

## 1 Bayesian fusion

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

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

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

$$\pi_i = P(c_i), \quad \kappa_i(\mathbf{x}_1, \dots, \mathbf{x}_N) = \frac{P(\mathbf{x}_1, \dots, \mathbf{x}_N|c_i)}{P(\mathbf{x}_1|c_i) \cdot \dots \cdot P(\mathbf{x}_N|c_i)},$$

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  $\frac{\kappa_i(\mathbf{x}_1, \dots, \mathbf{x}_N)}{\pi_i^{N-1}}$ , we obtain that:

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

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

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

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

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

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

If we assume that  $P(\mathbf{x}_1, \dots, \mathbf{x}_N | c_i) = P(\mathbf{x}_1 | c_i) \cdot \dots \cdot P(\mathbf{x}_N | c_i)$ , and avoid using  $\kappa_i(\mathbf{x}_1, \dots, \mathbf{x}_N)$  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 Bayesian fusion when  $\ln \kappa(\mathbf{x}_1, \dots, \mathbf{x}_N) = 0$ . Assume we have  $z_{\mathbf{x}_j} = \mathbf{W}_j v_{\mathbf{x}_j} + \mathbf{b}_j$  for all  $N$  classifiers, as well as the block-matrices  $\mathbf{W} = [\mathbf{W}_1 | \dots | \mathbf{W}_N]$ , and  $v = [v_{\mathbf{x}_1}^T | \dots | v_{\mathbf{x}_N}^T]^T$ , and the bias  $\mathbf{b} = (1 - N) \ln \pi + \sum_{j=1}^N \mathbf{b}_j$ . We then see that

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