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) for any norm we always have. $\begin{cases} \|x\| \ge 0. \ \|x\| = 0 \text{ iff } x = 0 & ① \\ \|cx\| = |c|\,\|x\| & ② \\ \|x+y\| \le \|x\| + \|y\| & ③ \end{cases}$

$\theta f(x) + (1-\theta) f(y) \ge f(\theta x + (1-\theta) y)$

for $x, y \in \text{dom}$, we have θx, $(1-\theta)y$ also $\in$ dom, $\theta \in [0,1]$

with property ③ we have:

$f(\theta x + (1-\theta) y) \le f(\theta x) + f((1-\theta) y)$

$\Rightarrow f(\theta x + (1-\theta) y) \le \theta f(x) + (1-\theta) f(y)$ property ④

so according to definition of convex function, we proved.

(b) $H = f'' = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} \end{bmatrix} = \begin{bmatrix} \frac{2}{x_2} & -2 x_1 x_2^{-2} \\ -2 x_1 x_2^{-2} & 2 x_1^2 x_2^{-3} \end{bmatrix}$

$|H| = 4 x_1^2 \cdot x_2^{-4} - 4 x_1^2 x_2^{-4} = 0 \qquad \frac{2}{x_2} > 0.$

$\because$ H is symmetric and the determinant is 0 so is semi positive definite matrix So convex

(c)

$$|H| = \begin{vmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} \end{vmatrix} = \begin{vmatrix} 0 & -x_2^{-2} \\ -x_2^{-2} & 2x_1 x_2^{-3} \end{vmatrix} = -x_2^{-4} \leq 0.$$

so the Hessian matrix is not positive definite

which shows the function not convex

(d)

$$E(x) = \frac{b_i}{\sum_{i=1}^{n} b_i} \frac{a_i}{b_i} \quad \text{and} \quad f(x) = x \cdot \log x$$

we find. $f''(x) \geq 0$ $f(x)$ convex

$\Rightarrow$ let $\lambda_i = \frac{b_i}{\sum b_i}$ $x_i = \frac{a_i}{b_i}$ $\because \sum_{i=1}^{n} \lambda_i = 1$

$\therefore$ we have $f\left(\sum_1^n \lambda_i (x_i)\right) \leq \sum_1^n \lambda_i f(x_i)$

$$\Rightarrow \sum_1^n \frac{b_i}{\sum b_i} \cdot \frac{a_i}{b_i} \cdot \log \sum_1^n \frac{b_i}{\sum b_i} \frac{a_i}{b_i} \leq \sum_1^n \frac{b_i}{\sum b_i} \cdot \frac{a_i}{b_i} \log \frac{a_i}{b_i}$$

$$\Rightarrow \sum_1^n a_i \cdot \log \frac{\sum a_i}{\sum b_i} \leq \left(\sum_1^n a_i\right) \log \frac{a_i}{b_i}$$

$$\Rightarrow \sum_1^n a_i \cdot \log \frac{\sum a_i}{\sum b_i} \leq \left(\sum_1^n a_i\right) \log \frac{a_i}{b_i}$$

prove

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) $L(\beta) = \prod_{i=1}^{n} \prod_{c=1}^{k} \left(p(y_c^i \mid x^i, \beta) \right)^{y_c^i} = \prod_{i=1}^{n} \prod_{c=1}^{k} \left[\frac{\exp(\beta_c^\top x^i)}{\sum_{c'} \exp(\beta_{c'}^\top x^i)} \right]^{y_c^i}$

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

$= \sum_{i=1}^{n} \sum_{c=1}^{k} \left(\ln \exp(\beta_c^\top x^i) \cdot y_c^i - y_c^i \ln \left(\sum_{c'} \exp(\beta_{c'}^\top x^i) \right) \right)$

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

(b) $\sum_{i=1}^{n} y_1^i (\beta_1 \cdot x^{i\top}) - y_1^i \cdot \frac{\exp(\beta_1^\top x^i) \cdot x^{i\top}}{\sum_{c'} \exp(\beta_{c'}^\top x^i)}$

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) $P(\lambda \mid D) = \dfrac{P(D \mid \lambda) P(\lambda)}{P(D)}$

$P(D) = \int_{-\infty}^{+\infty} P(D, \lambda) \cdot d\lambda$

$\Rightarrow P(\lambda \mid D) = \dfrac{\lambda^n e^{-\lambda(\sum_{i=1}^n x_i)} \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\lambda\beta}}{\int_{-\infty}^{+\infty} \lambda^n e^{-\lambda(\sum_{i=1}^n x_i)} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\lambda\beta} \cdot d\lambda}$

