COMS W4721 Spring 2020 Homework 1: Maximum Likelihood, Linear Regression, Bias-Variance Tradeoffs

Mengjia Lyu (ml4420@columbia.edu)

February 16, 2020

For this assignment I collaborated with the following people. Blank entries in this table means that I have worked on the corresponding parts on my own.

Problem	Collaborators with their UNIs	Part
Problem 2	Collaborator 1 (uni1@columbia.edu) Collaborator 2 (uni2@columbia.edu)	Part (a), (b) Part (b)
Problem 3		
Problem 4	Collaborator 3 (uni3@columbia.edu)	Part (c)

Problem 1

Introduction: Please briefly describe your academic/career goals and your expectations about the course.

Answer: My career goal is to become a machine learning engineer. My expectations about the course is to learn things that I cannot learn purely from free online resources.

Problem 2

In this problem we will review the principle of *maximum likelihood estimation.*

(a) We are given a coin which falls its heads up with probability $0 < \theta < 1$. Each throw is a Bernoulli random variable $x = \begin{cases} 1, \text{if falls heads up} \\ 0, \text{if falls tails up} \end{cases}$.

For a Bernoulli random variable x: the probability mass function of x is given by:

$$\Pr(x;\theta) = \theta^x (1-\theta)^{(1-x)}$$

Suppose we repeat the coin toss N times to collect the data $\{x^{(i)}\}_{i=1}^N$. Write the log likelihood function $\ln \mathcal{L}(\theta; \{x^{(i)}\}_{i=1}^N)$ and the maximum likelihood estimation of θ, $\hat{\theta}_{\mathsf{MLE}}$.

Solution:

First we write the likelihood function for Bernoulli distribution

$\mathcal{L}(\theta; \{x^{(i)}\}_{i=1}^N) = \prod_{i=1}^{N} \theta^{x^{(i)}} (1-\theta)^{1-x^{(i)}}$

Then we write the log likelihood function for Bernoulli distribution

$$\begin{aligned} \ln \mathcal{L}(\theta; \{x^{(i)}\}_{i=1}^N) &= \prod_{i=1}^{N} \ln \theta^{x^{(i)}} (1-\theta)^{1-x^{(i)}} \\ &= \prod_{i=1}^{N} x^{(i)} \ln \theta + (1 - x^{(i)}) \ln(1-\theta) \\ &= y \ln \theta + (N - y) \ln(1-\theta) \end{aligned} \tag{1}$$

To find the maximum likelihood estimation of θ, $\hat{\theta}_{\mathsf{MLE}}$, we need to find the first derivative of the log likelihood function and set it equal to 0.

$$\frac{\partial \ln \mathcal{L}(\theta; \{x^{(i)}\}_{i=1}^N)}{\partial \theta} = \frac{\partial y \ln \theta + (n-y)\ln(1-\theta)}{\partial \theta} = y\frac{1}{\theta} - (n-y)\frac{1}{1-\theta} = 0$$

Therefore we have

$$\hat{\theta}_{\mathsf{MLE}} = \frac{y}{N} = \frac{\sum_{i=1}^{N} x^{(i)}}{N}$$

(b) Suppose instead we are given a die with K sides with each side falling its heads up with probability $0 < \theta_k < 1$. While we can represent the result of a throw using a categorical variable $x \in \{1, \cdots, K\}$, we can use 1-of-K encoding:

$$x = k \quad \Leftrightarrow \quad \mathbf{x} = [0, \cdots, \underbrace{1}_{\substack{\text{k-th} \\ \text{position} \\ := x_k}}, \cdots, 0]$$

Each throw is a categorical random variable $\mathbf{x} \sim \text{Categorical}(\theta_1, \cdots, \theta_K)$ such that $\Pr(x_k = 1; \theta) = \theta_k$ and $\sum_{k=1}^{K} \theta_k = 1$. The probability mass function at $\mathbf{x}$ is given by:

$$\Pr(\mathbf{x}; \theta_1, \cdots, \theta_K) = \prod_{k=1}^{K} \theta_k^{x_k}$$

Suppose we throw the die N times and obtain the data $\{\mathbf{x}^{(i)}\}_{i=1}^{N}$. Write the log likelihood function $\ln \mathcal{L}(\theta; \{\mathbf{x}^{(i)}\}_{i=1}^{N})$ and the maximum likelihood estimation of θ, $\hat{\theta}_{\mathsf{MLE}}$.

