CARNEGIE MELLON UNIVERSITY 10-315
HOMEWORK 7

DUE: Saturday, April 1, 2023

https://www.cs.cmu.edu/~10315

INSTRUCTIONS

- **Format:** Use the provided LaTeX template to write your answers in the appropriate locations within the *.tex files and then compile a pdf for submission. We try to mark these areas with STUDENT SOLUTION HERE comments. Make sure that you don't change the size or location of any of the answer boxes and that your answers are within the dedicated regions for each question/part. If you do not follow this format, we may deduct points.

  You may also type you answer or write by hand on the digital or printed pdf. Illegible handwriting will lead to lost points. However, we suggest that try to do at least some of your work directly in LaTeX.

  Programming components

- **How to submit written component:** Submit to Gradescope a pdf with your answers. Again, make sure your answer boxes are aligned with the original pdf template.
- **How to submit programming component:** See section Programming Submission for details on how to submit to the Gradescope autograder.
- **Policy:** See the course website for homework policies, including late policy, and academic integrity policies.

Name	
Andrew ID	
Hours to complete all components (nearest hour)	

1 [16 pts] MLE

Consider the following distribution with parameters k and α:

$$p(x \mid k, \alpha) \;=\; \mathbb{I}(x \geq k)\frac{\alpha k^{\alpha}}{x^{\alpha+1}} \;=\; \begin{cases} \frac{\alpha k^{\alpha}}{x^{\alpha+1}} & x \geq k \\ 0 & \text{otherwise} \end{cases}$$

We also have that $k \in (0, \infty)$ and $\alpha \in (0, \infty)$.

This distribution is often used for modeling the distribution of wealth in society, fitting the trend that a large portion of wealth is held by a small fraction of the population. This is due to its nature as a skewed, heavy-tailed distribution (notice that the probability decays polynomially in value of the random variable x, unlike Gaussian, exponential, Laplace or Poisson distributions where the probability decays exponentially in the value of the random variable).

Suppose you have a dataset $\mathcal{D}$ which contains N i.i.d samples $x^{(1)}$,$x^{(2)}$,...,$x^{(N)}$ drawn from the above distribution.

Note: For convenience in this optimization problem, let us define $\log 0$ to be $-\infty$; this won't affect the answers in the end; it essentially just acts as a notational convenience.

1. Write the likelihood $\mathcal{L}(k, \alpha; \mathcal{D})$ and log-likelihood $\ell(k, \alpha; \mathcal{D})$. You may keep the indicator function in your answers (rather than trying to write cases or otherwise avoid the indicator function).

Your Answer

2. What is the numerical value of the likelihood (not the log-likelihood) if we have $\alpha = 2$, $k = 5$, and $N = 4$ data points $\mathcal{D} = \{2, 3, 7, 4\}$?

3. Give the MLE for the parameter α, assuming the parameter k is fixed and $k \leq x^{(i)}\ \forall i$. *Hint*: There is a closed-form solution.

$\hat{\alpha}_{MLE}$

4. Next, give the MLE for the parameter k. Here α may be any fixed value or set to its MLE.

   Hint: You may be tempted to conclude that MLE of k is infinity, but when $k = \infty$, what happens to $p(x \mid k, \alpha)$?

$\hat{k}_{MLE}$

2 Priors and Regularization

In this problem, we are going to have you work through the proof that adding a Laplace prior to probabilistic linear regression is the exact same as adding an L1 regularization to the MSE optimization of linear regression!

