Homework Assignment 2b

Due: Friday, Mar. 3, 2024, 11:59 p.m. Mountain time
Total marks: 50

Question 1. [25 MARKS]

Imagine that you would like to predict if your favorite table will be free at your favorite restaurant. The only additional piece of information you can collect, however, is if it is sunny or not sunny. Therefore, you would like to predict whether the table will be free or not given the weather.

You collect paired samples from visit of the form (is sunny, is table free), where it is either sunny (1) or not sunny (0) and the table is either free (1) or not free (0). Your goal is to learn P(Table is Free|Weather Information), which then also automatically tells you
P(Table is Not Free|Weather Information) by taking 1-P(Table is Free|Weather Information). Given these learned probabilities, you can make predictions about whether your table is free or not by checking if P(Table is Free|Weather Information) > 0.5. If there is a greater than 50% probability your table is free, you are happy to risk the icy streets to get to your favorite restaurant.

Hint: To help you with this question, see Section 5.4, Example 24 about books.

(a) [15 MARKS]

Formulate this distribution learning problem as a maximum likelihood problem. Recall your goal is to learn distribution P(Table is Free|Weather Information). Start by stating what the random variables are for this problem and what values they can take. Then explain what distributions on these random variables you want to learn and what parameters need to be learned. Finally, write the negative log likelihood explicitly for those distributions and parameters. Recall, $p(\mathcal{D}|\mathbf{w})$ is the likelihood where $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ is a dataset of the collected paired samples and $\mathbf{w}$ are your parameters. The parameters are bold because they might consist of more than one parameter, and so would be a vector rather than a scalar. **Note:** You do not need to solve this maximum likelihood problem, just formalize it.

(b) [5 MARKS] Assume you have collected data for the last 10 days and computed the maximum likelihood solution $\mathbf{w}^*$ to the problem formulated in (a). You do not actually have to compute the solution, just assume that you did and now have estimated the parameters $\mathbf{w}^*$ for your distribution. If it is sunny today, then how would you predict if your table will be free? Be precise and explain how you would use your learned distribution and parameters to do this.

(c) [5 MARKS] Imagine you could further gather information about if it is morning, afternoon, or evening, and learn P(Table is Free|Weather Information, Time of Day). How does this change the maximum likelihood problem? You do not need to write the log likelihood explicitly for this question, just explain if you need any new random variables, what values it can take, and what parameters you will need.

Question 2. [25 MARKS]

We can combine simple distributions to produce more complex (multi-modal) distributions using *mixtures*. The below figure shows what occurs if we take a convex combination of two Gaussians. We can write the distribution for such a random variable X with density corresponding to an equal mixture of two Gaussians, with unknown means μ_1, μ_2 and unknown variances σ_1^2, σ_2^2, as follows.

$$p(x|\mathbf{w}) = 0.5\mathcal{N}(x|\mu_1, \sigma_1^2) + 0.5\mathcal{N}(x|\mu_2, \sigma_2^2) \tag{1}$$

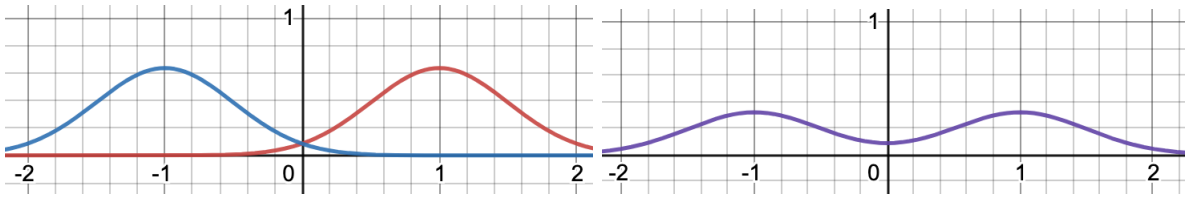

Figure 1: The blue curve is a Gaussian with $\mu_1 = -1$ and $\sigma_1 = 0.5$ and the red curve is a Gaussian with $\mu_1 = 1$ and $\sigma_1 = 0.5$. The purple curve is the mixture of the two, as in Equation (1). The purple curve allows us to model a bimodal distribution (two peaks), where now the two most likely values are -1 and 1, with density decreasing from these points. If we sample from this distribution, then we will see points centered around -1 and 1, with a reasonable likelihood for a point between the two (including at zero), and very low likelihood for points outside -2 and 2.