Hint: Once you obtain the log likelihood function $\mathcal{L}(\theta; \{\mathbf{x}^{(i)}\}_{i=1}^{N})$, you will need to add the Lagrangian multiplier part that takes the probability sum constraint $\sum_{k=1}^{K} \theta_k = 1$. For $\lambda \in \mathbb{R}$, the Lagrangian is:

$$\ln \mathcal{L}_\lambda(\theta; \{\mathbf{x}^{(i)}\}_{i=1}^{N}) = \ln \mathcal{L}(\theta; \{\mathbf{x}^{(i)}\}_{i=1}^{N}) + \lambda \left(1 - \sum_{k=1}^{K} \theta_k \right)$$

Take the partial derivative of $\ln \mathcal{L}(\theta; \{\mathbf{x}^{(i)}\}_{i=1}^{N})$ with respect to each θ_k and λ, set them to zero, and solve for each variable of θ_k. Appendix E of Bishop's book may be helpful for this problem.

Solution:

First we write the likelihood function for Bernoulli distribution

$$\mathcal{L}(\theta; \{\mathbf{x}^{(i)}\}_{i=1}^{N}) = \prod_{i=1}^{N} [\prod_{k=1}^{K} \theta_k^{x_k^{(i)}}] = \prod_{k=1}^{K} \theta_k^{N_k}$$

where $N_k = \sum_{i=1}^{N} x_k^{(i)}$is the number of times $x = k$

Then we write the log likelihood function for Bernoulli distribution

$$\begin{aligned} \ln \mathcal{L}(\theta; \{\mathbf{x}^{(i)}\}_{i=1}^{N}) &= \ln \prod_{i=1}^{N} [\prod_{k=1}^{K} \theta_k^{x_k^{(i)}}] \\ &= \ln \prod_{k=1}^{K} \theta_k^{N_k} \\ &= \sum_{k=1}^{K} N_k \ln \theta_k \end{aligned} \tag{2}$$

To find the maximum likelihood estimation of θ, $\hat{\theta}_{\mathsf{MLE}}$, usually we find the first derivative of the log likelihood function and set it equal to 0.

Notice that the function is subject to the probability sum constraint $\sum_{k=1}^{K} \theta_k - 1 = 0$. Therefore, we can use the Lagrangian function $\ln \mathcal{L}_\lambda(\theta; \{\mathbf{x}^{(i)}\}_{i=1}^{N})$ with parameter $\lambda \in \mathbb{R}$ to find the maximum likelihood estimate $\hat{\theta}_{\mathsf{MLE}}$. The constrained cost function becomes

$$\ln \mathcal{L}_\lambda(\theta; \{\mathbf{x}^{(i)}\}_{i=1}^{N}) = \sum_{k=1}^{K} N_k \ln \theta_k + \lambda \left(1 - \sum_{k=1}^{K} \theta_k\right)$$

Taking derivative with respect to θ_k yields

$$\frac{\partial \ln \mathcal{L}_\lambda(\theta; \{\mathbf{x}^{(i)}\}_{i=1}^{N})}{\partial \theta_k} = \frac{N_k}{\theta_k} - \lambda = 0$$

Taking derivative with respect to λ yields the original constraint

$$\frac{\partial \ln \mathcal{L}_\lambda(\theta; \{\mathbf{x}^{(i)}\}_{i=1}^{N})}{\partial \lambda} = \left(1 - \sum_{k=1}^{K} \theta_k\right) = 0$$

Using this sum-to-one constraint we have

$$
\begin{gathered}
N_k = \lambda \theta_k \\
\sum_{k=1}^{K} N_k = \lambda \sum_{k=1}^{K} \theta_k \\
N = \lambda \\
\hat{\theta}_{\mathsf{MLE}}^{(k)} = \frac{N_k}{N} = \frac{\sum_{i=1}^{N} x_k^{(i)}}{N}
\end{gathered}
$$

(c) In class we derived a MLE estimator for the univariate Gaussian assumption. For $\{x^{(i)} \in \mathbb{R}\}_{i=1}^N$ i.i.d, we chose $p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$ and solved for $\hat{\mu}_{\mathsf{MLE}}$ and $\hat{\sigma}^2_{\mathsf{MLE}}$. Repeat this exercise for the multivariate case: now assume $\{\mathbf{x}^{(i)} \in \mathbb{R}^d\}_{i=1}^N$ and choose $p(\mathbf{x}|\mu, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{\mathsf{det}(\Sigma)}} \exp\left(\frac{-(\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)}{2}\right)$. Write the log likelihood function $\ln \mathcal{L}(\{\mu, \Sigma\}; \{x^{(i)}\}_{i=1}^N)$ and solve for $\hat{\mu}_{\mathsf{MLE}}$ and $\hat{\Sigma}_{\mathsf{MLE}}$.

