of “rice”, what is the probability that it is a recipe?

- Probabilities make it possible to combine evidence from multiple sources systematically

Probability Theory

- Random experiment (trial): an experiment with uncertain outcome
  - e.g., flipping a coin, picking a word from text
- Sample space: the set of all possible outcomes for an experiment
  - e.g., flipping 2 fair coins, $\Omega = \{HH, HT, TH, TT\}$
- Event: a subset of the sample space, $E \subseteq \Omega$
  - E happens iff the outcome is in E, e.g.,
    - $E = \{HH\}$ (all heads)
    - $E = \{HH, TT\}$ (same face)

Events

- Probability of Event : $0 \leq P(E) \leq 1$, s.t.
  - $P(A \cup B) = P(A) + P(B)$, if $(A \cap B) = \emptyset$
  - e.g., A=same face, B=different face
- $\emptyset$ is the impossible event (empty set)
  - $P(\emptyset) = 0$
- Ω is the certain event (entire sample space)
  - $P(\Omega) = 1$

Example: Roll a Die

- Sample space: $\Omega = \{1,2,3,4,5,6\}$

- Fair die: $P(1) = P(2) = \cdots = P(6) = 1/6$
- Unfair die: $P(1) = 0.3, \; P(2) = 0.2, \ldots$

- N-dimensional die: $\Omega = \{1, 2, 3, 4, \ldots, N\}$

- Example in modeling text:
  - Roll a die to decide which word to write in the next position
  - $\Omega = \{cat, \; dog, \; tiger, \ldots\}$

Example: Flip a Coin

- Sample space: $\Omega = \{Heads,\ Tails\}$

- Fair coin: $P(H) = 0.5, P(T) = 0.5$
- Unfair coin: $P(H) = 0.3, P(T) = 0.7$

- Flipping three coins:
  - $\Omega = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$

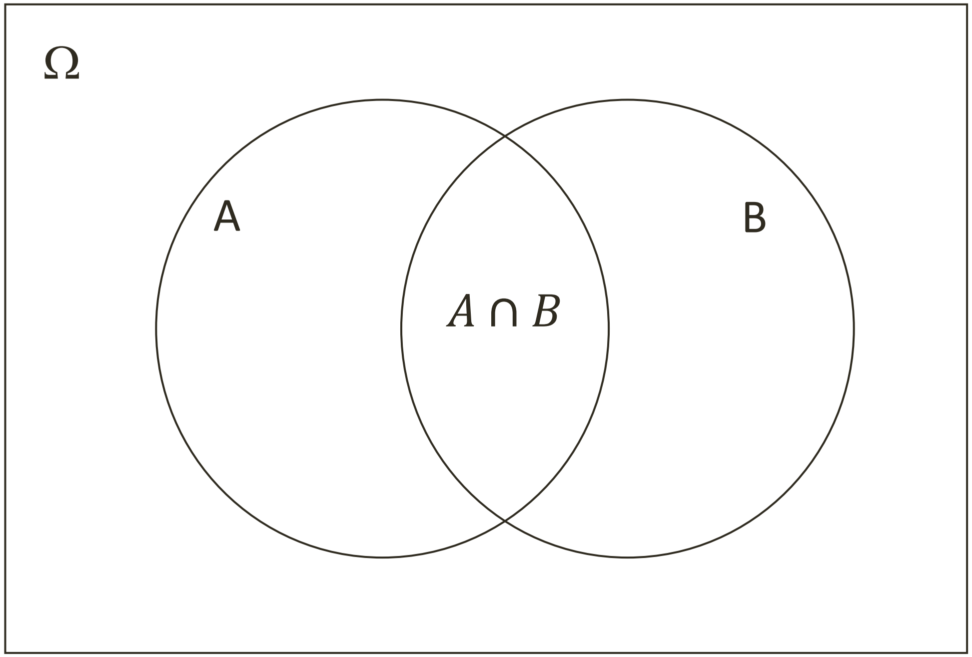
- Example in modeling text:
  - Flip a coin to decide whether or not to include a word in a document
  - Sample space = {appear, absence}

Probabilities

- Probability distribution
  - a function that distributes a probability mass of 1 throughout the sample space Ω
  - $0 \leq P(\omega) \leq 1$ for each outcome $\omega \in \Omega$
  - $\sum_{\omega \in \Omega} P(\omega) = 1$
- Probability of an event E
  - $P(E) = \sum_{\omega \in E} P(\omega)$
- Example: a fair coin is flipped three times
  - What is the probability of 3 heads?
  - What is the probability of 2 heads?

