Naïve Bayes Model

- For simplicity, we only consider binary features

$$x_i \in \{0,1\}, \quad \text{i.e.,} \quad \boldsymbol{x} \in \{0,1\}^d$$

- The generation model is

$$t \sim \text{Categorical}(\pi_1, \cdots, \pi_K)$$

$$\boldsymbol{x}_i | t = k \sim \text{Bernoulli}(p_{k,i})$$

Here: A Bernoulli distribution parametrized by π means that
$\Pr[X = 1] = \pi$ and $\Pr[X = 0] = 1 - \pi$.
It is a special case of categorical distributions in that only two cases are considered.

- Such a model can be used to represent a document in text classification. For example, the target indicates Spam or NotSpam. The feature indicates if a word in the vocabulary occurs in the document.

Problem 1. Please show that the parameters of naïve Bayes decompose, i.e., the probability factorizes (for the same reason as Gaussian mixture models).

$$\begin{aligned}
\log L(D) &= \log \prod_{m=1}^{M} p(x^{(m)}, t^{(m)}) \\
&= \log \prod_{m=1}^{M} p(x^{(m)} \mid t^{(m)}) \, p(t^{(m)}) \\
&= \sum_{m=1}^{M} \log p(x^{(m)} \mid t^{(m)}) + \sum_{m=1}^{M} \log p(t^{(m)}) \\
&= \sum_{k=1}^{K} \sum_{m: t^{(m)}=k} \log p(x_i^{(m)} \mid t^{(m)} = k ; p_{k,i}) + \sum_{m=1}^{M} \log p(t^{(m)}; \pi_1, \cdots \pi_K)
\end{aligned}$$

Problem 2. Write out the MLE for naïve Bayes (which is simply counting).

Hint: No proof is needed for the second part, because the MLE for categorical distribution has been clear in the Gaussian mixture models.

$$\hat{\pi}_k = \frac{\sum_{m=1}^{M} \mathbb{1}\{t^{(m)} = k\}}{M}$$

$$\hat{p}_{k,i} = \frac{\sum_{m=1}^{M} \mathbb{1}\{t^{(m)} = k, x_i = 1\}}{\mathbb{1}\{t^{(m)} = k\}}$$

END OF W8