Solution:

$$\begin{aligned}
\ln \mathcal{L}(\{\mu, \mathbf{\Sigma}\}; \{\mathbf{x}^{(\mathbf{i})}\}_{i=1}^N &= \ln \prod_{i=1}^N p(\mathbf{x}|\mu, \mathbf{\Sigma}) \\
&= \sum_{i=1}^N \ln p(\mathbf{x}^{(\mathbf{i})}|\mu, \mathbf{\Sigma}) \\
&= \sum_{i=1}^N \ln \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{\mathsf{det}(\Sigma)}} \exp\left(\frac{-(\mathbf{x}^{(\mathbf{i})}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}^{(\mathbf{i})}-\mu)}{2}\right)
\end{aligned} \tag{3}$$

Since $\frac{1}{(2\pi)^{\frac{d}{2}}}$ is a constant, we can use $C = \frac{1}{(2\pi)^{\frac{d}{2}}}$ to denote it for the sake of simplicity.

$$\begin{aligned}
\ln \mathcal{L}(\{\mu, \mathbf{\Sigma}\}; \{\mathbf{x}^{(\mathbf{i})}\}_{i=1}^N &= \sum_{i=1}^N \ln \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{\mathsf{det}(\Sigma)}} \exp\left(\frac{-(\mathbf{x}^{(\mathbf{i})}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}^{(\mathbf{i})}-\mu)}{2}\right) \\
&= \sum_{i=1}^N \ln \left(C\frac{1}{\sqrt{\mathsf{det}(\Sigma)}}\right) \exp\left(\frac{-(\mathbf{x}^{(\mathbf{i})}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}^{(\mathbf{i})}-\mu)}{2}\right) \\
&= \sum_{i=1}^N \left[\ln C + \ln \frac{1}{\sqrt{\mathsf{det}(\Sigma)}} + \ln \exp\left(\frac{-(\mathbf{x}^{(\mathbf{i})}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}^{(\mathbf{i})}-\mu)}{2}\right)\right] \\
&= \sum_{i=1}^N \left[\frac{-(\mathbf{x}^{(\mathbf{i})}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}^{(\mathbf{i})}-\mu)}{2} - \frac{1}{2}\ln \mathsf{det}(\Sigma) + \ln C\right] \\
&= -\frac{1}{2}\sum_{i=1}^N [(\mathbf{x}^{(\mathbf{i})}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}^{(\mathbf{i})}-\mu)] - \frac{N}{2}\ln \mathsf{det}(\Sigma) + N\ln C
\end{aligned} \tag{4}$$

To solve for $\hat{\mu}_{\mathsf{MLE}}$, we take the derivative of the log likelihood function with respect to μ and set it to 0.

$$
\begin{aligned}
\frac{\partial \ln \mathcal{L}(\{\mu, \mathbf{\Sigma}\}; \{\mathbf{x}^{(\mathbf{i})}\}_{i=1}^{N}}{\partial \mu} &= \frac{\partial - \frac{1}{2}\sum_{i=1}^{N}[(\mathbf{x}^{(\mathbf{i})} - \mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x}^{(\mathbf{i})} - \mu)] - \frac{N}{2}\ln \mathsf{det}(\Sigma) + N \ln C}{\partial \mu} \\
&= \frac{\partial - \frac{1}{2}\sum_{i=1}^{N}[(\mathbf{x}^{(\mathbf{i})} - \mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x}^{(\mathbf{i})} - \mu)}{\partial \mu}
\end{aligned}
\tag{5}
$$

To further simplify the derivative, we introduce the following property

- If $\mathbf{A}$ is symmetric,

  $$\frac{\partial \mathbf{x}^T \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = \mathbf{x}^T(\mathbf{A} + \mathbf{A}^T) = 2\mathbf{A}\mathbf{x}$$

  ***Proof*:**

  By product rule, we have

  $$\frac{\partial \mathbf{x}^T \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = \mathbf{x}^T \frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} + (\mathbf{A}\mathbf{x})^T \frac{\partial \mathbf{x}}{\partial \mathbf{x}}$$

  Notice that

  $$\mathbf{A}\mathbf{x} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} = \begin{bmatrix} a_{11}x1+ & \cdots & a_{1n}xn \\ \vdots & & \\ a_{n1}x1+ & \cdots & a_{nn}xn \end{bmatrix}$$

  Hence $\frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{A}$ and $\frac{\partial \mathbf{x}^T \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{x}^T\mathbf{A} + (\mathbf{A}\mathbf{x})^T = \mathbf{x}^T(\mathbf{A} + \mathbf{A}^T) = 2\mathbf{A}\mathbf{x}$
  If $\mathbf{A}$ is symmetric, $\mathbf{A} = \mathbf{A}^T$ and $\frac{\partial \mathbf{x}^T \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}$
  QED.