Probabilities

- Joint probability: $P(A \cap B)$
  - also written as $P(A, B)$
- Conditional probability: $P(A|B) = \frac{P(A \cap B)}{P(B)}$



Conditional Probability

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

- $P(A \cap B) = P(A)P(B|A) = P(B)P(A|B)$

- So, $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$ (Bayes' Rule)

- For independent events, $P(A \cap B) = P(A)P(B)$, so $P(A|B) = P(A)$

Conditional Probability

- Six-sided fair die
  - $P(D\ even) =?$
    - $1/2$
  - $P(D \geq 4) =?$
    - $1/2$
  - $P(D\ even | D \geq 4) =?$
    - $\frac{2/6}{1/2} = 2/3$
  - $P(D\ odd | D \geq 4) =?$
    - $\frac{1/6}{1/2} = 1/3$
- Multiple conditions: $P(D\ odd | D \geq 4,\ \ D \leq 5) =?$
    - $\frac{1/6}{2/6} = 1/2$

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

Independence

- *What are some examples of independent events?*

- *What about dependent events?*

Response

Naïve Bayes Classifier

- We use Baye's rule: $P(C|D) = \frac{P(D|C)P(C)}{P(D)}$
  - Here $C = Class,\ D = Document$
- We can simplify and ignore $P(D)$ since it is independent of class choice

$$P(C|D) \cong \boxed{P(D|C)P(C) \cong P(C)\prod_{i=1,n} P(w_i|C)}$$

  - This estimates the probability of D being in class C assuming that D has n tokens and w is a token in D.

But Wait…

- What is D?
  - $D = w_1 w_2 w_3 \ldots w_n$
- So what is $P(D|C)$, really?
  - $P(D|C) = P(w_1 w_2 w_3 \ldots w_n|C)$
    - But $w_1 w_2 w_3 \ldots w_n$ is not in our training set so we don't know its probability
    - How can we simplify this?

Conditional Probability Revisited

- Recall the definition of conditional probability
  1. $P(A|B) = \frac{P(AB)}{P(B)}$, or equivalently
  2. $P(AB) = P(A|B)P(B)$

- What if we have more than two events?
  - $P(ABC \ldots N) = P(A|BC \ldots N) * P(B|C \ldots N) * \cdots * P(M|N) * P(N)$
  - This is the chain rule for probability
  - We can prove this rule by induction on N

Independence Assumption

- So what is P(D|C)?
  - $= P(w_1 w_2 w_3 \dots w_n | C) = P(w_1 | w_2 w_3 \dots w_n C) * P(w_2 | w_3 \dots w_n C) * \cdots * P(w_n | C) * P(C)$
- This is still not very helpful…

- We make the "naïve" assumption that all words occur independently of each other
  - Recall that for independent events w_1 and w_2 we have $P(w_1 | w_2) = P(w_1)$
  - That's this step! $P(D|C) \cong \prod_{i=1,n} P(w_i | C)$

Independence Assumptions

- *Is the Naïve Bayes assumption a safe assumption?*
- *What are some examples of words that might be dependent on other words?*

Response

Using Labeled Training Data

- $P(C|D) \cong P(C) \prod_{i=1,n} P(w_i|C)$
  - $P(C) = \frac{D_c}{D}$
    - the number of training documents with label C
    - divided by the total number of training documents
  - $P(w_i|C) = \frac{Count(w_i C)}{\sum_{v_i \in V} Count(v_i C)}$
    - the number of times word w_i occurs with label C
    - divided by the number of times all words in the vocabulary V occur with label C

- This is the maximum-likelihood estimate (MLE)

Using Labeled Training Data

- *Can you think of ways to improve this model?*

- *Some issues to consider…*
  - *What if there are words that do not appear in the training set?*
  - *What if the plural of a word never appears in the training set?*
  - *How are extremely common words (e.g., “the”, “a”) handled?*

Response

A Quick Note on the MLE…

- The counts seem intuitively right, but how do we know for sure?
- We are trying to find values of $P(C)$ and $P(w_i|C)$ that maximize the likelihood of the training set
- i.e. we want the largest possible value of $P(T) = \prod_{t \in T}\left[P(c_t) \prod_{w_i \in t} P(w_i|c_t)\right]$
    - Here T is the training set and t is a training example
- We can find these values by taking the log, then taking the derivative, then solving for 0

Questions?

Reading for next time

- C 5.1 – 5.5, Speech and Language