Text Processing

Lecture 6:

Artificial Intelligence

Principles of

CS 530:
Perception: Two basic problems:

- Multiple levels of structure
  - Multi-pass or blackboard
  - Reduce dimensionality of data
- Noise in sensor data
  - Discriminatory functions
  - Bayes rule
  - I.e., summary parameters
Text Classification

- Search a collection of documents for those "about", e.g., food commodities like soybeans and wheat
- Problem: how do we tell if a document is about a subject?
- "crop predictions have been pessimistic" vs "if you crop a dog's tail" vs "producers of distilled spirits" vs "in the spirit of détente" vs "in the spirit of détente"
- "crop predictions have been pessimistic" vs "if you crop a dog's tail" vs "producers of distilled spirits" vs "in the spirit of détente" vs "in the spirit of détente"
Text Classification

- Heuristic: View document as a set of words
- Define class of interest ("about food commodities") by an example set of words
- Search for "similar" documents
- Boolean variables, one per word
  - \( x_1: \) apples: true or false

\[
\begin{align*}
\vec{x} = & x_1 \ldots x_n \\
\end{align*}
\]
Boolean vector space
Also a set of booleans for classes of documents:

- \( \neg w = w_1 \) means document about food commodities, \( w = w_2 \) means not about food commodities.

\( w \neq w_1 \) means document about food commodities, \( \neg w = w_2 \) means not about food commodities.
Empirical Data

- Apriori probabilities of classes
  \[ P(w=w_1) = P(w_1) \]

- Conditional probabilities of variable values
  \[ P(x_i=t | w_1 = w_1^t) = \frac{P(x_i=t, w_1 = w_1^t)}{P(w_1 = w_1^t)} \]

- Conditional probabilities of variable values of classes

- Apriori probabilities of classes
Assumption of independence

- Can use Bayes rule to get $P(w | x_1, \ldots, x_n)$

$$P(w | x_1, \ldots, x_n) = \frac{P(x_1, w) \cdot \cdots \cdot P(x_n, w)}{P(x_1, \ldots, x_n)}$$

If $x_i$ conditionally independent given $w$, then

$$P(x_1, \ldots, x_n | w) = P(x_1 | w) \cdot \cdots \cdot P(x_n | w)$$

Assumption of independence
$P(C|x) > P(\sim C|x)$

What we really need to know is if

What is $P(1.8)$? Don’t really care

$P(r=1.8 | \text{circle}) \cdot P(\text{ball}) / P(1.8) = P(\text{ball}) = P(\text{ball})$ * $P(C|x) \cdot P(x|C) / P(x)$

Bayes Rule