Using $\frac{\partial \mathbf{x}^T \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{x}^T(\mathbf{A} + \mathbf{A}^T)$ and the fact that $\mathbf{\Sigma}$ is symmetric, we have

$$
\begin{aligned}
\frac{\partial \ln \mathcal{L}(\{\mu, \Sigma\}; \{x^{(i)}\}_{i=1}^{N}}{\partial \mu} &= \frac{\partial - \frac{1}{2}\sum_{i=1}^{N}[(\mathbf{x^{(i)}} - \mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x^{(i)}} - \mu)}{\partial \mu} \\
&= -\sum_{i=1}^{N} \mathbf{\Sigma}^{-1}(\mathbf{x^{(i)}} - \mu)
\end{aligned}
\tag{6}
$$

Setting the derivative to 0, we have

$$
\begin{aligned}
-\sum_{i=1}^{N} \mathbf{\Sigma}^{-1}(\mathbf{x^{(i)}} - \mu) = 0 \\
\mathbf{\Sigma}^{-1}(\sum_{i=1}^{N} \mathbf{x^{(i)}} - N\mu) = 0
\end{aligned}
$$

Since Σ is positive definite, we can divide both sides by $\mathbf{\Sigma}$

$$
\begin{aligned}
\sum_{i=1}^{N} \mathbf{x^{(i)}} - N\mu = 0 \\
\hat{\mu}_{\mathsf{MLE}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x^{(i)}}
\end{aligned}
$$

To solve for $\hat{\Sigma}_{\mathsf{MLE}}$, we take the derivative of the log likelihood function with

respect to Σ^{-1} and set it to 0.

$$\begin{aligned}
\frac{\partial \ln \mathcal{L}(\{\mu, \mathbf{\Sigma}\}; \{x^{(i)}\}_{i=1}^{N}}{\partial \mathbf{\Sigma}^{-1}} &= \frac{\partial \left[-\frac{1}{2}\sum_{i=1}^{N}[(\mathbf{x^{(i)}}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x^{(i)}}-\mu)] - \frac{N}{2}\ln \mathsf{det}(\mathbf{\Sigma}) + N\ln C\right]}{\partial \mathbf{\Sigma^{-1}}} \\
&= \frac{\partial \left[-\frac{1}{2}\sum_{i=1}^{N}[(\mathbf{x^{(i)}}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x^{(i)}}-\mu) - \frac{N}{2}\ln \mathsf{det}(\mathbf{\Sigma})\right]}{\partial \mathbf{\Sigma^{-1}}} \\
&= \frac{\partial \left[-\frac{1}{2}\sum_{i=1}^{N}\mathsf{trace}\left(\mathbf{\Sigma}^{-1}(\mathbf{x^{(i)}}-\mu)(\mathbf{x^{(i)}}-\mu)^T\right) - \frac{N}{2}\ln \mathsf{det}(\mathbf{\Sigma})\right]}{\partial \mathbf{\Sigma^{-1}}} \\
&= \frac{\partial \left[-\frac{1}{2}\mathsf{trace}\left(\mathbf{\Sigma}^{-1}\sum_{i=1}^{N}[(\mathbf{x^{(i)}}-\mu)(\mathbf{x^{(i)}}-\mu)^T]\right) - \frac{N}{2}\ln \mathsf{det}(\mathbf{\Sigma})\right]}{\partial \mathbf{\Sigma^{-1}}} \\
&= \frac{\partial \left[-\frac{1}{2}\mathsf{trace}\left(\mathbf{\Sigma}^{-1}\sum_{i=1}^{N}[(\mathbf{x^{(i)}}-\mu)(\mathbf{x^{(i)}}-\mu)^T]\right)\right]}{\partial \mathbf{\Sigma^{-1}}} + \frac{\partial\left[-\frac{N}{2}\ln \mathsf{det}(\mathbf{\Sigma})\right]}{\partial \mathbf{\Sigma^{-1}}}
\end{aligned} \tag{7}$$

To further simplify the derivative, we introduce the following properties

- $$\frac{\partial \ln \mathsf{det} M}{\partial M} = M^{-T}$$

  ***Proof*: Using chain rule, we have**

  $$\frac{\partial \ln \mathsf{det} M}{\partial M} = \frac{\partial \ln \mathsf{det} M}{\partial \mathsf{det} M}\frac{\partial \mathsf{det} M}{\partial M}$$

  Note that $\frac{\partial \ln \mathsf{det} M}{\partial M}$ **and** $\frac{\partial \mathsf{det} M}{\partial M}$ **are matrices and** $\frac{\partial \ln \mathsf{det} M}{\partial \mathsf{det} M}$ **is a scalar.**

  Since $\mathsf{det}(M)$**is a scalar,** $\frac{\partial \ln \mathsf{det} M}{\partial \mathsf{det} M} = \frac{1}{\mathsf{det} M}$

