Introduction to Machine Learning, Fall 2023
Homework 2

(Due Tuesday Nov. 14 at 11:59pm (CST))

October 25, 2023

1. [10 points] [Convex Optimization Basics]
   (a) Proof any norm $f : \mathbb{R}^n \to \mathbb{R}$ is convex. [2 points]
   (b) Determine the convexity (i.e., convex, concave or neither) of $f(x_1, x_2) = x_1^2/x_2$ on $\mathbb{R} \times \mathbb{R}_{>0}$. [2 points]
   (c) Determine the convexity of $f(x_1, x_2) = x_1/x_2$ on $\mathbb{R}^2_{>0}$. [2 points]
   (d) Recall Jensen's inequality $f(\mathbb{E}(X)) \leq \mathbb{E}(f(X))$ if f is convex for any random variable X. Proof the log sum inequality:
   $$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i\right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$$
   where $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ are positive numbers. Hints: $f(x) = x \log x$ is strictly convex. [4 points]

Solution:

(a) let x, y be two points in R^n and let λ be a scalar in $[0,1]$

the point $\lambda x + (1-\lambda) y$ lies on the line connecting x and y

By the triangle inequality: $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$

The inequality holds for any norm f because of the properties of norm

So any norm $f: R^n \to R$ is convex

(b) $f_{x_1} = \frac{2x_1}{x_2}$ $\quad f_{x_2} = -\frac{x_1^2}{x_2^2}$ $\quad f_{x_1x_1} = \frac{2}{x_2}$ $\quad f_{x_2x_2} = \frac{2x_1^2}{x_2^3}$

since x_2 must be greater than 0

$\frac{2}{x_2}$ is always non-negative and $\frac{2x_1^2}{x_2^3}$ is also non-negative

for all x_1 and x_2 in the domain

so $f(x_1, x_2) = \frac{x_1^2}{x_2}$ is convex on $R \times R_{>0}$

(c) $f_{x_1} = \frac{1}{x_2}$ $f_{x_2} = -\frac{x_1}{x_2^2}$ $f_{x_1x_1} = 0$ $f_{x_2x_2} = \frac{2x_1}{x_2^3}$

since all (x_1, x_2) are in the domain $R^2_{>0}$

$\frac{2x_1}{x_2^3}$ is non-negative if x_1 and x_2 have the same sign

0 is always non-negative

So $f(x_1, x_2) = \frac{x_1}{x_2}$ is convex on $R^2_{>0}$ if x_1, x_2 are both positive or negative.

(d) $f'(x) = 1 + \log x$ $f''(x) = \frac{1}{x}$

since $f''(x) = \frac{1}{x} > 0$ for all $x > 0$, the function $f(x) = x\log x$ is strictly convex

In this case. X takes values $a_1, a_2, \cdots, a_n$ with probabilities $\frac{a_1}{\sum_{i=1}^n a_i}$, $\frac{a_2}{\sum_{i=1}^n a_i}$, ..., $\frac{a_n}{\sum_{i=1}^n a_i}$, applying Jensen's Inequality to the function $f(x) = x\log x$. we can get

$$\sum_{i=1}^n \left(\frac{a_i}{\sum_{i=1}^n a_i}\right) \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^n \frac{a_i}{\sum_{i=1}^n a_i}\right) \log\left(\frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i}\right)$$

$$\Rightarrow \sum_{i=1}^n a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^n a_i\right) \log \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i}$$

2. [10 points] [Linear Methods for Classification] Consider the "Multi-class Logistic Regression" algorithm. Given training set $\mathcal{D} = \{(x^i, y^i) \mid i = 1, \dots, n\}$ where $x^i \in \mathbb{R}^{p+1}$ is the feature vector and $y^i \in \mathbb{R}^k$ is a one-hot binary vector indicating k classes. We want to find the parameter $\hat{\beta} = [\hat{\beta}_1, \dots, \hat{\beta}_k] \in \mathbb{R}^{(p+1)\times k}$ that maximize the likelihood for the training set. Introducing the softmax function, we assume our model has the form

$$p(y_c^i = 1 \mid x^i; \beta) = \frac{\exp(\beta_c^\top x^i)}{\sum_{c'} \exp(\beta_{c'}^\top x^i)},$$

where y_c^i is the c-th element of y^i.

