Foundations of Data Science

DS 3001

Data Science Program

Department of Computer Science

Worcester Polytechnic Institute

Instructor: Prof. Kyumin Lee

Upcoming Schedule

- HW3
  - https://canvas.wpi.edu/courses/18106/assignments/133696
  - Due date: May 1st

- Project Checkpoint
  - https://canvas.wpi.edu/courses/18106/assignments/133778
  - Due date: May 1st

Clustering Evaluation

Evaluation

- How to evaluate clustering?
  - Internal:
    - Tightness and separation of clusters (e.g. k-means objective)
    - Fit of probabilistic model to data
  - External
    - Compare to known class labels on benchmark data

External criterion: Purity

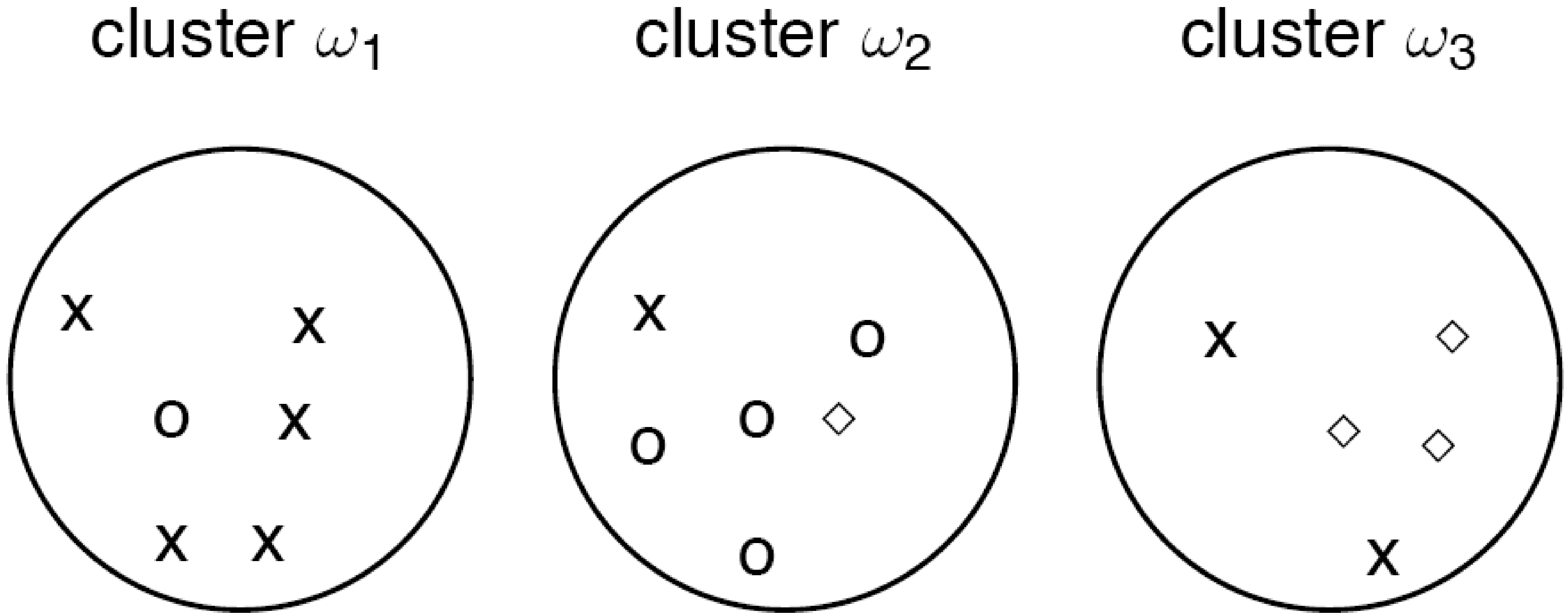
$$\text{purity}(\Omega, \Gamma) = \frac{1}{N} \sum_k \max_j |\omega_k \cap c_j|$$

$\Omega = \{\omega_1, \omega_2, \ldots, \omega_K\}$ is the set of clusters and
$\Gamma = \{c_1, c_2, \ldots, c_J\}$ is the set of classes.