Remember our probabilistic linear regression set-up, in which we assume the relationship between response $y \in \mathbb{R}$ and an input $\mathbf{x} \in \mathbb{R}^M$ is linear with zero-mean Gaussian noise added. Note that in this problem for simplicity, we are going to assume that this is no bias term at all.

$$y = \mathbf{w}^T\mathbf{x} + \epsilon \text{ with } \epsilon \sim \mathcal{N}(0, \sigma^2)$$

$$\text{or more concisely, } y \sim \mathcal{N}(\mathbf{w}^T\mathbf{x}, \sigma^2)$$

Recall that the pdf of a Gaussian random variable $\epsilon \sim \mathcal{N}(\mu, \sigma^2)$ is given by:

$$p(\epsilon \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\epsilon-\mu)^2}$$

Also, the pdf of a Laplacian random variable $z \sim \text{Laplace}(\mu, b)$ is given by:

$$p(z \mid \mu, b) = \frac{1}{2b} e^{-\frac{1}{b}(|z-\mu|)}$$

We are going to create our linear regression model with the assumption that each weight $w_j, j \in \{1 \dots M\}$ is drawn from $\text{Laplace}(0, b)$. Finally, assume we have a dataset of N i.i.d. samples $\mathcal{D} = \{(y^{(i)}, \mathbf{x}^{(i)})\}_{i=1}^N$

1. Derive the Laplace prior on the weights $p(\mathbf{w})$ given hyperparameter b. Write your answer in terms of the individual weights w_j, hyperparameter b, and dimension M.

$p(\mathbf{w})$

2. With your answer from the previous part, derive the conditional likelihood times the prior, $p\left(y^{(1)}, \ldots, y^{(N)} \mid \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}, \mathbf{w}\right) p(\mathbf{w})$, which is proportional to the posterior $p(\mathbf{w} \mid \mathcal{D})$.

   Note: Don't include any probability distribution shortcuts in your answer, e.g. make sure to write out the Gaussian distribution rather than stopping at an expression with $\mathcal{N}$ in it.

Your Answer

3. Derive the $\log$ of the conditional likelihood times the prior, $\log\left(p\left(y^{(1)}, \ldots, y^{(N)} \mid \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}, \mathbf{w}\right) p(\mathbf{w})\right)$.

Your Answer

4. Finally, prove that $\text{argmin}_{\mathbf{w}} ||\mathbf{y} - \mathbf{X}\mathbf{w}||_2^2 + \lambda ||\mathbf{w}||_1$ is equivalent to finding $\mathbf{w}$ by minimizing the negative log of the posterior with a Laplace prior.

   Specifically, write λ in terms of the log posterior hyperparameters, b and σ^2.

Your Answer

λ

3 MLE for Categorical Variables

The categorical distribution is a discrete distribution with K possible values (often corresponding to K discrete events). We'll represent this distribution as a one-hot vector of K random binary random variables, $Y = [Y_1, Y_2, \ldots, Y_K]^\top$, where Y_k takes on the value 1 if the outcome is in class k and 0 otherwise. It is described by a parameter vector $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_k)$ where each ϕ_k determines the probability that the random variable is in category k. Also note that $\phi_k \geq 0$ and $\sum_{k=1}^{K} \phi_k = 1$. Note that the Bernoulli distribution is a special case of the categorical distribution where $K = 2$.

For a dataset $\mathcal{D}$ of N i.i.d. samples drawn from the categorical distribution, the likelihood function can be written as:

$$\mathcal{L}(\boldsymbol{\phi}; \mathcal{D}) = \prod_{i=1}^{N} p\left(\mathbf{y}^{(i)} \mid \boldsymbol{\phi}\right) = \prod_{i=1}^{N} \prod_{k=1}^{K} \phi_k^{\mathbb{I}\left(y_k^{(i)}=1\right)}$$

To help us to remove the indicator function notation, we can define N_k as the number of points in $\mathcal{D}$ that belong to the k-th category.

1. Write the likelihood and log-likelihood functions $p(\mathcal{D} \mid \boldsymbol{\phi})$ and $\log p(\mathcal{D} \mid \boldsymbol{\phi})$ in terms of N_k.