(a) Complete the derivation of the conditional log likelihood for our model, which is

$$\ell(\beta) = \ln \prod_{i=1}^{n} p(y_t^i \mid x^i; \beta) = \sum_{i=1}^{n} \sum_{c=1}^{k} \left[y_c^i(\beta_c^\top x^i) - y_c^i \ln \left(\sum_{c'} \exp(\beta_{c'}^\top x^i) \right) \right].$$

For simplicity, we abbreviate $p(y_t^i = 1 \mid x^i; \beta)$ as $p(y_t^i \mid x^i; \beta)$, where t is the true class for x^i. [4 points]

(b) Derive the gradient of $\ell(\beta)$ w.r.t. β_1, i.e.,

$$\nabla_{\beta_1} \ell(\beta) = \nabla_{\beta_1} \sum_{i=1}^{n} \sum_{c=1}^{k} \left[y_c^i(\beta_c^\top x^i) - y_c^i \ln \left(\sum_{c'} \exp(\beta_{c'}^\top x^i) \right) \right].$$

Remark: Log likelihood is always concave; thus, we can optimize our model using gradient ascent. (The gradient of $\ell(\beta)$ w.r.t. $\beta_2, \dots, \beta_k$ is similar, you don't need to write them) [6 points]

Solution:

(a) the likelihood function for (x^i, y^i) is $p(y_t^i \mid x^i; \beta) = \prod_{c=1}^{k} p(y_c^i \mid x^i; \beta)^{y_c^i}$

Using softmax function: $p(y_c^i \mid x^i; \beta) = \dfrac{\exp(\beta_c^T x^i)}{\sum_{c'=1}^{k} \exp(\beta_{c'}^T x^i)}$

Substitute it to the likelihood function

$$p(y_t^i \mid x^i; \beta) = \prod_{c=1}^{k} \left(\frac{\exp(\beta_c^T x^i)}{\sum_{c'=1}^{k} \exp(\beta_{c'}^T x^i)} \right)^{y_c^i}$$

$$\ln p(y_t^i \mid x^i; \beta) = \sum_{c=1}^{k} (y_c^i \cdot \beta_c^T x^i) - y_c^i \ln \left(\sum_{c'=1}^{k} \exp(\beta_{c'}^T x^i) \right)$$

Sum all n points up

$$l(\beta) = \sum_{i=1}^{n} \sum_{c=1}^{k} \left[y_c^i (\beta_c^T x^i) - y_c^i \ln \left(\sum_{c'} \exp(\beta_{c'}^T, x^i) \right] \right.$$

(b) for $(\beta_1)^T x^i$, treat x^i as constants since we are differentiating with respect to β_1. So the derivative of $(\beta_1)^T x^i$ with respect to β_1 is just x^i.

Using the chain rule, the derivative of the second part with respect to $(\beta_1)^T x^i$ is $\dfrac{\exp(\beta_1^T x^i)}{\sum_{c'=1}^{K} \exp(\beta_{c'}^T x^i)}$

So $$\frac{\partial l(\beta)}{\partial \beta_1} = \sum_{i=1}^{n} \left[x^i - \frac{\exp(\beta_1^T x^i)}{\sum_{c'=1}^{K} \exp(\beta_{c'}^T x^i)} x^i \right]$$

3. [10 points] [Probability and Estimation] Suppose $\mathcal{D} = \{x_1, x_2, \ldots, x_n\}$ are i.i.d. samples from exponential distribution with parameter $\lambda > 0$, i.e., $X \sim \text{Expo}(\lambda)$. Recall the PDF of exponential distribution is

$$p(x \mid \lambda) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & \text{otherwise} \end{cases}.$$

(a) To derive the posterior distribution of λ, we assume its prior distribution follows gamma distribution with parameters $\alpha, \beta > 0$, i.e., $\lambda \sim \text{Gamma}(\alpha, \beta)$ (since the range of gamma distribution is also $(0, +\infty)$, thus it's a plausible assumption). The PDF of λ is given by

$$p(\lambda \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\lambda\beta},$$

where $\Gamma(\alpha) = \int_0^{+\infty} t^{\alpha-1} e^{-t} dt,\ \alpha > 0$. Show that the posterior distribution $p(\lambda \mid \mathcal{D})$ is also a gamma distribution and identify its parameters. Hints: Feel free to drop constants. [4 points]

