Faculty of Computer Science, Dalhousie University *5-Oct-2023*

CSCI 4152/6509 — Natural Language Processing

Lecture 10: Introduction to Probabilistic NLP

Location: Rowe 1011 Instructor: Vlado Keselj
Time: 16:05 – 17:25

Previous Lectures

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

---

Part III

Probabilistic Approach to NLP

11 Introduction to Probabilistic Approach to NLP

11.1 Logical versus Plausible Reasoning

Artificial Intelligence (AI) is an area of Computer Science related to the task of developing software and computers that exhibit some form of intelligent behaviour. NLP can be considered to be a sub-area of AI. Solving problems in AI generally involves implementing in a computer some form of reasoning or inference. The automated inference methods can be divided into two large groups: *logical reasoning*, and *plausible reasoning.* As we will see, this division can be adopted for the NLP area as well.

1. Logical reasoning is known also as classical, symbolic, or knowledge-based AI. The rule-based AI is also a form of logical reasoning. This form of automated inference is based on the principles that we can find in mathematical logic. We start with basic set of premises, known as axioms, and using a fixed set of rules, we derive new conclusions. This type of reasoning is called *monotonic* since with more evidence, i.e., more known facts, we can produce more conclusions. A consequence of this is that once we make conclusions, they cannot be retracted or canceled. In other words, if we conclude that something is true, we cannot conclude that it is wrong based on the new evidence because that would mean that we either had some wrong evidence in the first place, or we made a mistake in reasoning. We also describe this reasoning as *certain*, since the model is designed in such way that we can derive only certain (or definite, guaranteed) conclusions.

2. Plausible reasoning is an automated reasoning in which we typically derive plausible conclusions; i.e., conclusions that may be true, and we usually have some way of rating the truthfulness of such conclusions. There are different approaches to this kind of automated reasoning. The most widely used are the approaches based on the mathematical probability theory, where we try to make a mathematically sound estimate of the probability that our conclusions are true. Other approaches include fuzzy logic and neural networks. Unlike logical reasoning, plausible reasoning is *uncertain*, since the conclusions are not guaranteed to be correct. Logical reasoning is also

non-monotonic since based on new evidence, we can withdraw some conclusions, and with more evidence we may end up with less conclusions. Plausible reasoning allows for contradicting evidence.

Plausible Reasoning

- How to combine ambiguous, incomplete, and contradicting evidence to draw reasonable conclusions?
- Frequently approached as the task of making plausible inference of some hidden structure from observations.
- examples:

Input (observations)		Hidden Structure
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, ...

11.2 Review of Basics of Probability Theory

This is a brief and intuitive review of some basic notions from the theory of probability.

