Natural Language Processing CSCI 4152/6509 — Lecture 10 Introduction to Probabilistic NLP

Instructors: Vlado Keselj
Time and date: 16:05 – 17:25, 5-Oct-2023
Location: Rowe 1011

Previous Lectures

- Discussion about evaluation methods for classifiers
- Similarity-based Text Classification
- CNG classification method
- Edit distance:
  - introduction, properties, dynamic programming approach, example, algorithm

Edit Distance Example (to finish)

- distance between ‘there’ and ‘ythre’

Part III: Probabilistic Approach to NLP

Logical versus Plausible Reasoning

- As a part of AI (Artificial Intelligence), NLP follows two main approaches to *computer reasoning*, or *computer inference:*

1. logical reasoning
   - known also as classical, symbolic, knowledge-based AI
   - *monotonic:* once conclusion drawn, never retracted
   - *certain:* conclusions certain, given assumptions
2. plausible reasoning
   - examples: probabilistic, fuzzy logic, neural networks
   - *non-monotonic*
   - *uncertain*

Plausible Reasoning

- How to combine ambiguous, incomplete, and contradicting evidence to draw reasonable conclusions?
- Typical approach: make plausible inference of some hidden structure from observations
- Examples:

Observations (input)		Hidden Structure (output)
symptoms	$\rightarrow$	illness
pixel matrix	$\rightarrow$	object, relations
speech signal	$\rightarrow$	phonemes, words
word sequence	$\rightarrow$	meaning
sentence	$\rightarrow$	parse tree
word sequence	$\rightarrow$	POS tags, names, entities
words in e-mail Subject:	$\rightarrow$	Is message spam? Yes/No
text	$\rightarrow$	text category (class)

Probabilistic NLP as a Plausible Reasoning Approach

- Regular expressions and finite automata are example of logical or knowledge-based approach to NLP
- Plausible approaches to NLP:

   1. Probabilistic: use of Theory of Probability, also known as stochastic or statistical NLP
      - Alternative plausible approaches, examples:
   2. neural networks,
   3. kernel methods,
   4. fuzzy logic, fuzzy sets,
   5. Dempster-Shafer theory
   6. rough sets,
   7. default logic, ...

Review of Basics of Probability Theory

- You should have this background from previous courses; this is just a review,
  - discussed a bit in the textbook: [JM] 5.5, and [MS] 2.1
- Simple event or basic outcome
  - e.g., rolling a die, choosing a letter
- *Event space:* the set of all outcomes, usually denoted Ω
- *Event or outcome* is a set of simple events or basic outcomes
- In other words event is any subset of Ω; i.e., $A \subseteq \Omega$
- Each event is associated with a probability, which is a number between 0 and 1, inclusive: $0 \leq \mathrm{P}(A) \leq 1$

Probability Examples

- $\mathrm{P}(\text{“rolling a 6 with a die”}) = 1/6$
- Choosing a letter of English alphabet:
  - If we choose uniformly: $\mathrm{P}(\text{‘a’}) = 1/26 \approx 0.04$
  - Choosing from a text: $\mathrm{P}(\text{‘a’}) \approx 0.08$
  - Remember our output from “Tom Sawyer”:

```
35697 0.1204 e
28897 0.0974 t
23528 0.0793 a
23264 0.0784 o
20200 0.0681 n
...
```

Probability Axioms

- **(Nonnegativity)** $\mathrm{P}(A) \geq 0$, for any event A
- **(Additivity)** for disjoint events A and B, i.e., if $A, B \subset \Omega$ and $A \cap B = \emptyset$, then
  $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$
  or, more generally,
  $\mathrm{P}(A_1 \cup A_2 \cup \ldots) = \mathrm{P}(A_1) + \mathrm{P}(A_2) + \ldots$
- **(Normalization)** $\mathrm{P}(\Omega) = 1$, where Ω is the entire sample space.
- Some consequences of the above axioms are:
  $\mathrm{P}(\emptyset) = 0$ and $\mathrm{P}(\Omega - A) = 1 - \mathrm{P}(A)$

Independent and Dependent Events

- Independent events A and B (definition):
  $\mathrm{P}(A, B) = \mathrm{P}(A) \cdot \mathrm{P}(B)$
- Use of comma in: $\mathrm{P}(A, B) = \mathrm{P}(A \cap B)$
- Example: choosing two letters in text
  1. Choosing independently:
     $\mathrm{P}(\text{`t'}) = 0.1, \mathrm{P}(\text{`h'}) = 0.07, \mathrm{P}(\text{`t'}, \text{`h'}) = 0.007$
  2. Choosing two consecutive letters (dependent events):
     $\mathrm{P}(\text{`t'}, \text{`h'}) = 0.04$

Conditional Probability

- Conditional probability

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

- Expressing independency using conditional probability

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

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

  This is an alternative definition of independent events.

