- Recall the principle of Maximum Likelihood Estimation (MLE):

Probabilistic text classification using BOW features

5

$$
\begin{aligned}
\hat{\boldsymbol{\theta}} &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, P(\mathbf{X}, \mathbf{y}; \boldsymbol{\theta}) \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left(\prod_{i=1}^{M} P(\mathbf{x}_i, y_i\ ; \boldsymbol{\theta}) \right)^{\frac{1}{M}} \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \frac{1}{M} \sum_{i=1}^{M} \log P(\mathbf{x}_i, y_i\ ; \boldsymbol{\theta}) \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{i=1}^{M} \log P(\mathbf{x}_i, y_i\ ; \boldsymbol{\theta})
\end{aligned}
$$

$\boldsymbol{\theta}$ = parameters to estimate

$$\mathbf{X} \in \mathbb{R}^{M \times N}$$

$$\mathbf{y} \in \{1,\ldots,K\}^M$$

$$M = \text{number of observations}$$

K = number of categories/classes

where

N = vocabulary size

Note: P is a generative model (note the joint distribution x,y, not y|x)

Note: y is a scalar representation of label, not a one-hot vector, hence lower case

BOW features

Probabilistic text classification using BOW features

- Recall the principle of Maximum Likelihood Estimation (MLE):

$$
\begin{aligned}
\hat{\boldsymbol{\theta}} &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\; P(\mathbf{X}, \mathbf{y}; \boldsymbol{\theta}) \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left(\prod_{i=1}^{M} P(\mathbf{x}_i, y_i\; ; \boldsymbol{\theta}) \right)^{\frac{1}{M}} \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\; \frac{1}{M} \sum_{i=1}^{M} \log P(\mathbf{x}_i, y_i\; ; \boldsymbol{\theta}) \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{i=1}^{M} \log P(\mathbf{x}_i, y_i\; ; \boldsymbol{\theta})
\end{aligned}
$$

where

$\boldsymbol{\theta}$ = parameters to estimate

$\mathbf{X} \in \mathbb{R}^{M \times N}$ ← BOW features

$\mathbf{y} \in \{1, \ldots, K\}^M$

M = number of observations

N = vocabulary size

K = number of categories/classes

Note: P is a generative model (note the joint distribution x,y, not y|x)

Note: y is a scalar representation of label, not a one-hot vector, hence lower case

Naive Bayes' classifier

- Assumption 1: The text and label in one document does not affect those of another.

- Assumption 2: Words in a sentence are independent, conditioned on the class label

- Taken from Eisenstein, 2019, Chp 2

Algorithm 1 Generative process for the Naïve Bayes classification model

for Instance $i \in \{1, 2, \ldots, M\}$ **do**:
    Draw the label $y^{(i)} \sim \text{Categorical}(\boldsymbol{\mu})$;
    Draw the word counts $\boldsymbol{x}^{(i)} \mid y^{(i)} \sim \text{Multinomial}(\boldsymbol{\phi}_{y^{(i)}})$.

chain rule $P(\mathbf{x}, y) = P(\mathbf{x} \mid y)P(y)$

where $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_K]$ label probability

$\boldsymbol{\phi} = [\phi_1, \ldots, \phi_N]$ word probability

$$P_{mult}(\mathbf{x} \mid y; \phi) = B(\mathbf{x}) \prod_{j=1}^{N} \phi_j^{\mathbf{x}_j} \qquad B(\mathbf{x}) = \frac{\left(\sum_{j=1}^{N} \mathbf{x}_j\right)!}{\prod_{j=1}^{N} \mathbf{x}_j!}$$