This note explains essential stuff to understand a Naive Bayes classifier trained with both labeled data and unlabeled data [2, 3].

1 List of Important Terminologies

1. \( Q \) funtion of an EM algorithm is either of the following:

\[
\sum_{d \in D} \sum_{c \in C} \sum_{c \in C} P(c|d; \theta) \log P(c, d; \theta)
\]

\[
\sum_{d \in D} E_{P(c|d; \theta)}[\log P(c, d; \theta)]
\]

2. The objective function of a Naive Bayes classifier when conducting a maximum a posteriori estimation

\[
\log P(\theta) + \log P(D)
\]

3. Let \( q_{w,c} \) be a probability that a word \( w \) is chosen given class \( c \) i.e.

\[
q_{w,c} = P(W = w | C = c)
\]

2 Chain Rule

\[
P(A, B, C) = P(A) \frac{P(A, B)}{P(A)} \frac{P(A, B, C)}{P(A, B)} = P(A)P(B|A)P(C|A, B)
\]

3 Naive Bayes Models

Features of the data (in this case, words) are conditionally independent of given the class \( c \).

\( C \): a class of document e.g. positive/negative
\( D \): a document

We want to know the class \( C \) of a document \( D \) by calculating the following term:

\[
C' = \arg \max_C P(C|D)
\]

The model tries to calculate

\[
P(C|D) \propto P(C)P(D|C)
\]

So now, we want to know \( P(C) \) and \( P(D|C) \).
Assumes that a word \( w \) occur independently within a document given a class.

\[
P(D|C) = \prod_{w \in D} P(w|C)
\]
4 Notes

1. Since we want to handle probabilities, the constraint $\sum c p_c = 1$ is set. Therefore, we want to optimize using the method of Lagrange multiplier.

2. Assume that the data likelihood when incorporating both labeled data and unlabeled data is

$$\log P(D^l) P(D^u) = \log P(D^l) + \log P(D^u)$$

As a result, the objective function of a Naive Bayes classifier becomes

$$\max_{\theta} \log P(\theta) + \log P(D^l) + \log P(D^u)$$

subject to

$$\sum_{c \in C} p_c = 1$$

$$\sum_w q_{w,c} = \sum_w P(W = w|C = c) = 1$$

References