  By Jacobi's formula, denoting matrix of all ones by J, we have

  $$\begin{aligned}
  \frac{\partial \mathsf{det} M}{\partial m} &= \mathsf{det} M * \mathsf{trace}\left(M^{-1}\frac{\partial M}{\partial m}\right) \\
  &= \mathsf{det} M * \mathsf{trace}\left(M^{-1}J^{ij}\right)
  \end{aligned} \tag{8}$$

Since trace of a square matrix which is the product of two matrices can be rewritten as the sum of entry-wise products of their elements, we have $\mathsf{trace}(MJ^{ij}) = M^T J^{ij} = (M^T)^{ij}$**.**

Therefore $\mathsf{trace}(M^{-1}J^{ij}) = (M^{-1})^{ji}$**.**

Hence

$$\begin{aligned}\frac{\partial \ln \mathsf{det} M}{\partial M} &= \frac{\partial \ln \mathsf{det} M}{\partial \mathsf{det} M}\frac{\partial \mathsf{det} M}{\partial M} \\ &= \frac{1}{\mathsf{det} M}\mathsf{det} M^{-T} \\ &= M^{-T}\end{aligned} \tag{9}$$

QED.

- $$\frac{\partial}{\partial A}\mathsf{trace}[AB] = B^T$$

***Proof*:**

$$\mathsf{trace}[AB] = \mathsf{trace}\left[\begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_n \end{bmatrix}\begin{bmatrix} b_1 b_2 b_3 \ldots b_n \end{bmatrix}\right]$$

$$= \mathsf{trace}\begin{bmatrix} a_1b_1 & a_1b_2 & \cdots & a_1b_n \\ a_2b_1 & a_2b_2 & \cdots & a_2b_n \\ \vdots & \vdots & \ddots & \vdots \\ a_nb_1 & a_nb_2 & \cdots & a_nb_n \\ \end{bmatrix} \tag{10}$$

$$= \sum_{i=1}^{m} a_{1i}b_{i1} + \sum_{i=1}^{m} a_{2i}b_{i2} + ... + \sum_{i=1}^{m} a_{ni}b_{in}$$

Hence

$$
\begin{aligned}
\frac{\partial}{\partial a_{ij}} \mathsf{trace}[AB] &= b_{ji} \\
\frac{\partial}{\partial A} \mathsf{trace}[AB] &= B^T
\end{aligned}
$$

QED.

Using $\frac{\partial \ln \mathsf{det} M}{\partial M} = M^{-T}$**, we have**

$$
\begin{aligned}
\frac{\partial \left[-\frac{N}{2} \ln \mathsf{det}(\boldsymbol{\Sigma})\right]}{\partial \boldsymbol{\Sigma^{-1}}} &= \frac{\partial \left[\frac{N}{2} \ln \mathsf{det}(\boldsymbol{\Sigma^{-1}})\right]}{\partial \boldsymbol{\Sigma^{-1}}} \\
&= \frac{N}{2} (\boldsymbol{\Sigma}^{-1})^{-T} \\
&= \frac{N}{2} \boldsymbol{\Sigma}^T \\
&= \frac{N}{2} \boldsymbol{\Sigma}
\end{aligned}
\tag{11}
$$

Using $\frac{\partial}{\partial A} \mathsf{trace}[AB] = B^T$**, we have**

$$
\frac{\partial \left[-\frac{1}{2} \mathsf{trace} \left(\boldsymbol{\Sigma}^{-1} \sum_{i=1}^{N} [(\mathbf{x}^{(\mathbf{i})} - \mu)(\mathbf{x}^{(\mathbf{i})} - \mu)^T] \right)\right]}{\partial \boldsymbol{\Sigma^{-1}}} = -\frac{1}{2} \sum_{i=1}^{N} [(\mathbf{x}^{(\mathbf{i})} - \mu)(\mathbf{x}^{(\mathbf{i})} - \mu)^T]
\tag{12}
$$

Combining the results, we have

$$\begin{aligned}\frac{\partial \ln \mathcal{L}(\{\mu, \boldsymbol{\Sigma}\}; \{x^{(i)}\}_{i=1}^{N}}{\partial \boldsymbol{\Sigma}^{-1}} &= \frac{\partial \left[-\frac{1}{2}\mathsf{trace}\left(\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{N}[(\mathbf{x}^{(\mathbf{i})}-\mu)(\mathbf{x}^{(\mathbf{i})}-\mu)^{T}]\right)\right]}{\partial \boldsymbol{\Sigma}^{\mathbf{-1}}} + \frac{\partial\left[-\frac{N}{2}\ln \mathsf{det}(\boldsymbol{\Sigma})\right]}{\partial \boldsymbol{\Sigma}^{\mathbf{-1}}} \\ &= \frac{N}{2}\boldsymbol{\Sigma} - \frac{1}{2}\sum_{i=1}^{N}[(\mathbf{x}^{(\mathbf{i})}-\mu)(\mathbf{x}^{(\mathbf{i})}-\mu)^{T}] \\ &= \frac{1}{2}\left[N\boldsymbol{\Sigma} - \sum_{i=1}^{N}(\mathbf{x}^{(\mathbf{i})}-\mu)(\mathbf{x}^{(\mathbf{i})}-\mu)^{T}\right]\end{aligned} \tag{13}$$