For each cluster ω_k : find class c_j with most members n_{kj} in cluster

Sum all n_{kj} and divide by total number of points

Example



$$\text{good\_docs}(\omega_1) = \max(5, 1, 0) = 5$$
$$\text{good\_docs}(\omega_2) = \max(1, 4, 1) = 4$$
$$\text{good\_docs}(\omega_3) = \max(2, 0, 3) = 3$$
$$\text{purity}(\Omega) = 1/17 \cdot (5 + 4 + 3) = 12/17$$

Naive Bayes Classifier
(Text Classifier)

The Naive Bayes Classifier

- The Naive Bayes classifier is a probabilistic classifier
- We compute the probability of a document *d* being in a class *c* as follows:

Prior

$$\underbrace{P(c|d)}_{\text{Posterior}} \propto \overbrace{P(c)}^{\text{Prior}} \prod_{1 \le k \le n_d} P(t_k|c)$$

- $P(c)$ is the prior probability of *c*.
- n_d is the length of the document. (number of tokens)
- $P(t_k|c)$ is the conditional probability of term t_k occurring in a document of class *c*
- $P(t_k|c)$ as a measure of how much evidence t_k contributes that *c* is the correct class.
- If a document's terms do not provide clear evidence for one class vs. another, we choose the *c* with highest *P(c)* probability.

Maximum a posteriori class

- Our goal in Naive Bayes classification is to find the “best” class.
- The best class is the most likely or maximum a posteriori (MAP) class c_{map}:

$$c_{\text{map}} = \arg\max_{c \in \mathbb{C}} \hat{P}(c|d) = \arg\max_{c \in \mathbb{C}} \hat{P}(c) \prod_{1 \leq k \leq n_d} \hat{P}(t_k|c)$$

Parameter estimation take 1: Maximum likelihood

- Estimate parameters $\hat{P}(c)$ and $\hat{P}(t_k|c)$ from train data: How?
- Prior:

$$\hat{P}(c) = \frac{N_c}{N}$$

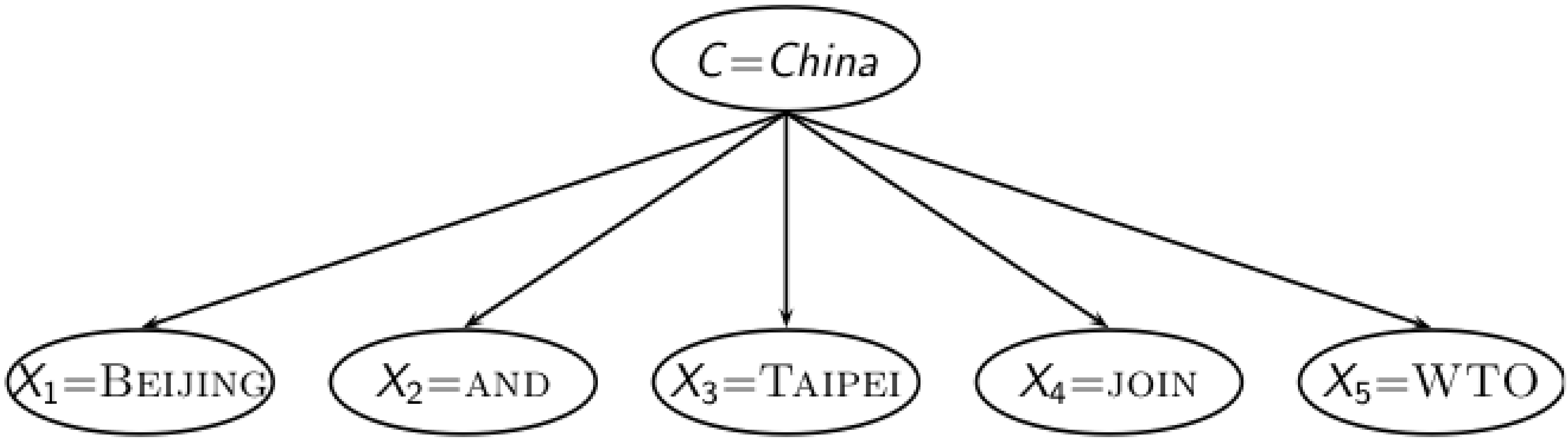
- N_c : number of docs in class *c*; *N*: total number of docs
- Conditional probabilities:

$$\hat{P}(t|c) = \frac{T_{ct}}{\sum_{t' \in V} T_{ct'}}$$

- T_{ct} is the number of tokens of *t* in training documents from class *c* (includes multiple occurrences)
- We've made a Naive Bayes independence assumption here:

$$\hat{P}(t_{k_1}|c) = \hat{P}(t_{k_2}|c)$$

The problem with maximum likelihood estimates: Zeros



$$P(China|d) \propto P(China) \cdot P(\texttt{BEIJING}|China) \cdot P(\texttt{AND}|China) \cdot P(\texttt{TAIPEI}|China) \cdot P(\texttt{JOIN}|China) \cdot P(\mathrm{WTO}|China)$$

$$\hat{P}(\mathrm{WTO}|China) = \frac{T_{China,\mathrm{WTO}}}{\sum_{t' \in V} T_{China,t'}} = \frac{0}{\sum_{t' \in V} T_{China,t'}} = 0$$

The problem with maximum likelihood estimates: Zeros (cont)

- If there were no occurrences of WTO in documents in class China, we'd get a zero estimate:

$$\hat{P}(\mathrm{WTO}|China) = \frac{T_{China,\mathrm{WTO}}}{\sum_{t' \in V} T_{China,t'}} = 0$$

- → We will get P(China|d) = 0 for any document that contains WTO!
- Zero probabilities cannot be conditioned away.

