CIS4930 - Math for Machine Learning

Homework 7

Fernando Sckaff

April 23, 2023

Question 1

You are given the following energy function

$$E(s_1, s_2) = -Js_1s_2 + h_1s_1 + h_2s_2$$

with $s_1, s_2 \in \{0, 1\}$. Write down the Gibbs distribution $\Pr(S_1 = s_1, S_2 = s_2) = \frac{\exp\{-\beta E(s_1,s_2)\}}{Z}$. Work out the expression for Z and the expected value $\mathcal{E}(s_1s_2)$. Show all steps.

Solution:

Gibbs distribution:

$$\Pr(S_1 = s_1, S_2 = s_2) = \frac{\exp\{-\beta(Js_1s_2 - h_1s_1 - h_2s_2)\}}{Z}$$

Z:

$$\begin{aligned} Z &= \sum_{\{s_1,s_2\}} \exp\{-\beta E(s_1, s_2)\} \\ &= \sum_{\{s_1,s_2\}} \exp\{-\beta[Js_1s_2 - h_1s_1 - h_2s_2]\} \end{aligned}$$

There are 4 configurations for $\{s_1, s_2\}$. That is $\{0, 0\}, \{1, 0\}, \{0, 1\}$ and $\{1, 1\}$. Therefore,

$$\begin{aligned} Z &= \exp\{-\beta[0]\} + \exp\{\beta[-h_1]\} + \exp\{\beta[-h_2]\} + \exp\{-\beta[J - h_1 - h_2]\} \\ Z &= 1 + \exp\{-\beta h_1\} + \exp\{-\beta h_2\} + \exp\{-\beta[J - h_1 - h_2]\} \end{aligned}$$

Moreover, let's calculate $\mathcal{E}(s_1s_2)$:

$$\mathcal{E}(s_1s_2) = (0 * 0)\Pr(S_1 = 0, S_2 = 0) + (1 * 0)\Pr(S_1 = 1, S_2 = 0) + (0 * 1)\Pr(S_1 = 0, S_2 = 1) + (1 * 1)\Pr(S_1 = 1, S_2 = 1)$$
$$\mathcal{E}(s_1s_2) = \Pr(S_1 = 1, S_2 = 1)$$

Since the Gibbs distribution is:

$$\Pr(S_1 = s_1, S_2 = s_2) = \frac{\exp\{-\beta(Js_1s_2 - h_1s_1 - h_2s_2)\}}{1 + \exp\{-\beta h_1\} + \exp\{-\beta h_2\} + \exp\{-\beta[J - h_1 - h_2]\}}$$

Then,

$$\mathcal{E}(s_1s_2) = \frac{\exp\{-\beta(J - h_1 - h_2)\}}{1 + \exp\{-\beta h_1\} + \exp\{-\beta h_2\} + \exp\{-\beta[J - h_1 - h_2]\}}$$

Question 2

Cover and Thomas, Chapter 12, problem 2: Minimize $D(p\|q)$ under constraints on p. We wish to find the (parametric form) of the probability mass function $p(x), x \in \{1, 2, ...\}$ that minimizes the KL divergence $D(p\|q)$ over $p(x)$ such that $\sum p(x)g_i(x) = \alpha_i, i = 1, 2, ...$ Here $q(x)$ is any given probability mass function (though usually can be taken as the true distribution of the data).

Solution:

The KL Divergence for discrete distributions is

$$D(p||q) = \sum_{k=1}^{K} p_k \log \frac{p_k}{q_k} \geqslant 0$$

where K is the number of possibilities.

Let's find the minimum entropy for $D(p\|q)$. Consider the constraint $\sum_i^m p(x)g_i(x) = \alpha_i$ where m is the number of multipliers. Then:

$$L(p, \lambda) = D(p||q) + \sum_{i=1}^{m} \lambda_i \left(p(x)g_i(x) - \alpha_i\right)$$

We want to find the values of $p(x)$ that minimize $L(p, \lambda)$. To do this, we take the derivative of $L(p, \lambda)$ with respect to $p(x)$ and set it to zero:

$$\frac{\partial L(p,\lambda)}{\partial p(x)} = \frac{\partial}{\partial p(x)}\left[p(x)\log\frac{p(x)}{q(x)} + \sum_{i=1}^{m}\lambda_i\left(p(x)g_i(x) - \alpha_i\right)\right] = 0$$

After simplifying,

$$0 = \log\frac{q(x)}{p(x)} + 1 + \sum_{i=1}^{m}\lambda_i g_i(x)$$
$$p(x) = q(x)\exp\left(-1 - \sum_{i=1}^{m}\lambda_i g_i(x)\right)$$

Is the parametric form of the probability mass function. 😊

Question 3

Is the Bregman divergence $\phi(y) - \phi(x) - (y - x)\phi'(x)$ convex w.r.t. both x and y [for a three times differentiable $\phi(x)$]? Mathematically justify your answer. You could evaluate the matrix of second H and check if it is non-negative definite, i.e. $u^T H u \geqslant 0, \forall u$.