Setting it to zero

$$\begin{aligned} N\boldsymbol{\Sigma} - \sum_{i=1}^{N}(\mathbf{x}^{(\mathbf{i})}-\mu)(\mathbf{x}^{(\mathbf{i})}-\mu)^{T} = 0 \\ \hat{\Sigma}_{\mathsf{MLE}} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}^{(\mathbf{i})}-\hat{\mu}_{\mathsf{MLE}})(\mathbf{x}^{(\mathbf{i})}-\hat{\mu}_{\mathsf{MLE}})^{T}\end{aligned}$$

In summary, we obtain

$\hat{\mu}_{\mathsf{MLE}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}^{(\mathbf{i})} = \bar{\mathbf{x}}$

$\hat{\Sigma}_{\mathsf{MLE}} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}^{(\mathbf{i})}-\hat{\mu}_{\mathsf{MLE}})(\mathbf{x}^{(\mathbf{i})}-\hat{\mu}_{\mathsf{MLE}})^{T}$

Problem 3

In this problem we will use a numerical optimization routine to obtain maximum likelihood estimate of parameters.

Suppose $\{x^{(i)} \in \mathbb{R}\}_{i=1}^{N}$ with $x^{(i)} \sim p(x; x_0, \gamma)$ defined as:

$$p(x; x_0, \gamma) = \frac{1}{\pi \exp(\gamma) \left[1 + \left(\frac{x - x_0}{\exp(\gamma)}\right)^2\right]}$$

(a) Prove that $p(x; x_0, \gamma)$ is a probability density function.

Solution: To show that $p(x; x_0, \gamma)$ is a probability density function, we need to show

1. $p(x; x_0, \gamma) > 0$ for all x.

2. $\int p(x; x_0, \gamma) dx = 1$

1.Since for all x

$$\pi > 0$$
$$\exp(\gamma) > 0$$
$$1 + \left(\frac{x - x_0}{\exp(\gamma)}\right)^2 > 0$$

We have for all x

$$p(x; x_0, \gamma) = \frac{1}{\pi \exp(\gamma) \left[1 + \left(\frac{x - x_0}{\exp(\gamma)}\right)^2\right]} > 0 \tag{14}$$

2.

Denote $\frac{x - x_0}{\exp(\gamma)}$ by t for simplicity.

We know $dx = d(t\exp(\gamma) + x_0) = \exp(\gamma)dt$

$$\begin{aligned}
\int_{-\infty}^{\infty} p(x; x_0, \gamma)dx &= \int_{-\infty}^{\infty} \frac{1}{\pi\exp(\gamma)\left[1 + \left(\frac{x-x_0}{\exp(\gamma)}\right)^2\right]}dx \\
&= \int_{-\infty}^{\infty} \frac{\exp(\gamma)}{\pi\exp(\gamma)\left[1+t^2\right]}dt \\
&= \int_{-\infty}^{\infty} \frac{1}{\pi\left[1+t^2\right]}dt \\
&= \frac{1}{\pi}arctan(t)\mid_{-\infty}^{\infty} \\
&= \frac{1}{\pi}(\frac{\pi}{2} - \frac{-\pi}{2}) \\
&= 1
\end{aligned} \tag{15}$$

Therefore, $p(x; x_0, \gamma)$ **is a probability density function.**

(b) Prove that the mean $\mathbb{E}_{x\sim p(x;x_0,\gamma)}[x]$ is undefined.

