Introduction to Machine Learning, Fall 2023
Homework 2
(Due Tuesday Nov. 14 at 11:59pm (CST))

December 3, 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) *Proof.* $\forall x, y \in \mathbf{dom} f,\ 0 \leq \theta \leq 1$
$$f(\theta x + (1-\theta)y) \leq f(\theta x) + f((1-\theta)y) = \theta f(x) + (1-\theta) f(y).$$
The inequality follows from triangle inequality and the equality follows from homogeneity of norm. □

(b) The Hessian of f is
$$\nabla^2 f(x_1, x_2) = \begin{bmatrix} \frac{2}{x_2} & -\frac{2x_1}{x_2^2} \\ -\frac{2x_1}{x_2^2} & \frac{2x_1^2}{x_2^3} \end{bmatrix} \succeq 0.$$
The Hessian is positive semidefinite (By Sylvester's criterion, which states that a Hermitian matrix is positive semidefinite if and only if all its principle minors are nonnegative. In this case, let denote determinant by $|\cdot|$, then $|\ \frac{2}{x_2}\ | \geq 0$, $|\ \frac{2x_1^2}{x_2^3}\ | \geq 0$ and $|\ \nabla^2 f(x_1, x_2)\ | = \frac{4x_1^2}{x_2^4} - \frac{4x_1^2}{x_2^4} = 0$. You can also check all the eigenvalues of the Hessian is nonnegative.), hence f is convex.

(c) The Hessian of f is
$$\nabla^2 f(x_1, x_2) = \begin{bmatrix} 0 & -\frac{1}{x_2^2} \\ -\frac{1}{x_2^2} & \frac{2x_1}{x_2^3} \end{bmatrix}.$$
The Hessian is indefinite (Use Sylvester's criterion to determine the Hessian is neither positive semidefinite nor negative semidefinite. Note that to determine the negative semidefiniteness, the necessary and sufficient condition becomes for every principle minor $\mathbf{M}_k$ of size k, $(-1)^k\ |\ \mathbf{M}_k\ | \geq 0$. You can also check eigenvalues of the Hessian have mixed sign.), hence f is neither convex nor concave.

(d) *Proof.* By Jensen's inequality applied to the convex function $f(x) = x \log x$, we have
$$\sum_i z_i f(t_i) \geq f\left(\sum_i z_i t_i\right)$$
where $z_i \geq 0$ and $\sum_i z_i = 1$. Setting $z_i = \frac{b_i}{\sum_j b_j}$ and $t_i = \frac{a_i}{b_i}$, we obtain
$$\sum_i \frac{a_i}{\sum b_j} \log \frac{a_i}{b_i} \geq \sum_i \frac{a_i}{\sum b_j} \log \sum_i \frac{a_i}{\sum b_j},$$
which is the log sum inequality. □

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) Since y^i is a one-hot binary vector, it follows that

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

(b) Note the first term in the summation is linear in β_1, then we have

$$\nabla_{\beta_1} \sum_{i=1}^{n} \sum_{c=1}^{k} y_c^i (\beta_c^\top x^i) = \sum_{i=1}^{n} y_1^i x^i.$$

For fixed $i \in [n]$, applying chain rule, we have

$$\begin{aligned}
&\nabla_{\beta_1} \sum_{c=1}^{k} y_c^i \ln \left(\sum_{c'} \exp(\beta_{c'}^\top x^i) \right) = \nabla_{\beta_1} y_t^i \ln \left(\sum_{c'} \exp(\beta_{c'}^\top x^i) \right) \\
=&\nabla_{\beta_1} \ln \left(\sum_{c'} \exp(\beta_{c'}^\top x^i) \right) = \frac{\exp(\beta_1^\top x^i)}{\sum_{c'} \exp(\beta_{c'}^\top x^i)} x^i = p(y_1^i \mid x^i; \beta) x^i.
\end{aligned}$$

Combining both terms, we derive

$$\nabla_{\beta_1} \ell(\beta) = \sum_{i=1}^{n} (y_1^i - p(y_1^i \mid x^i; \beta)) x^i.$$

