Introduction to Machine Learning, Fall 2023
Homework 2

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

November 14, 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 $f(\mathbb{E}(X)) \le \mathbb{E}(f(X))$ if f is convex for any random variable X. Proof the log sum inequality:

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

   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 $f(x)$ is p-norm, so $f(\theta x) = \theta f(x)$, and $f(x+y) \le f(x) + f(y)$
then, $\theta f(x_1) + (1-\theta) f(x_2) = f(\theta x_1) + f((1-\theta)x_2) \ge f(\theta x_1 + (1-\theta)x_2)$
so, the norm is convex.

(b)

$$\frac{\partial^2 f}{\partial x_1^2} = \frac{2}{x_2}$$

$$\frac{\partial^2 f}{\partial x_1 x_2} = -\frac{2x_1}{x_2^2}$$

$$\frac{\partial^2 f}{\partial x_2^2} = \frac{2x_1^2}{x_2^3}$$

$$\nabla^2 f = \begin{bmatrix} \frac{2}{x_2} & -\frac{2x_1}{x_2^2} \\ -\frac{2x_1}{x_2^2} & \frac{2x_1^2}{x_2^3} \end{bmatrix}$$

$$y^T \nabla^2 f y = \frac{2y_1^2}{x_2} - \frac{4x_1y_1y_2}{x_2^2} + \frac{2x_1^2y_2^2}{x_2^3} = \frac{2}{x_2^3}(x_2y_1 - x_1y_2)^2 \ge 0$$

then, it is convex

(c)

$$\frac{\partial^2 f}{\partial x_1^2} = 0$$

$$\frac{\partial^2 f}{\partial x_1 x_2} = -\frac{1}{x_2^2}$$

$$\frac{\partial^2 f}{\partial x_2^2} = \frac{2x_1}{x_2^3}$$

$$\nabla^2 f = \begin{bmatrix} 0 & -\frac{1}{x_2^2} \\ -\frac{1}{x_2^2} & \frac{2x_1}{x_2^3} \end{bmatrix}$$

$$y^T \nabla^2 f y = -\frac{2y_1y_2}{x_2^2} + \frac{2x_1y_2^2}{x_2^3} = \frac{2y_2}{x_2^2}(\frac{x_1y_2}{x_2} - y_1)$$

$$y^T\nabla^2 - fy = \frac{2y_1y_2}{x_2^2} - \frac{2x_1y_2^2}{x_2^3} = \frac{2y_2}{x_2^2}(y_1 - \frac{x_1y_2}{x_2})$$

when $\frac{x_1y_2}{x_2} < y_1$, $y^T\nabla^2 fy < 0$. So $y^T\nabla^2 fy$ is not semipositive definite matrix. then, the function is not convex.

when $\frac{x_1y_2}{x_2} > y_1$, $y^T\nabla^2 - fy < 0$. So $y^T\nabla^2 - fy$ is not semipositive definite matrix. then, the function is not concave.

(d) let $x_i = \frac{a_i}{b_i}$, and $P(x = x_i) = \frac{b_1}{\sum_{i=1}^n b_i}$. then,

$$E(f(x)) = \sum_{i=1}^n a_i \log \frac{a_i}{b_i}, \qquad f(E(x)) = (\sum_{i=1}^n a_i) \log \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i}$$

and because $f(x)$ is convex function, so $E(f(x)) \geq f(E(x))$, then $\sum_{i=1}^n a_i \log \frac{a_i}{b_i} \geq (\sum_{i=1}^n a_i) \log \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i}$

2. [10 points] [Linear Methods for Classification] Consider the "Multi-class Logistic Regression" algorithm. Given training set $\mathcal{D} = \{(x^i, y^i) \mid i = 1, \ldots, n\}$ 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 $\hat{\beta} = [\hat{\beta}_1, \ldots, \hat{\beta}_k] \in \mathbb{R}^{(p+1)\times k}$ 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. $\beta_2, \ldots, \beta_k$ is similar, you don't need to write them) [6 points]

Solution:

(a)

$$\min \mathbf{KL}(p(y_c^i|x^i,\beta)||p(y_t^i|x^i,\beta)) = \int p(y_c^i|x^i,\beta) \log p(y_c^i|x^i,\beta) dx_i - \int p(y_t^i|x^i,\beta) \log p(y_c^i|x^i,\beta) dx_i$$

$$= C - \int p(y_t^i|x^i,\beta) \log p(y_c^i|x^i,\beta) dx_i = C - \frac{1}{N} \sum_{i=1}^{n} \log p(y_t^i|x^i,\beta)$$

$$\Rightarrow \ell(\beta) = \sum_{i=1}^{n} \log p(y_t^i|x^i,\beta) = \sum_{i=1}^{n} \sum_{c=1}^{k} \left[y_c^i (\beta_c^\top x^i) - y_c^i \ln \left(\sum_{c'} \exp(\beta_{c'}^\top x^i) \right) \right]$$

(b)

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

$$= \sum_{i=1}^{n} \left[y_1^i x^i - y_1^i \frac{x^i \exp(\beta_1^T x^i)}{\sum_{c'} \exp(\beta_{c'}^T x^i)} \right] = \sum_{i=1}^{n} y_1^i x^i \left[1 - \frac{\exp(\beta_1^\top x^i)}{\sum_{c'} \exp(\beta_{c'}^\top x^i)} \right]$$