Solution:

Denote $\frac{x-x_0}{\exp(\gamma)}$ by t for simplicity.

$$\begin{aligned}
\mathbb{E}_{x\sim p(x;x_0,\gamma)}[x] &= \int_{-\infty}^{\infty} xp(x;x_0,\gamma)dx \\
&= \int_{-\infty}^{\infty} \frac{x}{\pi\exp(\gamma)\left[1+\left(\frac{x-x_0}{\exp(\gamma)}\right)^2\right]}dx \\
&= \int_{-\infty}^{\infty} \frac{\exp(\gamma)t+x_0}{\pi\left[1+t^2\right]}dt \\
&= \int_{-\infty}^{\infty} \frac{\exp(\gamma)t}{\pi\left[1+t^2\right]}dt + \int_{-\infty}^{\infty} \frac{x_0}{\pi\left[1+t^2\right]}dt \\
&= \frac{\exp(\gamma)}{\pi}\int_{-\infty}^{\infty} \frac{t}{\left[1+t^2\right]}dt + \frac{x_0}{\pi}\int_{-\infty}^{\infty} \frac{1}{\left[1+t^2\right]}dt \\
&= \frac{\exp(\gamma)}{\pi}\int_{-\infty}^{\infty} \frac{1}{2\left[1+t^2\right]}dt^2 + \frac{x_0}{\pi}\int_{-\infty}^{\infty} \frac{1}{\left[1+t^2\right]}dt \\
&= \frac{\exp(\gamma)}{2\pi}\int_{-\infty}^{\infty} \frac{1}{\left[1+t^2\right]}dt^2 + \frac{x_0}{\pi}\int_{-\infty}^{\infty} \frac{1}{\left[1+t^2\right]}dt \\
&= \frac{\exp(\gamma)}{2\pi}log(t^2)\mid_{1}^{\infty} + \frac{x_0}{\pi}arctan(t)\mid_{-\infty}^{\infty} \\
&= \frac{\exp(\gamma)}{2\pi}(log(\infty)-log(1))+x_0
\end{aligned} \tag{16}$$

However, $log(\infty)$ is undefined.

Hence, $\mathbb{E}_{x\sim p(x;x_0,\gamma)}[x]$ is undefined.

(c) Write the log likelihood function $\ln \mathcal{L}(\{x_0, \gamma\}; \{x^{(i)}\}_{i=1}^N)$ and the expression for $\frac{\partial \ln \mathcal{L}(\{x_0,\gamma\};\{x^{(i)}\}_{i=1}^N)}{\partial x_0}$ and $\frac{\partial \ln \mathcal{L}(\{x_0,\gamma\};\{x^{(i)}\}_{i=1}^N)}{\partial \gamma}$. Plot the log likelihood value as a 3D surface plot: x-axis should run over x_0 and y-axis should run over γ. z-axis should correspond to the log likelihood value at the corresponding (x_0, γ) pair. Include the plot in your writeup. Do the stationary points (solutions to the maximum likelihood equations) have closed form solutions?

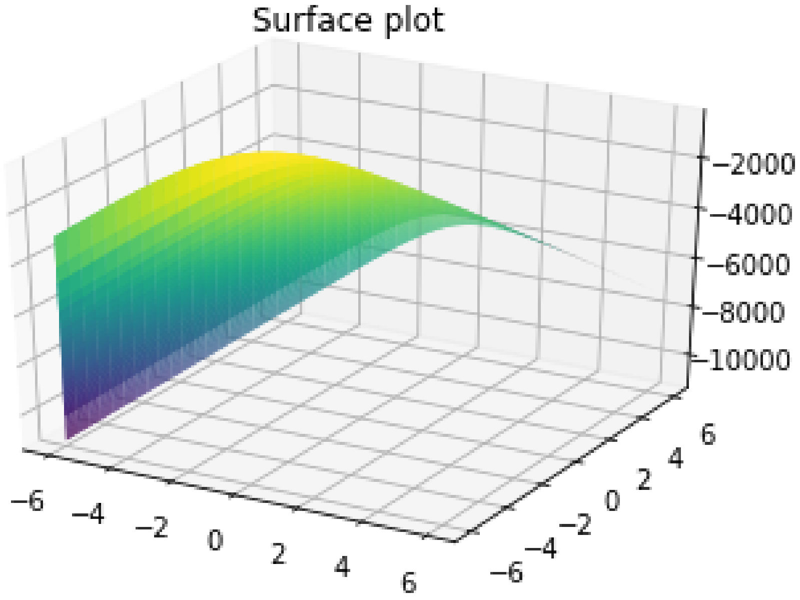
Solution:

