Probabilistic Machine Learning

Exercise Sheet #2

due on Monday, 1 May 2023, 10am sharp

1. **Theory Question:** In the lecture, we encountered the **Beta** distribution

$$p(x \mid a, b) = \mathcal{B}(x; a, b) := \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}$$

where the normalization constant $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ is the *Beta function*, using the (Euler) Gamma function. The Beta is **conjugate** to the **binomial** distribution

$$p(n, m \mid x) = \binom{n+m}{n} x^n (1-x)^m.$$

That is, the posterior on x arising from the Beta prior and the binomial likelihood is $\mathcal{B}(x; a+n, b+m)$. In this question, we will study the generalization to the **multinomial** case: Consider a data set $C = [c_1, c_2, \ldots, c_N]$ of discrete labels $c_i \in \{1, 2, 3, \ldots, K\}$. It is convenient to use the notation $n_k = |\{c_i \in C \mid c_i = k\}|$ for the number of labels observed in class k. We assume these data are drawn iid. from the multinomial distribution

$$p(C \mid \boldsymbol{x}) = \prod_{i=1}^{N} x_{c_i} = \frac{N!}{\prod_{\ell=1}^{K} n_\ell!} \prod_{k=1}^{K} x_k^{n_k}. \tag{1}$$

   (a) Show that the **Dirichlet** distribution (with parameter vector $\boldsymbol{\alpha} \in \mathbb{R}_+^K$)

$$\mathcal{D}(\boldsymbol{x} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k x_k^{\alpha_k - 1}$$

   is the conjugate prior for the multinomial (1). What is the associated posterior?

   (b) Show that the Dirichlet distribution has the *aggregation property*: If $p(\boldsymbol{x}) = \mathcal{D}(\boldsymbol{x}; \boldsymbol{\alpha})$, then

$$p([x_1, x_2, \ldots, x_i + x_j, \ldots x_{j-1}, x_{j+1}, \ldots, x_K]) \\ = \mathcal{D}(\cdot; [\alpha_1, \alpha_2, \ldots, \alpha_i + \alpha_j, \ldots, \alpha_{j-1}, \alpha_{j+1}, \ldots, \alpha_K]).$$

   (c) Show that the logarithm of the likelihood (1), as a function of $\boldsymbol{x}$, is the **cross-entropy loss**[1] for $\boldsymbol{x}$. What is the **maximum-likelihood** estimate for $\boldsymbol{x}$? How does it relate to the maximum[2] of the **posterior**?

   (d) Consider the following situation: You are on an e-commerce marketplace, which uses ratings on the common "five stars" scale. You have the choice between two vendors there. The first vendor has only three ratings so far, but they are all five-star ratings (i.e. their rating "vector" is $C_1 = [0, 0, 0, 0, 3]$). The second vendor has many more ratings, their rating vector is $C_2 = [1, 0, 12, 43, 354]$. Let's assume (questionably) that these ratings are drawn iid. from two multinomial distributions with parameter $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, one for each vendor. Use the results above to answer the following questions:

      i. What are the maximum likelihood estimates for $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$?

      ii. Under a uniform Dirichlet prior ($\boldsymbol{\alpha} = [1, 1, 1, 1, 1]$) what is the probability for *your* rating to be five stars? What is the probability for your rating to be *four or five* stars? (Assume you pick either one of the vendors, and your future rating will be distributed just as the existing ones. To answer this question, you need to compute integrals of the form $\int x_i \mathcal{D}(\boldsymbol{x} \mid \boldsymbol{\alpha})\, d\boldsymbol{x}$. Think about how to do this.)

2. **Practical Question:** can be found in `Ex02.ipynb`

[1] if necessary, look it up.
[2] derive it!