where we write the four parameters $\mathbf{w} = (\mu_1, \mu_2, \sigma_1, \sigma_2)$ and

$$\mathcal{N}(x|\mu, \sigma^2) = (2\pi)^{-1/2}\sigma^{-1}\exp(-(x-\mu)^2/(2\sigma^2)). \tag{2}$$

It is easy to show that $p(x|\mathbf{w})$ is a valid density, because

$$\begin{aligned}\int p(x|\mathbf{w})dx &= \int 0.5\mathcal{N}(x|\mu_1, \sigma_1^2) + 0.5\mathcal{N}(x|\mu_2, \sigma_2^2)dx \\ &= 0.5\int \mathcal{N}(x|\mu_1, \sigma_1^2)dx + 0.5\int \mathcal{N}(x|\mu_2, \sigma_2^2)dx \\ &= 0.5 + 0.5 = 1.\end{aligned}$$

You set forth to learn this distribution $p(x|\mathbf{w})$. However, now when you take the log-likelihood, you find that the log does not help as much, because the sum gets in the way of the log being applied to the exponentials.

$$\ln p(x|\mathbf{w}) = \ln\left(0.5\mathcal{N}(x|\mu_1, \sigma_1^2) + 0.5\mathcal{N}(x|\mu_2, \sigma_2^2)\right)$$

The log still helps convert the product over samples into a sum, for a given dataset of n iid samples from this distribution $\mathcal{D} = \{x_i\}_{i=1}^n$,

$$\ln p(\mathcal{D}|\mathbf{w}) = \ln\prod_{i=1}^n p(x_i|\mathbf{w}) = \sum_{i=1}^n \ln p(x_i|\mathbf{w})$$

Despite this difficulty, you are determined to learn this distribution, because you are confident it will do a better job of modeling your data. Your goal in this question is to obtain a procedure to estimate $\mathbf{w} = (\mu_1, \mu_2, \sigma_1, \sigma_2)$.

(a) [20 MARKS] Compute the gradient (partial derivatives) of your negative log likelihood objective $c(\mathbf{w}) \doteq -\ln p(\mathcal{D}|\mathbf{w})$. Start by computing the gradient of $\ln p(x_i|\mathbf{w})$. To simplify notation, consider defining $g_i(\mu_j, \sigma_j) = \sigma_j^{-1}\exp(-(x_i - \mu_j)^2/(2\sigma_j^2))$.

(b) [5 MARKS] Write the (first-order) gradient descent update rule for your parameters, using the gradient you compute, assuming you start from current point $\mathbf{w}_t$ and have stepsize η_t.

Homework policies:

Your assignment should be submitted as one pdf document on eClass. The pdf must be written legibly and scanned or must be typed (e.g., Latex). This .pdf should be named Firstname_LastName_Sol.pdf,

Because assignments are more for learning, and less for evaluation, grading will be based on coarse bins. **The grading is atypical**. For grades between (1) 81-100, we round-up to 100; (2) 61-80, we round-up to 80; (3) 41-60, we round-up to 60; and (4) **0-40, we round down to 0**. The last bin is to discourage quickly throwing together some answers to get some marks. The goal for the assignments is to help you learn the material, and completing less than 50% of the assignment is ineffective for learning.

We will not accept late assignments. There is no late penalty policy. The assignments must be submitted electronically via eClass on time, by 11:59 pm Mountain time on the due date. There is a grace period of 48 hours when assignments will be accepted. No submissions will be accepted after 48 hours after the deadline, and the assignment will be considered as incomplete if not submitted.

All assignments are individual. All the sources used for the problem solution must be acknowledged, e.g. web sites, books, research papers, personal communication with people, etc. Academic honesty is taken seriously; for detailed information see the University of Alberta Code of Student Behaviour.

Good luck!