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 $f(\mathbb{E}(X)) \le \mathbb{E}(f(X))$ 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) Any norm : $\mathbb{R}^n \to \mathbb{R}$ has three properties.

1° $\forall x \in \mathbb{R}^n$, $f(x) \geq 0$.

2° $\forall x \in \mathbb{R}^n$ $\forall \alpha \in \mathbb{R}^n$, $f(\alpha x) = |\alpha| f(x)$.

3° $\forall x, y \in \mathbb{R}^n$, $f(x+y) \leq f(x) + f(y)$

If we want to prove that it is convex, we need to prove that $\forall x, y$ $\forall \lambda$. $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$. $\lambda \in [0,1]$.

According to property 3°. we have. $f(\lambda x + (1-\lambda)y) \leq f(\lambda x) + f((1-\lambda)y)$

According to property 2°. we have $f(\lambda x) + f((1-\lambda)y) \leq |\lambda| f(x) + |1-\lambda| f(y)$

Because $\lambda \in [0,1]$ $\therefore$ $|\lambda| f(x) + |1-\lambda| f(y) = \lambda f(x) + (1-\lambda) f(y)$.

$\therefore$ $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$ #

(b) consider the Heissen matrix $\nabla^2 f(x)$,

$\nabla^2 f(x)$ is semi positive $\Longleftrightarrow$ $f(x)$ convex.

$$\frac{\partial f(x_1, x_2)}{\partial x_1} = \frac{\partial \frac{x_1^2}{x_2}}{\partial x_1} = \frac{2x_1}{x_2}.$$

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

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

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

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

$$\det\left(\nabla^2 f(x) - \lambda I\right) = \begin{vmatrix} \frac{2}{x_2} - \lambda & \frac{-2x_1}{x_2^2} \\ \frac{-2x_1}{x_2^2} & \frac{2x_1^2}{x_2^3} - \lambda \end{vmatrix}$$

$$= \left(\frac{2}{x_2} - \lambda\right)\left(\frac{2x_1^2}{x_2^3} - \lambda\right) - \left(\frac{-2x_1}{x_2^2}\right)^2$$

$$= \frac{4x_1^2}{x_2^4} - \lambda\left(\frac{2}{x_2} + \frac{2x_1^2}{x_2^3}\right) - \frac{4x_1^2}{x_2^4} + \lambda^2 = 0.$$

$$\lambda = \frac{2}{x_2} + \frac{2x_1^2}{x_2^3} \quad \because x_2 > 0 \quad \therefore \lambda > 0$$

$\therefore$ $\nabla^2 f(x)$ is semi positive

$\therefore$ $f(x_1, x_2)$ on $\mathbb{R} \times \mathbb{R}_{>0}$ is convex.

(c). consider the Heissen matrix $\nabla^2 f(x)$,

$\nabla^2 f(x)$ is semi positive $\Longleftrightarrow$ $f(x)$ convex.

$$\frac{\partial f(x_1, x_2)}{\partial x_1} = \frac{\partial \frac{x_1}{x_2}}{\partial x_1} = \frac{1}{x_2}.$$

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

$$\frac{\partial f(x_1, x_2)}{\partial x_1^2} = 0$$

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

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

$$\det\left(\nabla^2 f(x) - \lambda I\right) = \begin{vmatrix} 0 - \lambda & -\frac{1}{x_2^2} \\ -\frac{1}{x_2^2} & \frac{2x_1}{x_2^3} - \lambda \end{vmatrix}$$

$$= -\lambda\left(\frac{2x_1}{x_2^3} - \lambda\right) - \left(\frac{-1}{x_2^2}\right)^2$$

$$= \lambda^2 - \lambda\frac{2x_1}{x_2^3} - \frac{1}{x_2^4} = 0.$$

$$\Delta = \frac{4x_1^2}{x_2^6} + 4\frac{1}{x_2^4} \qquad \lambda_1 \cdot \lambda_2 = -\frac{1}{x_2^4} < 0.$$

$\therefore$ $\nabla^2 f(x)$ is not semi positive

$\therefore$ $f(x_1, x_2)$ on $\mathbb{R} \times \mathbb{R}_{>0}$ is not. convex.

(d). According to Jason inequality, $f(E(X)) \leq E(f(X))$ if f is convex

Let $\lambda \geq 0$. and $\sum_{i=1}^{n} \lambda_i = 1$, then $\quad f\left(\sum_{i=1}^{n} \lambda_i x_i\right) \leq \sum_{i=1}^{n} \lambda_i f(x_i)$

Let $x_i = \frac{a_i}{b_i}$ and $\lambda_i = \frac{b_i}{\sum_{i=1}^{n} b_i}$

then $$f\left(\sum_{i=1}^{n} \frac{b_i}{\sum_{i=1}^{n} b_i} \frac{a_i}{b_i}\right) \leq \sum_{i=1}^{n} \frac{b_i}{\sum_{i=1}^{n} b_i} \frac{a_i}{b_i} \log\left(\frac{a_i}{b_i}\right).$$

