

## CSCI 4152/6509 — Natural Language Processing

7-Oct-2009

### Lecture 12: Probabilistic Modeling

Room: FASS 2176  
Time: 11:35 – 12:25

#### Previous Lecture

- Aside: Introduction to IR on-line book <http://nlp.stanford.edu/IR-book/html/htmledition/irbook.html>
- Evaluation methods for classification (cont.): n-fold cross-validation,
- Parser evaluation:
  - PARSEVAL measures, labeled and unlabeled precision and recall, F-measure;
- Text clustering:
  - task definition, the simple k-means method, hierarchical clustering, divisive and agglomerative clustering;
- Evaluation of clustering:
  - inter-cluster similarity, cluster purity, use of entropy or information gain;
- CNG — Common N-Grams classification method

#### 6.6 CNG—Common N-Gram analysis for text classification

- Method based on character n-grams
- Language independent
- Based on creating n-gram based author profiles
- kNN method ( $k$  Nearest Neighbours)
- similarity measure:

$$\sum_{g \in D_1 \cup D_2} \left( \frac{f_1(g) - f_2(g)}{\frac{f_1(g) + f_2(g)}{2}} \right)^2 = \sum_{g \in D_1 \cup D_2} \left( \frac{2 \cdot (f_1(g) - f_2(g))}{f_1(g) + f_2(g)} \right)^2 \quad (2)$$

where  $f_i(g) = 0$  if  $g \notin D_i$ .

### 7 Elements of Probability Theory

- simple event

Examples: rolling a dice, choosing a letter

$P('a') = ?$ ; probability of choosing a letter;  $1/26 \approx 0.04$ . However, in a typical English text, it is about 0.08.

This is a brief and intuitive review of some basic notions from the theory of probability. 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  $\Omega$  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  $P(A)$ . A set of outcomes  $A$  is called an event.

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

- **(Nonnegativity)**  $P(A) \geq 0$ , for any event  $A$

- **(Additivity)** for disjoint events  $A$  and  $B$ , i.e., if  $A, B \subset \Omega$  and  $A \cap B = \emptyset$ , then

$$P(A \cup B) = P(A) + P(B).$$

More generally, for a possibly infinite sequence of disjoint events  $A_1, A_2, \dots$ ,

$$P(A_1 \cup A_2 \cup \dots) = P(A_1) + P(A_2) + \dots$$

- **(Normalization)**  $P(\Omega) = 1$ , where  $\Omega$  is the entire sample space.
- some consequences of the above axioms are:  $P(\emptyset) = 0$  and  $P(\Omega - A) = 1 - P(A)$
- independent events (definition):  $P(A, B) = P(A) \cdot P(B)$
- use of comma in:  $P(A, B) = P(A \cap B)$   
Example: choosing two letters vs. choosing two consecutive letters  
choosing t: 0.1, h:0.07; choosing 't' and 'h' independently: 0.007; consecutive 'th' = 0.04; not independent events
- random variables, independent random variables  
Two random variables  $V_1$  and  $V_2$  are independent if:

$$P(V_1 = x_1, V_2 = x_2) = P(V_1 = x_1) \cdot P(V_2 = x_2)$$

- conditional probability

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

- expressing independency using conditional probability  
Two events  $A$  and  $B$  are independent if and only if:

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

This is an alternative definition of independent events.

- Bayes' theorem (one form):

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

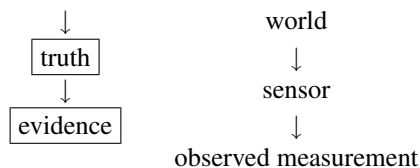
- Conditionally independent variables

Random variables  $V_1$  and  $V_2$  are *conditionally independent given*  $V_3$  if  $P(V_1 = x_1, V_2 = x_2 | V_3 = x_3) = P(V_1 = x_1 | V_3 = x_3)P(V_2 = x_2 | V_3 = x_3)$  for all  $x_1, x_2, x_3$ . Equivalently, if  $P(V_1 = x_1 | V_2 = x_2, V_3 = x_3) = P(V_1 = x_1 | V_3 = x_3)$  for all  $x_1, x_2, x_3$ .

## 8 Probabilistic Modelling

### 8.1 Generative Models

- also known as Forward generative model
- Bayesian inference
- one way of representing knowledge with a probabilistic model



**Bayesian Inference**

- principle of evidence combination: Bayesian inference

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

- application to speech recognition: acoustic model and language model