1 Multimodal fusion by Bayesian inference

Theorem 1. *Given N observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ and logits $z_{\boldsymbol{x}_1}, \ldots, z_{\boldsymbol{x}_N}$ such that for all relevant i, j:* $\text{softmax}_i(z_{\boldsymbol{x}_j}) = P(c_i|\boldsymbol{x}_j)$, *and assume for all classes c_i that $P(\boldsymbol{x}_1, ..., \boldsymbol{x}_N, c_i) > 0$. Then*

$$P(c_i|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) = \text{softmax}_i\left(\sum_{j=1}^{N} z_{\boldsymbol{x}_j} + \ln \boldsymbol{\kappa}\left(\boldsymbol{x}_1, ..., \boldsymbol{x}_N\right) - (N-1)\ln \boldsymbol{\pi}\right),$$

where $\boldsymbol{\pi}$ and $\boldsymbol{\kappa}\left(\boldsymbol{x}_1, ..., \boldsymbol{x}_N\right)$ are vectors in $\mathbb{R}^C$ with elements

$$\pi_i = P(c_i), \quad \kappa_i\left(\boldsymbol{x}_1, ..., \boldsymbol{x}_N\right) = \frac{P(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N|c_i)}{P(\boldsymbol{x}_1|c_i) \cdot \ldots \cdot P(\boldsymbol{x}_N|c_i)},$$

with C being the number of classes, and the logarithm is applied element-wise.

2 Proof of Theorem 1

Proof. Using Bayes' rule and multiplying by $\frac{\kappa_i(\boldsymbol{x}_1,...,\boldsymbol{x}_N)}{\kappa_i(\boldsymbol{x}_1,...,\boldsymbol{x}_N)}$, we obtain that:

$$\begin{aligned} P(c_i|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) &= \frac{P(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N|c_i)P(c_i)}{P(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)} \\ &= \left(\prod_{j=1}^{N} P(c_i|\boldsymbol{x}_j)\right) \frac{\kappa_i\left(\boldsymbol{x}_1, ..., \boldsymbol{x}_N\right)}{\pi_i^{N-1}} \frac{\prod_{j=1}^{N} P(\boldsymbol{x}_j)}{P(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)}. \end{aligned} \tag{1}$$

We notice that the fraction $\frac{P(\boldsymbol{x}_1,...,\boldsymbol{x}_N)}{\prod_{j=1}^{N} P(\boldsymbol{x}_j)}$ can be rewritten as such:

$$\begin{aligned} \frac{P(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)}{\prod_{j=1}^{N} P(\boldsymbol{x}_j)} &= \sum_{i=1}^{C} \frac{P(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N|c_i)P(c_i)}{P(\boldsymbol{x}_1) \cdot \ldots \cdot P(\boldsymbol{x}_N)} \\ &= \sum_{i=1}^{C} \frac{\kappa_i\left(\boldsymbol{x}_1, ..., \boldsymbol{x}_N\right)}{\pi_i^{N-1}} \prod_{j=1}^{N} P(c_i|\boldsymbol{x}_j) \\ &= \sum_{i=1}^{C} \frac{\kappa_i\left(\boldsymbol{x}_1, ..., \boldsymbol{x}_N\right)}{\pi_i^{N-1}} \prod_{j=1}^{N} \frac{e^{z_{\boldsymbol{x}_j,i}}}{S\left(z_{\boldsymbol{x}_j}\right)} \\ &= \frac{S(\sum_{j=1}^{N} z_{\boldsymbol{x}_j} + \ln \boldsymbol{\kappa}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) - (N-1)\ln \boldsymbol{\pi})}{S(z_{\boldsymbol{x}_1}) \cdot \ldots \cdot S(z_{\boldsymbol{x}_N})}, \end{aligned} \tag{2}$$

where $S(\boldsymbol{z}) = \sum_{i=1}^{C} e^{z_i}$. Now we substitute softmax for $P(c_i|\boldsymbol{x}_j)$ as well as the above result into equation 1 to get:

$$P(c_i|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) = \text{softmax}_i\left(\sum_{j=1}^{N} z_{\boldsymbol{x}_j} + \ln \boldsymbol{\kappa}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) - (N-1)\ln \boldsymbol{\pi}\right). \qquad \square$$

Remark 1. If we assume $P(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, c_i) = 0$ for some possible realization, then $\ln \kappa_i(\boldsymbol{x_1}, \ldots, \boldsymbol{x_N})$ or $\ln \pi_i$ is undefined and we have $P(c_i|\boldsymbol{x_1}, \ldots, \boldsymbol{x_N}) = 0$.

If we assume that $P(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N|c_i) = P(\boldsymbol{x}_1|c_i) \cdot \ldots \cdot P(\boldsymbol{x}_N|c_i)$, and avoid using $\kappa_i(\boldsymbol{x_1}, \ldots, \boldsymbol{x_N})$ to resolve dependencies in the derivation, we get the same result as in equation (1). Here we can relax the assumption to $P(c_i) > 0$.

Remark 2. We see that for ordinary logistic regression on the concatenated embeddings, a weight and a bias exist such that it is equal to the naive Bayes fusion (i.e. when $\ln \kappa\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\right) = 0$). Assume we have $z_{\boldsymbol{x}_j} = \boldsymbol{W}_j v_{\boldsymbol{x}_j} + \boldsymbol{b}_j$ for all N classifiers, as well as the block-matrices $\boldsymbol{W} = [\boldsymbol{W}_1| \ldots |\boldsymbol{W}_N]$, and $v = [v_{\boldsymbol{x}_1}^T| \ldots |v_{\boldsymbol{x}_N}^T]^T$, and the bias $\boldsymbol{b} = (1-N)\ln \boldsymbol{\pi} + \sum_{j=1}^N \boldsymbol{b}_j$. We then see that

$$\operatorname{softmax}\left(\boldsymbol{W}v + \boldsymbol{b}\right) = \operatorname{softmax}\left((1-N)\ln \pi + \sum_{j=1}^{N} z_{\boldsymbol{x}_j}\right).$$