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

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

$$\therefore \left(\sum_{i=1}^{n} a_i\right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} \le \sum_{i=1}^{n} a_i \log \frac{a_i}{b_i}. \qquad \#$$

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). $l(\beta) = \ln \prod_{i=1}^n p(y_t^i \mid x^i; \beta)$

$$= \ln \prod_{i=1}^n \frac{\exp(\beta_c^T x^i)}{\sum_{c'} \exp(\beta_{c'}^T x^i)}$$

$$= \sum_{i=1}^n \ln \frac{\exp(\beta_c^T x^i)}{\sum_{c'} \exp(\beta_{c'}^T x^i)}$$

$$= \sum_{i=1}^n \left(\beta_c^T x^i - \ln\left(\sum_{c'} \beta_{c'}^T x^i\right)\right)$$

$\because y^i \in \mathbb{R}^k$ is a one-hot binary vector.

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

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

$$= \sum_{i=1}^n \left[y_j^i x^i - \sum_{c=1}^k y_1^i \cdot \frac{\exp(\beta_j^T x^i)\, x^i}{\sum_{c'} \exp(\beta_{c'}^T x^i)} \right]$$

$$\therefore \nabla_{\beta_1} l(\beta) = \sum_{i=1}^n \left[y_1^i x_i - \sum_{c=1}^k y_1^i \cdot \frac{\exp(\beta_j^T x^i) x^i}{\sum_{c'} \exp(\beta_{c'}^T x^i)} \right]$$

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

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

$\because X \sim \text{Expo}(\lambda)$ 且 $\lambda > 0$.

$\therefore p(x|\lambda) = \lambda e^{-\lambda x}$, $x > 0$.

求后验

(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

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

$f(x|\lambda) = \lambda e^{-\lambda x}$, $x > 0$

$$f\left(\sum_{n=1}^{n} x_i \,\middle|\, n, \lambda\right) = \frac{\lambda^n \left(\sum_{n=1}^{n} x_i\right)^{n-1} e^{-\lambda \sum_{n=1}^{n} x_i}}{(n-1)!}$$

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]

找到使后验分布最大化的 λ 值

(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). $p(\lambda|D) = \dfrac{P(D|\lambda)P(\lambda)}{P(D)} \propto \underbrace{P(D|\lambda)}_{①}\underbrace{P(\lambda)}_{②}$. (drop constant)

① 其中 $P(D|\lambda)$ 是给定 λ 下观察到数据 D 的概率.

$\because P(x|\lambda) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & \text{otherwise.} \end{cases}$ 且 $D = \{x_1, x_2 \ldots x_n\}$ and iid.

$\therefore P(D|\lambda) = \lambda^n e^{-\lambda(x_1 + \cdots + x_n)} = \lambda^n e^{-\lambda \sum_{i=1}^n x_i}$

② 其中 $P(\lambda)$ 可以得到 $\lambda \sim \text{Gamma}(\alpha, \beta)$.

$P(\lambda|\alpha, \beta)$ 表示没有观测到数据前对参数 λ 的分布.

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

$$\therefore P(\lambda|D) \propto \lambda^n e^{-\lambda \sum_{i=1}^n x_i} \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\lambda\beta}$$

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

Gamma 分布形式是

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

令 $\alpha' = n + \alpha$, $\beta' = \beta + \sum_{i=1}^n x_i$ $\therefore P(\lambda|D) \sim \text{Gamma}(\alpha', \beta')$

$\therefore P(\lambda|D) \sim \text{Gamma}(\alpha + n, \beta + \sum_{i=1}^n x_i)$

(b). 后验分布 $p(\lambda|D,\alpha,\beta) = \frac{P(D|\lambda)P(\lambda|\alpha,\beta)}{P(D)} \propto P(D|\lambda)P(\lambda|\alpha,\beta)$.

$$P(\lambda, D|\alpha,\beta) = P(D|\lambda,\alpha,\beta)\,P(\lambda|\alpha,\beta)$$
$$= P(D|\lambda)\,P(\lambda|\alpha,\beta).$$

$\therefore \log P(\lambda|D,\alpha,\beta) \propto \log P(D|\lambda) + \log P(\lambda|\alpha,\beta)$.

$\because P(D|\lambda) \propto \lambda^n e^{-\lambda\sum_{n=1}^{n} x_i}$

$P(\lambda|\alpha,\beta) \propto \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\lambda\beta}$

$\therefore \log P(\lambda|D,\alpha,\beta) \propto n\log\lambda - \lambda\sum_{n=1}^{n} x_i + (\alpha-1)\log\lambda - \lambda\beta$

to find the max λ, we let $\frac{d\log P(\lambda|D,\alpha,\beta)}{d\lambda} = 0$

$$\frac{d\log P(\lambda|D,\alpha,\beta)}{d\lambda} = \frac{n}{\lambda} - \sum_{n=1}^{n} x_i + \frac{\alpha-1}{\lambda} - \beta$$

$$= \frac{n+\alpha-1}{\lambda} - \sum_{n=1}^{n} x_i + \beta = 0.$$

$$\lambda = \frac{n+\alpha-1}{\sum_{n=1}^{n} x_i - \beta}$$

(C), $\because \sum_{x=1}^{n} x_i \sim \text{Gamma}(n, \lambda)$ (Gamma 函数. $\Gamma(\alpha) = (\alpha-1)!$).