3. [10 points] [Probability and Estimation] Suppose $\mathcal{D} = \{x_1, x_2, \dots, x_n\}$ 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) if all the x is greater than 0:

$$P(\lambda|\mathcal{D}) = \frac{P(\mathcal{D}|\lambda)P(\lambda)}{P(D)} = P(\lambda|\alpha,\beta)\prod_{i=1}^n P(x_i|\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\lambda\beta}\prod_{i=1}^n(\lambda e^{-\lambda x_i})$$

$$= \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha+n-1}e^{-\lambda\sum_{i=1}^n x_i + \lambda\beta} = \lambda^{\alpha+n-1}e^{-\lambda\sum_{i=1}^n x_i + \lambda\beta}$$

$$= \frac{(\sum_{i=1}^n x_i + \beta)^{\alpha+n}}{\Gamma(\alpha+n)}\lambda^{\alpha+n-1}e^{-\lambda(\sum_{i=1}^n x_i + \beta)} \sim \textbf{Gamma}(\alpha+n, \sum_{i=1}^n x_i + \beta)$$

if there exist one x is not greater than 0, then $P(x|\lambda) = 0$, so the posterior $P(\lambda|\mathcal{D}) = 0$

(b)

$$\hat{\lambda}^{\text{MAP}} = \arg\max_\lambda P(\lambda|\mathcal{D})$$

for the Gamma distribution $\textbf{Gamma}(\alpha, \beta)$, it's PDF is $f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-x\beta}$, the maximum value is at

$$\frac{df}{dx} = \frac{\beta^\alpha}{\Gamma(\alpha)}((\alpha-1)x^{\alpha-2}e^{-x\beta} - \beta x^{\alpha-1}e^{-x\beta}) = 0$$

so its max point is $x = \frac{\alpha-1}{\beta}$

because $\hat{\lambda}^{\text{MAP}} \sim \textbf{Gamma}(\alpha+n, \sum_{i=1}^n x_i + \beta)$, so its max point is max $\hat{\lambda}^{\text{MAP}} = \frac{\alpha'-1}{\beta'} = \frac{\alpha+n-1}{\sum_{i=1}^n x_i + \beta}$

$$\Rightarrow \hat{\lambda}^{\text{MAP}} = \arg\max_\beta = \frac{\alpha+n-1}{\sum_{i=1}^n x_i + \beta}$$

(c)

$$\mathbb{E}(\frac{n-1}{n}\hat{\lambda}_{\text{MLE}}) = \mathbb{E}(\frac{n-1}{\sum_{i=1}^n x_i}) = (n-1)\mathbb{E}(\frac{1}{\sum_{i=1}^n x_i})$$

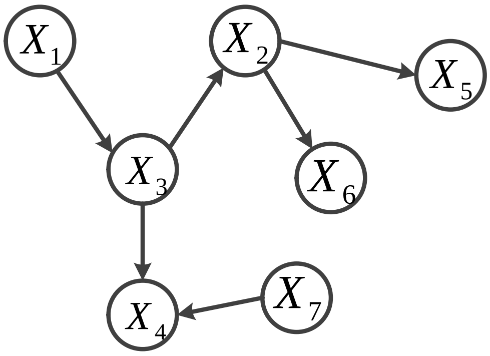
$$\sum_{i=1}^n x_i \sim \textbf{Gamma}(n,\lambda) \Rightarrow \mathbb{E}(\frac{1}{\sum_{i=1}^n x_i}) = \int_0^\infty \frac{1}{x}\cdot\frac{\lambda^n}{\Gamma(n)}x^{n-1}e^{-\lambda x}dx = \frac{\lambda^n}{\Gamma(n)}\int_0^\infty x^{n-2}e^{-x\lambda}dx$$

$$= \frac{\lambda}{\Gamma(n)}\int_0^\infty (\lambda x)^{n-1-1}e^{-\lambda x}d\lambda x = \frac{\lambda}{\Gamma(n)}\Gamma(n-1) = \frac{\lambda}{n-1}$$

$$\Rightarrow \mathbb{E}(\frac{n-1}{n}\hat{\lambda}_{\text{MLE}}) = \lambda$$

Therefore, the MLE for λ is 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)
$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7) = P(X_1)P(X_2)P(X_3|X_1)P(X_4|X_3)P(X_5|X_2)P(X_6|X_2)P(X_7)$$

Note: Different starting points may result in different final answers.

(b) yes. because X_1 to X_5 is a head-to-tail path. so given X_2, X_1 and X_5 are not connected. so $X_1 \perp\!\!\!\perp X_5|X_2$

(c) yes. if given X_3, then the path between X_5 and X_7 must be broken, then they're disconnected. so $X_5 \perp\!\!\!\perp X_7|X_3, X_4$

(d) no. because X_4 is a head-to-head node, so given X_4, the path between X_5 and X_7 is connected. so $X_5 \not\!\perp\!\!\!\perp X_7|X_4$

(e) Markov blanket is containing its co-parent, children and parent. they are X_1, X_2, X_4, X_7