- 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

  The event space is also known as the sample space. For example, the event space for rolling a die has six elements, corresponding to the six different outcomes; or an event space for choosing a letter of the English alphabet has 26 elements if we ignore letter case.
- *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
...
```

As we saw from the "Tom Sawyer" example, the probability of the letter 'a' in a typical English text is about 0.08.

The probability can be seen as a function that maps a set of possible experiment outcomes to a real number between 0 and 1, including 0 and 1, i.e. to a number from the interval $[0, 1]$. We always have in mind certain space of outcomes Ω and we assume that in each experiment (trial, instance, or model configuration), one of those outcomes will happen. The probability that one of the outcomes from a set A ($A \subset \Omega$) will happen, is denoted $\mathrm{P}(A)$. A set of outcomes A is called an event.

Probability Axioms

- Probability axioms: nonnegativity, additivity, normalization:

  The basic properties of the probability, known as the probability axioms, are:
  - **(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

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B).$$

    More generally, for a possibly infinite sequence of disjoint events A_1, A_2, ...,

$$\mathrm{P}(A_1 \cup A_2 \cup \ldots) = \mathrm{P}(A_1) + \mathrm{P}(A_2) + \ldots$$

  - **(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: $\mathrm{P}(\text{'t'}) = 0.1, \mathrm{P}(\text{'h'}) = 0.07, \mathrm{P}(\text{'t', 'h'}) = 0.007$
  2. Choosing two consecutive letters (dependent events): $\mathrm{P}(\text{'t', 'h'}) = 0.04$

Conditional Probability

- Conditional probability

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

  The probability $\mathrm{P}(A|B)$ (it is read "probability of A given B") is the probability of event A happening given that the event B happens.

– 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):

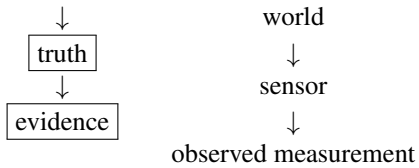
$$\mathrm{P}(A|B) = \frac{\mathrm{P}(B|A) \cdot \mathrm{P}(A)}{\mathrm{P}(B)}$$

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

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

11.3 Bayesian Inference and Generative Models

In our further discussion, we will use *Bayesian Inference* applied to the *Generative Models.* These models are called the generative models because they describe how a particular event that we are studying is generated; i.e., what were the causal relations that were involved in producing certain obeservations. We can use the following visual representation to describe this model in general:



We assume that there is certain true structure or information that caused some events, which can be observed through some sensors. We will simply refer to this true information as *truth* and information that is available to us as *observable evidence,* or simply as *evidence.*

Notation Remark: 'max' and 'arg max' Operators. Before presenting the main use of the Bayes theorem in Bayesian inference, let us first clarify the use of 'max' and 'arg max' operators in formulae.

The operator 'max' is applied to an expression depending on an index variable x (denoted $\max_x$) and it returns the maximum value of the expression over all possible values of the variable x. On the other hand, the operator 'arg max' gives one value of the index variable x (denoted $\arg\max_x$) for which the expression achieves the maximum value. Figure 3 illustrates graphically these operators for a function $y = f(x)$.

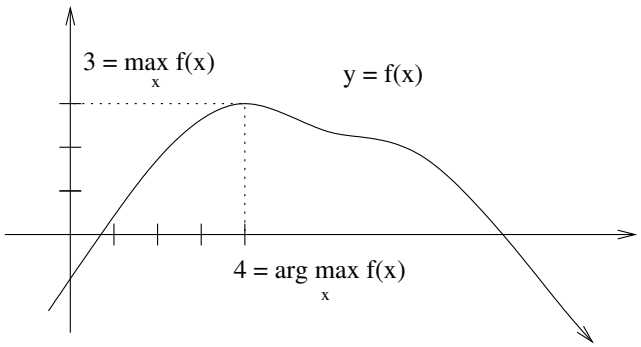


Figure 3: Illustrating operators $\max$ and arg max

Bayesian Inference: Using Bayes' Theorem

– Bayesian inference is a principle of combining evidence

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

– application to speech recognition: acoustic model and language model

The above equations describe Bayesian inference. To understand steps in the equations, let us go over them step by step:

$$\text{conclusion} = \underset{\text{possible truth}}{\arg\max}\ P(\text{possible truth}|\text{evidence})$$

The above equation states that we reach the conclusion about the best possible truth by finding the maximal probability of that 'truth' given the evidence, according to our model.

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

In the above equation, we directly apply the Bayes' theorem.

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

It is important to note that the reason why the above equation holds is because the denominator $\mathrm{P}(\text{evidence})$ does not depend on 'possible truth'; i.e., it is constant for different values of 'possible truth'. A frequent mistake is to assume that the equation is true because the $\mathrm{P}(\text{possible truth})$ is 1.

If you are not clear why 'evidence' does not depend on different 'possible truths', it will be more clear once we express this in terms of random variables.

Bayesian Inference: Speech Recognition Example

Let us look at the speech recognition as an example of the problem that can be addressed using Bayesian Inference on a generative model.

Slide notes:

Bayesian Inference: Speech Recognition Example

- evidence $\rightarrow$ sound
- possible truth $\rightarrow$ utterance (words spoken)
- our best guess about utterance $\rightarrow$ utterance*

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

At the end, we need to estimate two probabilities to recognize the utterance base on the sound: $\mathrm{P}(\text{sound}|\text{utterance})$ which is estimated using an *acoustic model*, and $\mathrm{P}(\text{utterance})$ which is estimated using a *language model.*