To avoid zeros: Add-one smoothing

- Before:

$$\hat{P}(t|c) = \frac{T_{ct}}{\sum_{t' \in V} T_{ct'}}$$

- Now: Add one to each count to avoid zeros:

$$\hat{P}(t|c) = \frac{T_{ct} + 1}{\sum_{t' \in V}(T_{ct'} + 1)} = \frac{T_{ct} + 1}{(\sum_{t' \in V} T_{ct'}) + B}$$

- B is the number of different words (in this case the size of the vocabulary: $|V| = M$)

To avoid zeros: Add-one smoothing

- Estimate parameters from the training corpus using add-one smoothing
- For a new document, for each class, compute sum of (i) log of prior and (ii) logs of conditional probabilities of the terms
- Assign the document to the class with the largest score

Example

	docID	words in document	in $c = China$?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Japan Chinese Chinese Chinese Tokyo	?

- What do we need?
  - Class priors: P(c), P(not c)
  - Conditional probabilities: P(t|c), P(t|not c)

Example

	docID	words in document	in c = *China*?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Japan Chinese Chinese Chinese Tokyo	?

$$c_{\text{map}} = \arg\max_{c \in \mathbb{C}} \hat{P}(c|d) = \arg\max_{c \in \mathbb{C}} \hat{P}(c) \prod_{1 \leq k \leq n_d} \hat{P}(t_k|c)$$

$$c_{\text{map}} = \arg\max_{c \in \mathbb{C}} [\log \hat{P}(c) + \sum_{1 \leq k \leq n_d} \log \hat{P}(t_k|c)]$$

To avoid underflow,
Apply logarithm into the formula

$$\hat{P}(c) = \boxed{} \text{ and } \hat{P}(\overline{c}) = \boxed{}$$

$$\begin{aligned}
\hat{P}(\text{Chinese}|c) &= \\
\hat{P}(\text{Tokyo}|c) = \hat{P}(\text{Japan}|c) &= \\
\hat{P}(\text{Chinese}|\overline{c}) &= \\
\hat{P}(\text{Tokyo}|\overline{c}) = \hat{P}(\text{Japan}|\overline{c}) &=
\end{aligned}$$

Example

	docID	words in document	in c = *China*?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Japan Chinese Chinese Chinese Tokyo	?

$$c_{\text{map}} = \arg\max_{c \in \mathbb{C}} [\log \hat{P}(c) + \sum_{1 \leq k \leq n_d} \log \hat{P}(t_k|c)]$$

$$\hat{P}(c) = 3/4 \text{ and } \hat{P}(\overline{c}) = 1/4$$

$$\begin{aligned}
\hat{P}(\text{Chinese}|c) &= (5+1)/(8+6) = 6/14 = 3/7 \\
\hat{P}(\text{Tokyo}|c) = \hat{P}(\text{Japan}|c) &= (0+1)/(8+6) = 1/14 \\
\hat{P}(\text{Chinese}|\overline{c}) &= (1+1)/(3+6) = 2/9 \\
\hat{P}(\text{Tokyo}|\overline{c}) = \hat{P}(\text{Japan}|\overline{c}) &= (1+1)/(3+6) = 2/9
\end{aligned}$$

Example

	docID	words in document	in c = *China*?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Japan Chinese Chinese Chinese Tokyo	?

$$c_{\text{map}} = \arg\max_{c \in \mathbb{C}} [\log \hat{P}(c) + \sum_{1 \leq k \leq n_d} \log \hat{P}(t_k|c)]$$

$$\hat{P}(c|d) \propto$$ log3/4 + log1/14 + 3log3/7 + log1/14 ≈ -3.52

$$\hat{P}(\overline{c}|d) \propto$$ log1/4 + log2/9 + 3log2/9 + log2/9 ≈ -3.86

Thus, the classifier assigns the test document to c = *China*.

PageRank

(Measure a relative score of a web page based on importance and authority by evaluating the quality and quantity of its links)

https://www.youtube.com/watch?v=Quk88piD8PM

https://www.youtube.com/watch?v=LVV_93mBfSU