# Natural Language Processing CSCI 4152/6509 — Lecture 9 P0 Topics Discussion; Introduction to Probabilistic Modeling

Instructors: Vlado Keselj
Time and date: 16:05 – 17:25, 7-Oct-2023
Location: Carleton Tupper Building Theatre C

# Previous Lecture

- CNG classification method
- Edit distance:
  - introduction, properties, dynamic programming approach, example, algorithm

# P0 Topics Discussion

- Discussion of individual projects as proposed in P0 submissions
- Projects discussed: P-01, P-03, P-04, P-05, P-06

# 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 $\Omega$
- *Event or outcome* is a set of simple events or basic outcomes
- In other words event is any subset of $\Omega$; 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

- P(“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 $\Omega$ 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

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

- Expressing independency using conditional probability

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

$$P(A|B) = P(A)$$

  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):

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

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

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

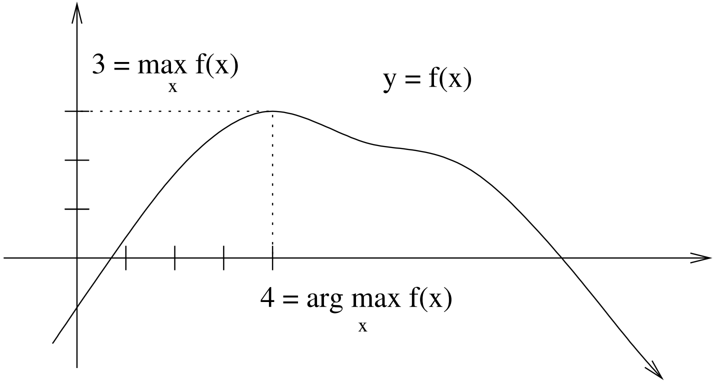
# Bayesian Inference and Generative Models

- 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

| | |
|---|---|
| ↓ | world |
| truth | ↓ |
| ↓ | sensor |
| evidence | ↓ |
| | observed measurement |

# 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

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

- 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*

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

# Probabilistic Modeling

- How do we create and use a probabilistic model?
- Model elements:
  - Random variables
  - Model configuration (Random configuration)
  - Variable dependencies
  - Model parameters
- Computational tasks

# Random Variables

- Random variable $V$, defining an event as $V = x$ for some value $x$ from a domain of values $D$; i.e., $x \in D$
- $V = x$ is usually not a **basic** event due to having more variables
- An event with two random variables: $V_1 = x_1, V_2 = x_2$
- Multiple random variables: $\mathbf{V} = (V_1, V_2, ..., V_n)$

# Model Configuration (Random Configuration)

- **Full Configuration:** If a model has $n$ random variables, then a Full Model Configuration is an assignment of all the variables:

$$V_1 = x_1, V_2 = x_2, \ldots, V_n = x_n$$

- **Partial configuration:** only some variables are assigned, e.g.:

$$V_1 = x_1, V_2 = x_2, \ldots, V_k = x_k \quad (k < n)$$

# Probabilistic Modeling in NLP

Probabilistic Modeling in NLP is a general framework for modeling NLP problems using random variables, random configurations, and an effective ways to reason about probabilities of these configurations.

# Variable Independence and Dependence

- Random variables $V_1$ and $V_2$ are *independent* if $\mathrm{P}(V_1\!=\!x_1, V_2\!=\!x_2) = \mathrm{P}(V_1\!=\!x_1)\mathrm{P}(V_2\!=\!x_2)$ for all $x_1, x_2$
- or expressed in a different way: $\mathrm{P}(V_1\!=\!x_1|V_2\!=\!x_2) = \mathrm{P}(V_1\!=\!x_1)$ for all $x_1, x_2, x_3$.
- Random variables $V_1$ and $V_2$ are *conditionally independent given* $V_3$ if, for all $x_1, x_2, x_3$: $\mathrm{P}(V_1\!=\!x_1, V_2\!=\!x_2|V_3\!=\!x_3) =$ $\mathrm{P}(V_1\!=\!x_1|V_3\!=\!x_3)\mathrm{P}(V_2\!=\!x_2|V_3\!=\!x_3)$
- or $\mathrm{P}(V_1\!=\!x_1|V_2\!=\!x_2, V_3\!=\!x_3) = \mathrm{P}(V_1\!=\!x_1|V_3\!=\!x_3)$

# Computational Tasks in Probabilistic Modeling

1. Evaluation: compute probability of a complete configuration
2. Simulation: generate random configurations
3. Inference: has the following sub-tasks:
   - 3.a Marginalization: computing probability of a partial configuration,
   - 3.b Conditioning: computing conditional probability of a completion given an observation,
   - 3.c Completion: finding the most probable completion, given an observation
4. Learning: learning parameters of a model from data.

# Illustrative Example: Spam Detection

- the problem of spam detection
- a probabilistic model for spam detection; random variables:

$$\begin{aligned} \mathit{Caps} &= \text{'Y' if the message subject line does not contain lowercase letter, 'N' otherwise,} \\ \mathit{Free} &= \text{'Y' if the word 'free' appears in the message subject line (letter case is ignored), 'N' otherwise, and} \\ \mathit{Spam} &= \text{'Y' if the message is spam, and 'N' otherwise.} \end{aligned}$$

- one random configuration represents one e-mail message

# Random Sample

- Data based on sample of 100 email messages

| *Free* | *Caps* | *Spam* | Number of messages |
|---|---|---|---|
| Y | Y | Y | 20 |
| Y | Y | N | 1 |
| Y | N | Y | 5 |
| Y | N | N | 0 |
| N | Y | Y | 20 |
| N | Y | N | 3 |
| N | N | Y | 2 |
| N | N | N | 49 |
| | | Total: | 100 |

What are examples of computational tasks in this example?