Annotation with More Events

- There is a bit of flexibility in using notation; e.g.,
- $\mathrm{P}(A, B, C) = \mathrm{P}(A \cap B \cap C)$
- $\mathrm{P}(A|B, C) = \mathrm{P}(A|B \cap C)$
- $\mathrm{P}(A, B, C|D, E, F) = \mathrm{P}(A \cap B \cap C|D \cap E \cap F)$
- and so on.
- Three independent events: $\mathrm{P}(A, B, C) = \mathrm{P}(A)\mathrm{P}(B)\mathrm{P}(C)$
- Conditionally independent events

$$\mathrm{P}(A, B|C) = \mathrm{P}(A|C) \cdot \mathrm{P}(B|C)$$

Bayes' Theorem

- Bayes' theorem (one form):

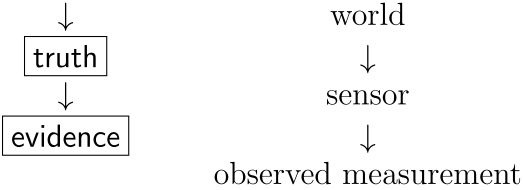
$$\mathrm{P}(A|B) = \frac{\mathrm{P}(B|A) \cdot \mathrm{P}(A)}{\mathrm{P}(B)}$$

- The second form is based on breaking the set B into disjoint sets $B = A_1 \cup A_2 \cup \ldots$:

$$\mathrm{P}(A_i|B) = \frac{\mathrm{P}(B|A_i) \cdot \mathrm{P}(A_i)}{\mathrm{P}(B)} = \frac{\mathrm{P}(B|A_i) \cdot \mathrm{P}(A_i)}{\sum_i \mathrm{P}(B|A_i)\mathrm{P}(A_i)}$$

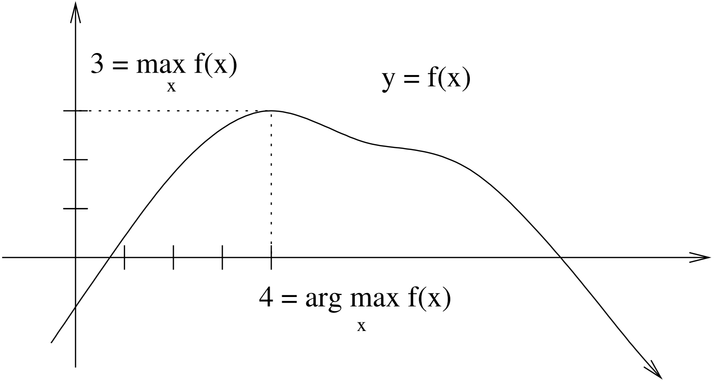
Bayesian Inference and Generative Models

- We will use Bayesian Inference on Generative Models
- Generative Models, also known as Forward Generative Models
- One way of representing knowledge with a probabilistic model



Notation Remark: max and argmax

- $\max$ is the maximum value of a function
- $\arg\max$ is an argument value for which function achieves the maximum



Bayesian Inference: Using Bayes' Theorem

- Bayesian inference is a principle of combining evidence

$$
\begin{aligned}
\text{conclusion} &= \mathop{\text{arg max}}_{\text{possible truth}} P(\text{possible truth}|\text{evidence}) \\
&= \mathop{\text{arg max}}_{\text{possible truth}} \frac{P(\text{evidence}|\text{possible truth})P(\text{possible truth})}{P(\text{evidence})} \\
&= \mathop{\text{arg max}}_{\text{possible truth}} P(\text{evidence}|\text{possible truth})P(\text{possible truth})
\end{aligned}
$$

- application to speech recognition: acoustic model and language model

Bayesian Inference: Speech Recognition Example

- evidence → sound
- possible truth → utterance (words spoken)
- our best guess about utterance → utterance*

$$
\begin{aligned}
\text{utterance}^* &= \operatorname*{arg\,max}_{\text{all utterances}} P(\text{utterance}|\text{sound}) \\
&= \operatorname*{arg\,max}_{\text{all utterances}} \frac{P(\text{sound}|\text{utterance})P(\text{utterance})}{P(\text{sound})} \\
&= \operatorname*{arg\,max}_{\text{utterance}} P(\text{sound}|\text{utterance})P(\text{utterance})
\end{aligned}
$$