WRITTEN EXERCISES

1. Maximum Likelihood and KL Divergence. In machine learning, we often need to assess the similarity between pairs of distributions. This is often done using the Kullback-Leibler (KL) divergence:

$$KL(p(x)||q(x)) = \mathbb{E}_{p(x)}\left[\log p(x) - \log q(x)\right]$$

The KL divergence is always non-negative, and equals zero when p and q are identical. This makes it a natural tool for comparing distributions.

This question explores connections between the KL divergence and maximum likelihood learning. Suppose we want to learn a supervised probabilistic model $p_\theta(y|x)$ (e.g., logistic regression) with parameters θ over a dataset $\mathcal{D} = \{(x^{(i)}, y^{(i)}) \mid i = 1, 2, ..., n\}$. Let $\hat{p}(x, y)$ denote the *empirical* distribution of the data, which is the distribution that assigns a probability of $1/n$ to each of the data points in $\mathcal{D}$ (and zero to all the other possible (x, y)):

$$\hat{p}(x,y) = \begin{cases} \frac{1}{n} & \text{if } (x,y) \in \mathcal{D} \\ 0 & \text{otherwise.} \end{cases}$$

The empirical distribution can be seen as a guess of the true data distribution from which the dataset $\mathcal{D}$ was sampled i.i.d.: it assigns a uniform probability to every possible training instance seen so far, and does not assign any probability to unseen training instances.

Prove that selecting parameters θ by maximizing the likelihood is equivalent to selecting θ that minimize the average KL divergence between the data distribution and the model distribution:

$$\arg\max_\theta \mathbb{E}_{\hat{p}(x,y)}\left[\log p_\theta(y|x)\right] = \arg\min_\theta \mathbb{E}_{\hat{p}(x)}\left[KL(\hat{p}(y|x)||p_\theta(y|x))\right].$$

Here, $\mathbb{E}_{p(x)} f(x)$ denotes $\sum_{x\in\mathcal{X}} f(x)p(x)$ if x is discrete and $\int_{x\in\mathcal{X}} f(x)p(x)dx$ if x is continuous.

$$\arg\max_\theta E_{\hat{p}(x,y)}[\log p_\theta(y|x)] = \arg\min_\theta E_{\hat{p}(x)}[KL(\hat{p}(y|x)||p_\theta(y|x)]$$

$$RHS = \arg\min_\theta E_{\hat{p}(x)}[E_{\hat{p}(y|x)}[\log p(y|x) - \log p_\theta(y|x)]]$$

$$= \arg\min_\theta E_{\hat{p}(x)}[E_{p(y|x)} \log \hat{p}(y|x) - E_{p(y|x)} \log p_\theta(y|x)]$$

$$= \arg\min_\theta E_{\hat{p}(x)}\left[\sum_{(x,y)\in X} \log \hat{p}(y|x) \cdot \frac{P(x,y)}{P(x)} - \sum_{(x,y)\in X} \log p_\theta(y|x) \cdot \frac{P(x,y)}{P(x)}\right]$$

$$= \arg\min_\theta E_{\hat{p}(x)}\left[\left(\sum_y \log \hat{p}(y|x) - \sum_y \log p_\theta(y|x)\right) \cdot \frac{P(x,y)}{P(x)}\right]$$

$$= \arg\min_\theta \sum_x \sum_y (\log \hat{p}(y|x) - \log p_\theta(y|x)) \cdot \frac{P(x,y)}{\cancel{P(x)}} \cdot \cancel{P(x)}$$

$$= \arg\min_\theta \sum_x \sum_y (\log \hat{p}(y|x) - \log p_\theta(y|x) \cdot P(x,y))$$

$\because$ θ is not dependent on $\log \hat{p}(y|x)$

$\therefore$ $LHS = \arg\min_{\theta} \sum_x \sum_y (-\log P_\theta(y|x) \cdot \hat{P}(x,y))$

$= \arg\min_{\theta} E_{\hat{p}(x,y)}(-\log p_\theta(y|x))$

$= \arg\max_{\theta} E_{\hat{p}(x,y)}[\log p_\theta(y|x)]$

$\therefore$ $RHS = LHS$.

2. Gradient and log-likelihood for logistic regression.

(a) Let $\sigma(a) = \dfrac{1}{1+e^{-a}}$ be the sigmoid function. Show that $\dfrac{d\sigma(a)}{da} = \sigma(a)(1-\sigma(a))$.