(b) Derive the maximum a posterior (MAP) estimation for λ under $\text{Gamma}(\alpha, \beta)$ prior. [3 points]

(c) For exponential distribution $\text{Expo}(\lambda)$, $\sum_{i=1}^n x_i \sim \text{Gamma}(n, \lambda)$ and the inverse sample mean $\frac{n}{\sum_{i=1}^n x_i}$ is the MLE for λ. Argue that whether $\frac{n-1}{n}\hat{\lambda}_{MLE}$ is unbiased ($\mathbb{E}(\frac{n-1}{n}\hat{\lambda}_{MLE}) = \lambda$). Hints: $\Gamma(z+1) = z\Gamma(z)$, $z > 0$. [3 points]

Solution: (a) $p(D \mid \lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^n x_i}$

$p(\lambda \mid \alpha, \beta) = \frac{\beta}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\lambda\beta}$

$p(D \mid \lambda)\, p(\lambda \mid \alpha, \beta) = \lambda^{n+\alpha-1} e^{-\lambda(\sum_{i=1}^n x_i + \beta)} = \lambda^{\alpha'-1} e^{-\lambda\beta'}$

So $\alpha' = n + \alpha$, $\beta' = \sum_{i=1}^n x_i + \beta$

The posterior distribution $p(\lambda \mid D)$ is also a gamma distribution with parameters $\alpha' = n+\alpha$ and $\beta' = \sum_{i=1}^n x_i + \beta$

(b) Taking the derivative of the log of the posterior distribution with respect to λ and setting it to zero

$\frac{d}{d\lambda}\left[(\alpha'-1)\log\lambda - \lambda\beta'\right] = 0$

we can get : $\frac{\alpha'-1}{\lambda} - \beta' = 0$

So $\lambda_{MAP} = \frac{\alpha'-1}{\beta'} = \frac{n+\alpha-1}{\sum_{i=1}^n x_i + \beta}$

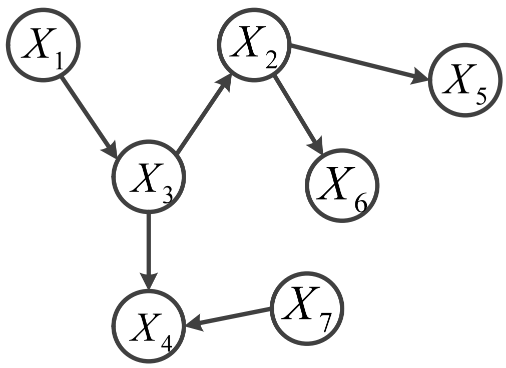
(c). $E\left(\frac{n-1}{n}\hat{\lambda}_{MLE}\right) = \frac{n-1}{n} E(\hat{\lambda}_{MLE}) = \frac{n-1}{n} \cdot E\left(\frac{n}{\sum_{i=1}^n x_i}\right)$

Since $\sum_{i=1}^n x_i \sim \text{Gamma}(n, \lambda)$, $E(\sum_{i=1}^n x_i) = n E(X_1) = n \cdot \frac{1}{\lambda}$

So $E\left(\frac{n-1}{n}\hat{\lambda}_{MLE}\right) = \frac{n-1}{n} \cdot n \cdot \frac{1}{\frac{n}{\lambda}} = \frac{n-1}{n}\lambda \neq \lambda$

Thus $\frac{n-1}{n}\hat{\lambda}_{MLE}$ is not unbiased.

4. [10 points] [Graphical Models] Given the following Bayesian Network,



answer the following questions.

(a) Factorize the joint distribution of $X_1, \cdots, X_7$ according to the given Bayesian Network. [2 points]

(b) Justify whether $X_1 \perp X_5 \mid X_2$? [2 points]

(c) Justify whether $X_5 \perp X_7 \mid X_3, X_4$? [2 points]

(d) Justify whether $X_5 \perp X_7 \mid X_4$? [2 points]

(e) Write down the variables that are in the Markov blanket of X_3. [2 points]

Solution:

(a) $P(X_1, \cdots, X_7) = P(X_1)\, P(X_2 \mid X_3)\, P(X_3 \mid X_1)\, P(X_4 \mid X_3, X_7)\, P(X_5 \mid X_2)\, P(X_6 \mid X_2)\, P(X_7)$

(b) YES

(c) YES

(d) NO

(e) X_1, X_2, X_4, X_7