Problem 1.

Consider softmax regression $\boldsymbol{y} = \mathrm{softmax}(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b})$, where $\boldsymbol{x} \in \mathbb{R}^d$, and $\boldsymbol{y}, \boldsymbol{b} \in \mathbb{R}^K$. The cross entropy loss for a sample is

$$J = -\sum_{k=1}^{K} t_k \log y_k$$

where t_k indicates whether the sample is in the kth category or not.

Derive the gradients $\frac{\partial J}{\partial w_{k,i}}$ and $\frac{\partial J}{\partial b_k}$.

Solution:

1° Consider the loss for one sample. Superscript (m) is omitted for simplicity

$$J = -\sum_{k'} t_{k'} \log y_{k'} \quad \text{where} \quad y_k = \frac{\exp(z_k)}{\sum_{k''} \exp(z_{k''})} \quad \text{and} \quad z_k = w_k^T x + b$$

$$= -\sum_{k'} t_{k'} \left[\log \exp(z_{k'}) - \log \sum_{k''} \exp(z_{k''}) \right]$$

$$= -\sum_{k'} t_{k'} \left[z_{k'} - \log \sum_{k''} \exp(z_{k''}) \right]$$

$$= -\sum_{k'} t_{k'} z_{k'} + \log \sum_{k''} \exp(z_{k''})$$

[because one t^K is one] $\sum_{k'}$ for the second term is not needed

$$\frac{\partial J}{\partial z_k} = -\frac{\partial}{\partial z_k} \left[t_k z_k - \log \sum_{k''} \exp(z_{k''}) \right]$$

$$= -t_k + \frac{1}{\sum_{k''} \exp(z_{k''})} \cdot \exp(z_k) \cdot$$

$$= -t_k + y_k$$

Thus $\frac{\partial J}{\partial w_{k,i}} = \frac{\partial J}{\partial z_k} \cdot \frac{\partial z_k}{\partial w_{k,i}} = (y_k - t_k)\, x_i$

$$\frac{\partial J}{\partial b_k} = \frac{\partial J}{\partial z_k} \cdot \frac{\partial z_k}{\partial b_k} = y_k - t_k$$

Note: In this question, J is defined as the loss of a particular sample. If we compute the derivative of the total loss for multiple samples, we need to sum over different samples.

Problem 2.

Represent the above gradient in matrix-vector forms. In other words, write out the expressions for $\frac{\partial J}{\partial \boldsymbol{W}}$, $\frac{\partial J}{\partial \boldsymbol{b}}$.

Solution:

$$\frac{\partial J}{\partial W} = (y - t)\, x^{\top}$$

$$\frac{\partial J}{\partial b} = y - t$$

Note: Again, here we deal with the partial derivative of the loss for one sample. If multiple samples are considered, the $\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{x}$ vectors will be extended to a matrix.

Problem 3.

Consider a k-way classification. The predicted probability of a sample is $\boldsymbol{y} \in \mathbb{R}^K$, where y_k is the predicted probability of the kth category. Suppose correctly predicting a sample of category k leads to a utility of u_k. Incorrect predictions do not have utilities or losses.

Give the decision rule, i.e., a mapping from $\boldsymbol{y}$ to $\hat{t}$, that maximizes the total expected utility.

Solution:

3°
$$\mathop{\mathbb{E}}_{t\sim y}[u] = \sum_k y_k u_k \cdot \mathbb{1}\{\hat{t} = k\}$$

To maximize the utility

$$\hat{t} = \arg\max_k \; y_k u_k$$

Problem 4 (Naïve Bayes Model).

- For simplicity, we only consider binary features
  $x_i \in \{0,1\}$, i.e., $\boldsymbol{x} \in \{0,1\}^d$
- 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.

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

Solution:

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

b) 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.

Solution:

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

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