$\therefore f(\sum_{i=1}^{n} x_i) = \frac{\lambda^n}{\Gamma(n)}(\sum_{i=1}^{n} x_i)^{n-1} e^{-\lambda(\sum_{i=1}^{n} x_i)}$

$$E\left(\frac{1}{\sum_{n=1}^{n} x_i}\right) = \int_0^\infty \frac{1}{\sum_{i=1}^{n} x_i} f(\sum_{i=1}^{n} x_i)\,dx.$$

$$= \int_0^\infty \frac{1}{\sum_{i=1}^{n} x_i}\frac{\lambda^n}{\Gamma(n)}(\sum_{i=1}^{n} x_i)^{n-1} e^{-\lambda(\sum_{i=1}^{n} x_i)}$$

$$= \int_0^\infty \frac{\lambda^n}{\Gamma(n)}(\sum_{i=1}^{n} x_i)^{n-2} e^{-\lambda(\sum_{i=1}^{n} x_i)}.$$

$$= \int_0^\infty \frac{\lambda \cdot \lambda^{n-1}}{(n-1)\Gamma(n-1)} \left(\sum_{i=1}^{n} x_i\right)^{n-2} e^{-\lambda\left(\sum_{i=1}^{n} x_i\right)}.$$

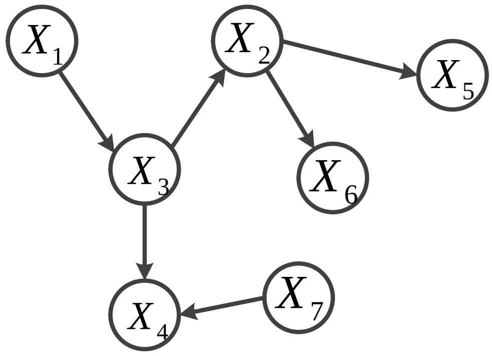
$$= \frac{\lambda}{n-1} \int_0^\infty \frac{\lambda^{n-1}}{\Gamma(n-1)} \left(\sum_{i=1}^{n} x_i\right)^{n-2} e^{-\lambda\left(\sum_{i=1}^{n} x_i\right)}$$

$$= \frac{\lambda}{n-1}$$

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

$\therefore \frac{n-1}{n}\hat{\lambda}_{MLE}$ 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) $P(X_1 \dots X_7) = P(X_1)P(X_3|X_1)P(X_2|X_3)P(X_6|X_2)$
$P(X_5|X_2)P(X_4|X_3, X_7)P(X_7)$.

(b). Yes. if given X_2, X_3 and X_5 are conditionally independent.

$$P(X_5|X_3, X_2) = \frac{P(X_3, X_5, X_2)}{P(X_2, X_3)}$$

$$= \frac{P(X_3)P(X_2|X_3)P(X_5|X_2)}{P(X_3)P(X_2|X_3)}.$$

$$= P(X_5|X_2).$$

∴ X_1 to X_5 is blocked, because the path contains a single inactive triple.

∴ $X_1 \perp\!\!\!\perp X_5 | X_2$.

(C) Yes

We want to prove that $P(X_5 \mid X_7, X_3, X_4) = P(X_5 \mid X_3, X_4)$

$$P(X_5 \mid X_7, X_3, X_4) = \frac{\sum_{x_2} P(X_5, X_7, X_3, X_4, X_2)}{P(X_3, X_4, X_7)}$$

$$= \frac{\sum_{x_2} P(X_2 \mid X_3)\, \cancel{P(X_3)}\, \cancel{P(X_4 \mid X_3, X_7)}\, \cancel{P(X_7)}\, P(X_5 \mid X_2)}{\cancel{P(X_3)}\, \cancel{P(X_4)}\, \cancel{P(X_7)}}$$

$$= \sum_{x_2} P(X_2 \mid X_3)\, P(X_5 \mid X_2),$$

$$= P(X_2 = 0 \mid X_3)\, P(X_5 \mid X_2 = 0) + P(X_2 = 1 \mid X_3)\, P(X_5 \mid X_2 = 1)$$

P_5 has no ~~deal~~ with P_7.

$\therefore$ $P_5 \perp\!\!\!\perp P_7 \mid X_3, X_4$.

X_4 is given, so that X_7 and X_3 are dependent.

X_3 is given, so that X_4 and X_2 are independent.

$\therefore$ X_4 to X_5 is blocked, because the path contains a single inactive triple.

d. No.

X_4 is given. X_7 and X_3 are dependent.

X_3 and X_5 are dependent.

$\therefore$ Given X_4. X_7 and X_5 are dependent.

(e) X_1, X_2, X_4, X_7.