$= \dfrac{\lambda^{n+\alpha-1} \cdot e^{-\lambda(\sum_{i=1}^n x_i + \beta)}}{\int_{-\infty}^{+\infty} \lambda^{n+\alpha-1} e^{-\lambda(\sum_{i=1}^n x_i + \beta)} \cdot d\lambda}$

for $\int_{-\infty}^{+\infty} \lambda^{n+\alpha-1} e^{-\lambda(\sum_{i=1}^n x_i + \beta)} \cdot d\lambda$

$\Rightarrow \int_{-\infty}^{+\infty} \left(\dfrac{t}{\sum_{i=1}^n x_i + \beta}\right)^{n+\alpha-1} \cdot e^{-t} \cdot \dfrac{1}{\sum_{i=1}^n x_i + \beta} \cdot dt$

$\Rightarrow \left(\dfrac{1}{\sum_{i=1}^n x_i + \beta}\right)^{n+\alpha} \cdot \int_{-\infty}^{+\infty} t^{n+\alpha-1} \cdot e^{-t} \cdot dt$

so for $P(\lambda \mid D)$ we have $\dfrac{(\sum_{i=1}^n x_i + \beta)^{n+\alpha} \cdot \lambda^{n+\alpha-1} \cdot e^{-\lambda(\sum_{i=1}^n x_i + \beta)}}{\Gamma(\alpha)}$

so $\alpha' = n + \alpha \quad \beta' = \sum_{i=1}^n x_i + \beta$

we have $\text{Gamma}(\alpha', \beta')$

(b)

$$0 = p(\lambda|D)' = \frac{(\beta+\sum_{i=1}^{n} x_i)^{n+\alpha}}{\Gamma(n+\alpha)} \cdot \left(\lambda^{n+\alpha-2}\cdot(n+\alpha-1)\cdot e^{-(\beta+\sum_{i=1}^{n} x_i)\cdot\lambda} + \lambda^{n+\alpha-1}\cdot(-\beta-\sum_{i=1}^{n} x_i)\cdot e^{-(\beta+\sum_{i=1}^{n} x_i)\cdot\lambda}\right)$$

$$= (n+\alpha-1)\cdot\lambda^{n+\alpha-2} + \lambda^{n+\alpha-1}\cdot(-\beta-\sum_{i=1}^{n} x_i) = 0$$

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

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

(c) $\dfrac{\wedge}{\sum_{i=1}^{n} x_i} = E(\hat{\lambda}_{MLE})$

$\because \sum_{i=1}^{n} x_i \sim Gamma(n, \lambda)$

$$\therefore E\left[\sum_{i=1}^{n} x_i\right] = \int_0^{\infty} \frac{\lambda^n}{\Gamma(n)} x^{n-1}\cdot e^{-\lambda x} \div x \div x\, dx = \frac{\lambda\Gamma(n-1)}{\Gamma(n)}\int_0^{\infty} \frac{\lambda^{n-1}}{\Gamma(n-1)}\cdot x^{n-2}\cdot e^{-\lambda x}\cdot dx$$

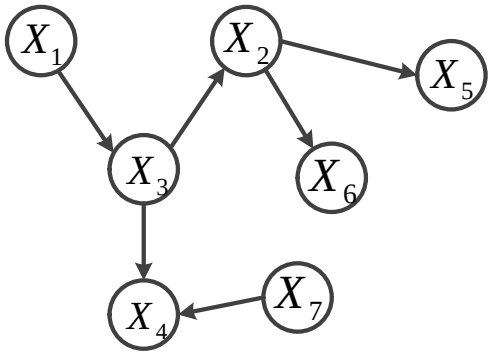
$$= \frac{\lambda}{n-1}$$

$$\therefore E\left[\frac{n}{\sum_{i=1}^{n} x_i}\right] = \frac{n\lambda}{n-1}$$

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

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

(b) $X_1 \to X_3 \to X_2$ is active triple. $\Rightarrow$ block
$X_3 \to X_2 \to X_5$ is inactive triple.
Yes

(c) $X_5 \perp X_7 \mid X_3, X_4$
$X_3 \to X_2 \to X_5$ is active triple
$X_4 \leftarrow X_3 \to X_2$ is inactive triple $\Rightarrow$ block
$X_5 \to X_4 \leftarrow X_7$ is active
Yes

(d) No
$X_3 \to X_2 \to X_5$ active.
$X_4 \leftarrow X_3 \to X_2$ active $\Rightarrow$ unblock.
$X_3 \to X_4 \leftarrow X_7$ active

(e) X_1 X_2 X_4 X_7