Example 3: Document Classification

Document Modelling

Bag-of-words: Describe a document as a D-dimensional binary vector $\mathbf{x}$, indicating the presence/absence of a word in a vocabulary $\mathcal{V}$.

Example: consider the following tiny vocabulary:

$$\mathcal{V} = \{\text{football, defence, strategy, goal, office}\}$$

Then, a sentence "Adam from UIC Registrar's Office scored two goals in a community football game." is represented as

$$\mathbf{x} = (1, 0, 0, 1, 1),$$

since it contains only the words "football", "office", and "goal"

- We do not care about the order of the words
- We do not care about the words that are not in the vocabulary

Example 3: Document Classification:

Binary Classification

We want to classify documents as being about sports ($\mathcal{C}_1$) or politics ($\mathcal{C}_2$).

A simple *model* for $p(\mathbf{x}|\mathcal{C}_j)$ is:

$$p(\mathbf{x}|\mathcal{C}_j) = \prod_{i=1}^{D} p(x_i|\mathcal{C}_j)$$

This is called Naive Bayes due to its unrealistic assumption of conditional independence of words given the class label

Example 3: Document Classification:

Conditional Probability Tables

Assume the vocabulary:

$$\mathcal{V} = \{\text{football, defence, strategy, goal, office}\}$$

and the conditional probability tables (CPTs) are given by:

$$\begin{array}{ll}
p(\mathcal{C}_1) = 0.5 & p(\mathcal{C}_2) = 0.5 \\
p(f = 1|\mathcal{C}_1) = 0.8 & p(f = 1|\mathcal{C}_2) = 0.1 \\
p(d = 1|\mathcal{C}_1) = 0.7 & p(d = 1|\mathcal{C}_2) = 0.7 \\
p(s = 1|\mathcal{C}_1) = 0.2 & p(s = 1|\mathcal{C}_2) = 0.8 \\
p(g = 1|\mathcal{C}_1) = 0.7 & p(g = 1|\mathcal{C}_2) = 0.3 \\
p(o = 1|\mathcal{C}_1) = 0.2 & p(o = 1|\mathcal{C}_2) = 0.7
\end{array}$$

A new document arrives and is described by $\mathbf{x} = (0, 1, 1, 1, 0)$.

What is the probability of this document being about sports?

Example 3: Document Classification:

Solution

$$
\begin{aligned}
p(\mathcal{C}_1|\mathbf{x}) &= \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1) + p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)} \\
&= \frac{\prod_{d=1}^{D} p(x_d|\mathcal{C}_1) \cdot p(\mathcal{C}_1)}{\prod_{d=1}^{D} p(x_d|\mathcal{C}_1) \cdot p(\mathcal{C}_1) + \prod_{d=1}^{D} p(x_d|\mathcal{C}_2) \cdot p(\mathcal{C}_2)} \\
&= \frac{\prod_{d=1}^{D} p(x_d|\mathcal{C}_1)}{\prod_{d=1}^{D} p(x_d|\mathcal{C}_1) + \prod_{d=1}^{D} p(x_d|\mathcal{C}_2)}
\end{aligned}
$$

Example 3: Document Classification:

Solution

$$
\begin{aligned}
p(\mathcal{C}_1|\mathbf{x}) &= \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1) + p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)} \\
&= \frac{\prod_{d=1}^{D} p(x_d|\mathcal{C}_1) \cdot p(\mathcal{C}_1)}{\prod_{d=1}^{D} p(x_d|\mathcal{C}_1) \cdot p(\mathcal{C}_1) + \prod_{d=1}^{D} p(x_d|\mathcal{C}_2) \cdot p(\mathcal{C}_2)} \\
&= \frac{\prod_{d=1}^{D} p(x_d|\mathcal{C}_1)}{\prod_{d=1}^{D} p(x_d|\mathcal{C}_1) + \prod_{d=1}^{D} p(x_d|\mathcal{C}_2)} \\
&= \frac{(0.2)(0.7)(0.2)(0.7)(0.8)}{(0.2)(0.7)(0.2)(0.7)(0.8) + (0.9)(0.7)(0.8)(0.3)(0.3)} \\
&\approx 0.26.
\end{aligned}
$$

We would classify this document as politics as $p(\mathcal{C}_2|\mathbf{x}) \approx 0.74$.