Lecture 23: Semi Supervised Learning

Madhavan Mukund

https://www.cmi.ac.in/~madhavan

Data Mining and Machine Learning
August–December 2020

Semi-supervised learning

- Supervised learning requires labelled training data
- What if we don't have enough labelled data?
- For a probabilistic classifier we can apply EM
  - Use available training data to assign initial probabilities
  - Label the rest of the data using this model — fractional labels
  - Add up counts and re-estimate the parameters

Semi-supervised topic classification

- Each document is a multiset or bag of words over a vocabulary $V = \{w_1, w_2, \ldots, w_m\}$
- Each topic c has probability $Pr(c)$
- Each word $w_i \in V$ has conditional probability $Pr(w_i \mid c_j)$, for $c_j \in C$
  - Note that $\sum_{i=1}^{m} Pr(w_i \mid c_j) = 1$
- Assume document length is independent of the class
- Only a small subset of documents is labelled
  - Use this subset for initial estimate of $Pr(c)$, $Pr(w_i \mid c_j)$

Semi-supervised topic classification

- Current model $Pr(c)$, $Pr(w_i \mid c_j)$
- Compute $Pr(c_j \mid d)$ for each unlabelled document d
  - Normally we assign the maximum among these as the class for d
  - Here we keep fractional values
- Recompute $Pr(c_j) = \dfrac{\sum_{d \in D} Pr(c_j \mid D)}{|D|}$
  - For labelled d, $Pr(c_j \mid d) \in \{0, 1\}$
  - For unlabelled d, $Pr(c_j \mid d)$ is fractional value computed from current parameters
- Recompute $Pr(w_i \mid c_j)$ — fraction of occurrences of w_i in documents labelled c_j
  - n_{id} — occurrences of w_i in d
  - $Pr(w_i \mid c_j) = \dfrac{\sum_{d \in D} n_{id}\ Pr(c_j \mid d)}{\sum_{t=1}^{m} \sum_{d \in D} n_{td}\ Pr(c_j \mid d)}$

Summary

- Use EM to bootstrap model from limited labelled data
- Compute initial parameters from labelled data
- Extrapolate (fractional) labels to remaining data
- Recompute the parameters and iterate
- Fractional classification is important to apply EM