(b) Using the previous result and the chain rule of calculus, derive the expression for the gradient of the log likelihood:

$$\nabla \ell(\theta) = [y - \sigma(\theta^T \mathbf{x})]\mathbf{x}$$

where

$$\ell(\theta) = y \log \sigma(\theta^T \mathbf{x}) + (1-y)\log(1-\sigma(\theta^T \mathbf{x}))$$

(a). $\dfrac{d\sigma(a)}{da} = \dfrac{d}{da}\left(\dfrac{1}{1+e^{-a}}\right)$

$= \dfrac{-1}{(1+e^{-a})^2} \cdot e^{-a} = \dfrac{e^{-a}}{(1+e^{-a})^2}$

$\because 1-\sigma(a) = 1 - \dfrac{1}{1+e^{-a}} = \dfrac{1+e^{-a}-1}{1+e^{-a}} = \dfrac{e^{-a}}{1+e^{-a}}$

$\therefore$ $\sigma(a) \cdot (1-\sigma(a)) = \dfrac{e^{-a}}{1+e^{-a}} \cdot \dfrac{1}{1+e^{-a}} = \dfrac{e^{-a}}{(1+e^{-a})^2} = \dfrac{d\sigma(a)}{da}$.

(b). $l(\theta) = y \log \sigma(\theta^T x) + (1-y) \log(1-\sigma(\theta^T x))$

$$= y \nabla \log \sigma(\theta^T x) + (1-y) \nabla \log(1-\sigma(\theta^T x))$$

$$= y \cdot \frac{1}{\sigma(\theta^T x)} \cdot \nabla \sigma(\theta^T x) + (1-y) \frac{1}{1-\sigma(\theta^T x)} \cdot \nabla (1-\sigma(\theta^T x))$$

$$= y \cdot \frac{1}{\cancel{\sigma(\theta^T x)}} \cdot x \cancel{\sigma(\theta^T x)} (1-\sigma(\theta^T x)) + (1-y) \cdot \frac{1}{1-\sigma(\theta^T x)} \cdot (-\sigma(\theta^T x)(1-\sigma(\theta^T x))) x$$

$$= y \cdot \frac{\sigma(\theta^T x)(1-\sigma(\theta^T x)) x}{\sigma(\theta^T x)} + (1-y) \cdot \frac{-\sigma(\theta^T x)(1-\sigma(\theta^T x)) x}{1-\sigma(\theta^T x)}$$

$$= xy - \cancel{y \sigma(\theta^T x)} x - x \sigma(\theta^T x) + \cancel{y \sigma(\theta^T x)} x$$

$$= xy - x\sigma(\theta^T x) = x(y - \sigma(\theta^T x))$$

3. Analytical solution of the Ordinary Least Squares Estimation. Consider we have a simple dataset of n labeled data $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, where data $x_i \in \mathbb{R}$ and $y_i \in \mathbb{R}$ is its corresponding label. We use a simple estimated regression function of:

$$\widehat{y_i} = \theta_0 + \theta_1 x_i$$

Instead of gradient descent which works in an iterative manner, we try to directly solve this problem. We define the cost function as the residual sum of squares, parameterized by θ_0, θ_1:

$$J(\theta_0, \theta_1) = \sum^{n} (y_i - \widehat{y_i})^2$$

(a) Calculate the partial derivative of $\frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1)$ and $\frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1)$.

(b) Consider the fact that $J(\theta_0, \theta_1)$ has an unique optimum, we can actually get the analytical solution of θ_0, θ_1 by the following normal equations:

$$\frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1) = 0$$
$$\frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1) = 0$$

prove the following proprieties that

$$\theta_0 = \bar{y} - \theta_1 \bar{x}$$

and

$$\theta_1 = \frac{\sum_{i=1}^n x_i (y_i - \bar{y})}{\sum_{i=1}^n x_i (x_i - \bar{x})}$$

(Note: $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ and $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$.)

(c) Calculate the sum of the residuals $\sum_{i=1}^n e_i = \sum_{i=1}^n (y_i - (\theta_0 + \theta_1 x_i))$. What can you learn from the value of $\sum_{i=1}^n e_i$?