$\mathcal{L}(\boldsymbol{\phi}; \mathcal{D})$

$\ell(\boldsymbol{\phi}; \mathcal{D})$

If we just took the derivative of the log-likelihood and set it equal to zero, we would get an estimate saying that each ϕ_k should be equal to infinity, regardless of the value of N_k (???). So, something is a little broken here... What we're missing is to account for the constraint that $\sum_{k=1}^{K} \phi_k = 1$. Here is the constrained optimization problem:

$$\operatorname*{argmin}_{\boldsymbol{\phi}} - \ell(\boldsymbol{\phi}; \mathcal{D})$$

$$\text{s.t. } \sum_{k=1}^{K} \phi_k = 1$$

To solve this constrained optimization problem, we're going to use a really cool trick called Lagrange multipliers. We'll give you just enough information on Lagrange multipliers in this write-up, but if you are curious and want more information, you could start with Lagrange multiplier tutorials in Khan Academy or Paul's Notes.

- **Super brief instructions for using Lagrange multipliers to solve a generic constrained optimization problem**

  Goal: Solve the following optimization:
  $$\operatorname*{argmin}_{\boldsymbol{\phi}} f(\boldsymbol{\phi})$$
  $$\text{s.t. } g(\boldsymbol{\phi}) = 0$$

  Step 1: Construct the "Lagrangian", $L(\boldsymbol{\phi}, \lambda)$, with the added scalar variable λ. Note that we're going to use the letter L for the Lagrangian rather than the typical $\mathcal{L}$ to avoid confusion with the notation for the likelihood.
  $$L(\boldsymbol{\phi}, \lambda) = f(\boldsymbol{\phi}) + \lambda g(\boldsymbol{\phi})$$

  Step 2: Find the "saddle point" of the Lagrangian, specifically the $\boldsymbol{\phi}$ and λ where:
  $$\nabla L(\boldsymbol{\phi}, \lambda) = \mathbf{0}$$

For our categorical MLE problem, we can formulate the Lagrangian (Step 1) with $f(\boldsymbol{\phi})$ as the negative log-likelihood and $g(x) = \sum_{k=1}^{K} \phi_k - 1$.

2. (Lagrange multipliers Step 2a) Write the partial derivatives of the Lagrangian with respect to parameter ϕ_k and with respect to λ:

$\partial L / \partial \phi_k$

$\partial L / \partial \lambda$

3. (Lagrange multipliers Step 2b) Set your expressions for $\partial L / \partial \phi_k$ and $\partial L / \partial \lambda$ both equal to zero and use them as a system of equations to solve for the MLE estimate for ϕ_k.

ϕ_k MLE

4. Finally, let's plug in some numbers to see if our results above match our intuition. For a dataset with $N = 100$ points $K = 4$ categories and $N_1 = 10, N_2 = 20, N_3 = 30$, and $N_4 = 40$, what is the MLE of the vector $\boldsymbol{\phi}$. Write your answer as a vector of numerical values.

$\boldsymbol{\phi}$ MLE

4 Programming

This programming assignment will be completed in the *hw7.ipynb* that is provided for you. It will involve programming machine learning models in PyTorch, which is an industry-standard and widely used Python package for deep learning. It is an incredibly useful and powerful tool, and this programming assignment aims to help give you an introduction to using it.

Please follow the instructions in *hw7.ipynb* given in the handout. Questions 0, 1, and 2 will be autograded while the others will not. The remaining questions 3 and 4 will be graded based off of the plots that are generated that you will provide in your writeup.

1. Provide the plots that were generated from training the `MobileNet` model from scratch and the plots generated from fine-tuning the `MobileNet` model (i.e. training it with pre-trained weights).

From Scratch	Fine-Tuning

2. Which model performed better? Training the model from scratch or fine-tuning the model? Why?

Your Answer

5 Collaboration Policy

After you have completed all other components of this assignment, report your answers to the following collaboration questions.

1. Did you receive any help whatsoever from anyone in solving this assignment? If so, include full details including names of people who helped you and the exact nature of help you received.

2. Did you give any help whatsoever to anyone in solving this assignment? If so, include full details including names of people you helped and the exact nature of help you offered.

3. Did you find or come across code that implements any part of this assignment? If so, include full details including the source of the code and how you used it in the assignment.