Solution:
We have:

$$\begin{aligned}\frac{\partial^2}{\partial x^2}(\phi(y) - \phi(x) - y\phi'(x) + x\phi'(x)) &= -\phi''(x) + y\phi''(x)\\ &= (y-1)\phi''(x)\end{aligned}$$

Similarly, we have:

$$\frac{\partial^2}{\partial y^2}(\phi(y) - \phi(x) - y\phi'(x) + x\phi'(x)) = \phi''(y)$$
$$\frac{\partial^2}{\partial x \partial y}(\phi(y) - \phi(x) - y\phi'(x) + x\phi'(x)) = 0$$

The Hessian matrix is then:

$$H = \begin{bmatrix}(y-1)\phi''(x) & 0\\ 0 & \phi''(y)\end{bmatrix}$$

To check if H is non-negative definite, we need to check if $u^T H u \geqslant 0$ for all vectors u. Let $u = [u_1, u_2]^T$ be an arbitrary vector. Then we have:

$$u^T H u = (y-1)\phi''(x)u_1^2 + \phi''(y)u_2^2 \geqslant 0$$

Since $\phi''(x)$ and $\phi''(y)$ are both non-negative (as $\phi(x)$ is assumed to be three times differentiable), it follows that $u^T H u \geqslant 0$ for all u. Therefore, H is non-negative definite. Since the Hessian matrix is positive semi-definite, the function is convex w.r.t. both x and y.

Question 4

You are in charge of estimating the probability distribution of midterm scores of a class. Assume there are N students in a class and their scores take K *discrete* possibilities (for the sake of simplicity). Further, assume that N_k students get a score k (with k ranging from 0 to $K-1$). If the student scores are considered to be independent and identically distributed (i.i.d) with each (random) student score denoted by the random variable $X_i, i \in \{1, ..., N\}$ and taking the value x_i (which is one of the K *discrete* possibilities), this yields a (taken to be TRUE) probability distribution $\Pr(X = k) = p_k = \frac{N_k}{N}, k = 0, 1, ..., K-1$. What is the relationship between the class entropy $-\sum_{k=0}^{K-1} p_k \log p_k$ and the quantity $-\frac{1}{N}\sum_{i=1}^{N} \log \Pr(x_i)$? Give a detailed explanation while noting that the entropy expression sums over the set of possibilities while the negative log-likelihood sums over the set of students.

Solution: The class entropy $H(X)$ is defined as:

$$H(X) = -\sum_{k=0}^{K-1} p_k \log p_k$$

where p_k is the probability that a student gets a score of k. The entropy measures the uncertainty in the distribution of scores across the class. Higher entropy means more uncertainty and lower entropy means less uncertainty.

On the other hand, the negative log-likelihood is defined as:

$$-\frac{1}{N}\sum_{i=1}^{N} \log \Pr(x_i)$$

where $\Pr(x_i)$ is the probability that a particular student gets the score x_i. The negative log-likelihood measures how well a particular probability distribution fits the observed data. Lower negative log-likelihood means a better fit of the distribution to the data.

Now, consider a set of N i.i.d. random variables $X_1, X_2, ..., X_N$. The joint probability of observing the values $x_1, x_2, ..., x_N$ is:

$$\Pr(X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N) = \prod_{i=1}^{N} \Pr(X_i = x_i)$$

Using the definition of p_k given in the question, we can write:

$$\Pr(X_i = x_i) = p_{x_i} = \frac{N_{x_i}}{N}$$

Substituting this in the expression for negative log-likelihood, we get:

$$\begin{aligned}
-\frac{1}{N}\sum_{i=1}^{N} \log \Pr(x_i) &= -\frac{1}{N}\sum_{i=1}^{N} \log p_{x_i} \\
&= -\frac{1}{N}\sum_{k=0}^{K-1}\sum_{i=1}^{N} I(x_i = k) \log\left(\frac{N_k}{N}\right) \\
&= -\sum_{k=0}^{K-1} p_k \log\left(\frac{N_k}{N}\right) \\
&= -\sum_{k=0}^{K-1} p_k \log p_k + \sum_{k=0}^{K-1} p_k \log N
\end{aligned}$$

where $I(x_i = k)$ is an indicator function that takes value 1 if $x_i = k$ and 0 otherwise (as it was shown in class as a way to iterate through all data). The sum of the indicator function over all students who get a score of k counts the number of students who get a score of k, since the indicator function takes the value 1 for each student who

gets a score of k, and 0 otherwise. Therefore, this sum is equal to the number of students who get a score of k, which is N_k. Dividing this by N gives p_k. The second term of the equation can be seen as a constant term, since the sum of the probabilities over all possibilities is 1. Therefore, we have:

$$-\frac{1}{N}\sum_{i=1}^{N}\log \Pr(x_i) = -\sum_{k=0}^{K-1} p_k \log p_k + \text{constant}$$

where the constant term does not depend on p_k. Therefore, the relationship between the negative log-likelihood and entropy is that they differ by a constant term, which does not affect the comparison of the two quantities. In other words, minimizing the negative log-likelihood is equivalent to maximizing the entropy, up to an additive constant. Hence, maximizing the entropy corresponds to finding the probability distribution that best fits the observed data.

Machine Learning for the win!

F.S.

Note:-

- Question 1: Correct
- Question 2: 20/25 No sum to one constraint.
- Question 3: 15/25. Incomplete algebra and conclusions are wrong.
- Question 4: Correct. However, I could have been tighter since $p_k = N_k/N$ and therefore the NLL = Entropy in this case.