(a). $\frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1) = \frac{\partial}{\partial_0} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

$= \frac{\partial}{\partial \theta_0} (y_i - \hat{y}_i)^2$

$= 2(y_i - \hat{y}_i) \cdot (y_i - \hat{y}_i)'$

$= 2(y_i - \hat{y}_i) \cdot (-\nabla \hat{y}_i)$

$= (2\hat{y}_i - 2y_i) \cdot (\nabla(\theta_0 + \theta_1 x_i))$

$= (2\theta_0 + 2\theta_1 x_i - 2y_i) \cdot 1$

$= 2\theta_0 + 2\theta_1 x_i - 2y_i = -2\sum_{y}^{n}(y_i - \theta_0 - \theta_i x_i)$

$\frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1) = 2(y_i - \hat{y}_i) \cdot (y_i - \hat{y}_i)'$

$= (2\hat{y}_i - 2y_i) \cdot (-\nabla \hat{y}_i)$

$= (2\hat{y}_i - 2y_i) \cdot (\nabla(\theta_0 + \theta_1 x_i))$

$= (2\theta_0 + 2\theta_i x_i - 2y_i) \cdot x_i$

$= -2x_i \sum^{n} y_i - \theta_1 x_i - \theta_0$

(b). ① $-2\sum_{y}^{n}(y_i - \theta_0 - \theta_i x_i) = 0$

$\sum_{i}^{n}(y_i - \theta_0 - \theta_1 x_i) = 0$

$\sum_{i}^{n} y_i - \sum_{i}^{n}\theta_0 - \sum_{i}^{n}\theta_1 x_i = 0$

$n\bar{y} - n\theta_0 - n\bar{x}\,\theta_1 = 0$

$n\theta_0 = n\bar{y} - n\bar{x}\,\theta_1$

$\theta_0 = \bar{y} - \theta_1 \bar{x}$

$\therefore \sum_{i=1}^{y}\theta_0 = \sum_{i=1}^{n}(\bar{y}_i - \theta_1 x_i)$

$\therefore \theta_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n}\sum_{i=1}^{n}\theta_1 x_i$

$\because \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad \therefore \theta_0 = \bar{y} - \theta_1 \bar{x}$

② $-2x_i \sum_{i=1}^{n} (y_i - \theta_1 x_i - \theta_0) = 0$

$$\sum_i^n x_i y_i - \sum_i^n x_i \theta_1 x_i - \sum_i^n x_i \theta_0 = 0$$

$$\sum_i^n x_i y_i - \sum_i^n x_i \theta_1 x_i - \sum_i^n x_i \sum_i^n (\bar{y}_i - \theta_1 \bar{x}) = 0$$

$$\sum_i^n x_i y_i - \sum_i^n x_i (\bar{y} - \theta_1 \bar{x}) - \sum_i^n x_i \theta_1 x_i = 0$$

$$\sum_i^n x_i y_i - \sum_i^n x_i \bar{y} + \sum_i^n x_i \theta_1 \bar{x} - \sum_i^n x_i \theta_1 x_i = 0$$

$$\Rightarrow \sum_i^n x_i (y_i - \bar{y}) + \sum_i^n \theta_1 x_i (\bar{x} - x_i) = 0$$

$$\therefore \quad \theta_1 = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\sum_{i=1}^{n} x_i (x_i - \bar{x})}$$

(c) $\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - (\bar{y} - \theta_1 \bar{x} + \theta_1 x_i))$

$$= \sum_{i=1}^{n} ((y_i - \bar{y}) + \theta_1 (x_i - \bar{x}))$$

$$= \sum_{i=1}^{n} (y_i - \bar{y}) - \theta_1 \sum_{i=1}^{n} (x_i - \bar{x})$$

$\because \theta_1 = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\sum_{i=1}^{n} x_i (x_i - \bar{x})} \Rightarrow \sum_{i=1}^{n} (y_i - \bar{y}) - \frac{\sum_{i=1}^{n} \cancel{x_i} (y_i - \bar{y})}{\cancel{x_i} \sum_{i=1}^{n} \cancel{(x_i - \bar{x})}} \cdot \sum_{i=1}^{n} \cancel{(x_i - \bar{x})}$

$$\Rightarrow = \sum_{i=1}^{n}(y_i - \bar{y}) - \sum_{i=1}^{n}(y_i - \bar{y}) = 0$$

$$\therefore \sum_{i=1}^{n} e_i = 0$$

$\because \sum_{i=1}^{n} e_i = 0 \Rightarrow$ the sum of redidual is 0, means that given a_0, a_1 the regression function $\hat{y}_i = a_0 + a_1 x_i$ can perfectly predict the data set, it's a perfect estimation.