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) Recall that the posterior is proportional to likelihood times prior, we have

$$\begin{aligned} p(\lambda \mid \mathcal{D}) &\propto p(\mathcal{D} \mid \lambda) p(\lambda) \\ &= \left(\prod_{i=1}^n \lambda e^{-\lambda x_i} \right) \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\lambda\beta} \\ &\propto \lambda^n e^{-\lambda \sum_{i=1}^n x_i} \lambda^{\alpha-1} e^{-\lambda\beta} \\ &= \lambda^{(n+\alpha)-1} e^{-\lambda(\beta + \sum_{i=1}^n x_i)}. \end{aligned}$$

Since the last term is proportional to the PDF of $\text{Gamma}(n+\alpha, \beta + \sum_{i=1}^n x_i)$; therefore, we must have $\lambda \mid \mathcal{D} \sim \text{Gamma}(n+\alpha, \beta + \sum_{i=1}^n x_i)$.

(b) The maximum a posterior (MAP) estimation for λ is given by

$$\hat{\lambda}_{MAP} = \operatorname*{argmin}_\lambda \ p(\lambda \mid \mathcal{D}) = \operatorname*{argmin}_\lambda \ \ln p(\lambda \mid \mathcal{D}).$$

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

$$\ln p(\lambda \mid \mathcal{D}) = ((n+\alpha) - 1) \ln \lambda - \lambda \left(\beta + \sum_{i=1}^n x_i \right) + C,$$

where C is a constant w.r.t. λ. Take $\nabla_\lambda \ln p(\lambda \mid \mathcal{D})$ and set it to 0, it follows that

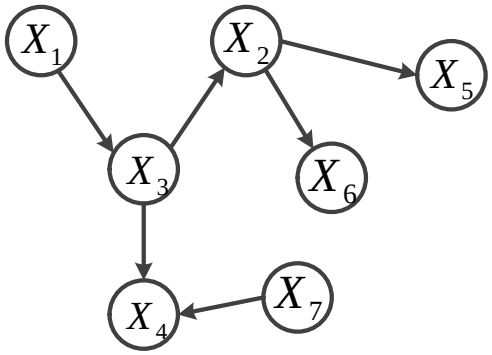
$$\frac{(n+\alpha) - 1}{\hat{\lambda}_{MAP}} - \left(\beta + \sum_{i=1}^n x_i \right) = 0 \implies \hat{\lambda}_{MAP} = \frac{n + (\alpha - 1)}{\beta + \sum_{i=1}^n x_i}.$$

(c) $\frac{n-1}{n}\hat{\lambda}_{MLE}$ is unbiased. Let z denote $\sum_{i=1}^n x_i$, since

$$\begin{aligned} \mathbb{E}(\hat{\lambda}_{MLE}) = \mathbb{E}(\frac{n}{z}) &= \int_0^{+\infty} \frac{n}{z} \frac{\lambda^n}{\Gamma(n)} z^{n-1} e^{-\lambda z} dz = n\lambda \frac{\Gamma(n-1)}{\Gamma(n)} \int_0^{+\infty} \frac{\lambda^{n-1}}{\Gamma(n-1)} z^{n-1-1} e^{-\lambda z} dz \\ &= n\lambda \frac{\Gamma(n-1)}{\Gamma(n)} = \frac{n}{n-1}\lambda, \end{aligned}$$

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

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) The joint distribution can be factorized as:

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

(b) **Yes**. There is only one path from X_1 to X_5. X_2 is along the path and observed. The arrows on X_2 is head-to-tail; thus, the path is blocked.

(c) **Yes**. There is only one path from X_7 to X_5. X_3 is along the path and observed. The path is blocked by X_3, since it's a tail-to-tail node.

(d) **No**. Similar to (c), but X_3 is unobserved. Therefore, the path is unblocked.

(e) X_1, X_2, X_4 and X_7 are in the Markov blanket of X_3.