

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) Prove any norm $f : \mathbb{R}^n \rightarrow \mathbb{R}$ 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 $\mathbb{R}_{>0}^2$. [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 . Prove 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, \dots, a_n and b_1, \dots, b_n are positive numbers. Hints: $f(x) = x \log x$ is strictly convex. [4 points]

Solution:

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. β_2, \dots, β_k is similar, you don't need to write them) [6 points]

Solution:

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

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

- (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

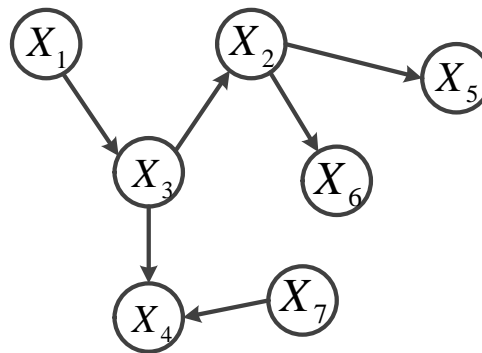
$$p(\lambda | \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\lambda\beta},$$

where $\Gamma(\alpha) = \int_0^{+\infty} t^{\alpha-1} e^{-t} dt$, $\alpha > 0$. Show that the posterior distribution $p(\lambda | \mathcal{D})$ 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:

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



answer the following questions.

- (a) Factorize the joint distribution of X_1, \dots, 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: