Homework 8

Deadline: Wednesday, March 29, at 11:59pm.

Submission: You must submit your solutions as a PDF file through MarkUs[1]. You can produce the file however you like (e.g. LaTeX, Microsoft Word, scanner), as long as it is readable.

Late Submission: MarkUs will remain open until 2 days after the deadline; until that time, you should submit through MarkUs. If you want to submit the assignment more than 2 days late, please e-mail it to `csc321ta@cs.toronto.edu`. The reason for this is that MarkUs won't let us collect the homeworks until the late period has ended, and we want to be able to return them to you in a timely manner.

Weekly homeworks are individual work. See the Course Information handout[2] for detailed policies.

1. **Categorial Distribution.** Let's consider fitting the categorical distribution, which is a discrete distribution over K outcomes, which we'll number 1 through K. The probability of each category is explicitly represented with parameter θ_k. For it to be a valid probability distribution, we clearly need $\theta_k \geq 0$ and $\sum_k \theta_k = 1$. We'll represent each observation $\mathbf{x}$ as a 1-of-K encoding, i.e, a vector where one of the entries is 1 and the rest are 0. Under this model, the probability of an observation can be written in the following form:

$$p(\mathbf{x}; \boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k^{x_k}.$$

   (a) **[2pts]** Determine the formula for the maximum likelihood estimate of the parameters in terms of the counts $N_k = \sum_i x_k^{(i)}$ of all the outcomes. You may assume all of the counts are nonzero. Note that your solution needs to obey the constraints $\theta_k \geq 0$ and $\sum_k \theta_k = 1$. You might want to see Khan Academy[3] for background on constrained optimization.

   (b) **[2pts]** Now consider Bayesian parameter estimation. For the prior, we'll use the Dirichlet distribution, which is defined over the set of probability vectors (i.e. vectors that are nonnegative and whose entries sum to 1). Its PDF is as follows:

$$p(\boldsymbol{\theta}) \propto \theta_1^{a_1 - 1} \cdots \theta_K^{a_k - 1}.$$

   A useful fact is that if $\boldsymbol{\theta} \sim \text{Dirichlet}(a_1, \ldots, a_K)$, then

$$\mathbb{E}[\theta_k] = \frac{a_k}{\sum_{k'} a_{k'}}.$$

   Determine the posterior distribution $p(\boldsymbol{\theta} \,|\, \mathcal{D})$, where $\mathcal{D}$ is the set of observations. From that, determine the posterior predictive probability that the next outcome will be k.

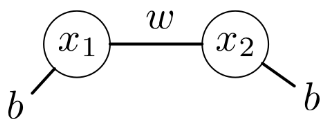
   (c) **[1pt]** Still assuming the Dirichlet prior distribution, determine the MAP estimate of the parameter vector $\boldsymbol{\theta}$.

[1] `https://markus.teach.cs.toronto.edu/csc321-2017-01`

[2] `http://www.cs.toronto.edu/~rgrosse/courses/csc321_2017/syllabus.pdf`

[3] `https://www.khanacademy.org/math/multivariable-calculus/applications-of-multivariable-derivatives/lagrange-multipliers-and-constrained-optimization/v/constrained-optimization-introduction`

2. **Gibbs Sampling.** Suppose we have a Boltzmann machine with two variables, with a weight w and the same bias b for both variables:



The variables both take values in $\{-1, 1\}$.

(a) **[2pts]** For the values $w = 1$ and $b = -0.2$, determine the probability distribution over all four configurations of the variables. Then determine the marginal probability $p(x_1 = 1)$. Now do the same thing for $w = 3$ and $b = -0.2$. (Give your answers to three decimal places.)

(b) We'll initialize both variables to 1, and then run Gibbs sampling to try to sample from the model distribution. On odd numbered steps, we will resample x_1 from its conditional distribution; on even numbered steps, we will resample x_2 from its conditional distribution. Denote by α_t the probability that the value chosen by Gibbs sampling on the t^{th} step is a 1. Since we are initializing to 1, we have the initial condition $\alpha_0 = 1$.

i. **[2pts]** Give a recurrence that expresses α_t as a function of α_{t-1}.

ii. **[1pts]** Using your recurrence, determine the value of α_{100} for each of the two cases given above (i.e. ($w = 1$, $b = -0.2$), and ($w = 3$, $b = -0.2$)). Give your answer to three decimal places. We encourage you to try to derive an explicit formula for α_t, but we'll give you full credit if you just write a program to compute it.