$$\begin{aligned}
\ln \mathcal{L}(\{x_0, \gamma\}; \{x^{(i)}\}_{i=1}^N) &= \ln \prod_{i=1}^{N} p(x; x_0, \gamma) \\
&= \ln \prod_{i=1}^{N} \frac{1}{\pi \exp(\gamma) \left[1 + \left(\frac{x^{(i)} - x_0}{\exp(\gamma)}\right)^2\right]} \\
&= \sum_{i=1}^{N} \ln \frac{1}{\pi \exp(\gamma) \left[1 + \left(\frac{x^{(i)} - x_0}{\exp(\gamma)}\right)^2\right]} \\
&= -\sum_{i=1}^{N} \ln \left[\pi \exp(\gamma) \left(1 + \left(\frac{x^{(i)} - x_0}{\exp(\gamma)}\right)^2\right)\right] \\
&= -\sum_{i=1}^{N} \ln \pi - \sum_{i=1}^{N} \ln \exp(\gamma) - \sum_{i=1}^{N} \ln \left[1 + \left(\frac{x^{(i)} - x_0}{\exp(\gamma)}\right)^2\right] \\
&= -N \ln \pi - N \ln \exp(\gamma) - \sum_{i=1}^{N} \ln \left[1 + \left(\frac{x^{(i)} - x_0}{\exp(\gamma)}\right)^2\right] \\
&= -N \ln \pi - N\gamma - \sum_{i=1}^{N} \ln \left[1 + \left(\frac{x^{(i)} - x_0}{\exp(\gamma)}\right)^2\right]
\end{aligned} \tag{17}$$

$$\begin{aligned}
\frac{\partial \ln \mathcal{L}(\{x_0, \gamma\}; \{x^{(i)}\}_{i=1}^{N})}{\partial x_0} &= \frac{\partial \left(-N \ln \pi - N\gamma - \sum\limits_{i=1}^{N} \ln \left[1 + \left(\frac{x^{(i)} - x_0}{\exp(\gamma)} \right)^2 \right] \right)}{\partial x_0} \\
&= \frac{\partial \left(- \sum\limits_{i=1}^{N} \ln \left[1 + \left(\frac{x^{(i)} - x_0}{\exp(\gamma)} \right)^2 \right] \right)}{\partial x_0} \\
&= - \sum_{i=1}^{N} \frac{2 \frac{x_0 - x^{(i)}}{\exp(2\gamma)}}{\left[1 + \left(\frac{x^{(i)} - x_0}{\exp(\gamma)} \right)^2 \right]}
\end{aligned} \tag{18}$$

$$\begin{aligned}
\frac{\partial \ln \mathcal{L}(\{x_0, \gamma\}; \{x^{(i)}\}_{i=1}^{N})}{\partial \gamma} &= \frac{\partial \left(-N \ln \pi - N\gamma - \sum\limits_{i=1}^{N} \ln \left[1 + \left(\frac{x^{(i)} - x_0}{\exp(\gamma)} \right)^2 \right] \right)}{\partial \gamma} \\
&= -N - \frac{\partial \sum\limits_{i=1}^{N} \ln \left[1 + \left(\frac{x^{(i)} - x_0}{\exp(\gamma)} \right)^2 \right]}{\partial \gamma} \\
&= -N - \sum_{i=1}^{N} \frac{-2 \frac{(x^{(i)} - x_0)^2}{\exp(2\gamma)}}{\left[1 + \left(\frac{x^{(i)} - x_0}{\exp(\gamma)} \right)^2 \right]}
\end{aligned} \tag{19}$$

Here is the plot:



From the expression derived for the gradients, it is obvious that we cannot solve for either of x_0 or γ easily. Therefore we conclude the stationary points do not have closed-form solutions.

x_0	γ
-20.000	-20.000
-19.910	-18.994
-19.552	-14.970
-18.117	1.126
-12.377	65.511
-17.296	10.335
-17.950	3.000
32.637	0.372
-8.872	2.529
-2.321	2.188
15.158	1.280
1.068	2.012
2.803	1.922
65.257	-5.513
6.151	1.524
3.478	1.842
3.362	1.802
2.897	1.641
1.038	0.996
-1.092	-1.284
0.701	0.635
1.073	0.180
2.565	-1.642
1.467	-0.301
2.798	-2.678
1.613	-0.562
1.574	-1.033
1.360	-1.915
1.482	-1.410
1.475	-2.491
1.553	-2.313
1.498	-2.439
1.491	-2.157
1.493	-2.211
1.493	-2.211
1.493	-2.211
1.493	-2.211
1.493	-2.211
1.493	-2.211

Table 1: Iteration History for (d)

(d) Write a program to obtain an estimate of $\theta = \{x_0, \gamma\}$ using the dataset **problem3.csv**. If you are using Python, it is helpful to utilize scipy.optimize.minimize function. Choose gradient descent optimizer or a quasi-Newton optimizer such as BFGS. List the optimizer that was chosen for this problem with the initial iterate. Tabulate the coordinates of the iterates of the optmization process and the final converged solution.

Solution:

Please refer to Table 1 for the tabulation of the iterates of the optimization process and the final converged solution.

The final converged solution is x_0=1.49348727, λ=-2.21055569.

The gradient descent optimizer that was chosen for this problem is BFGS.

The initial iterate starts from x_0=-20,λ=-20

Problem 4

In this problem you will implement ridge regression estimator using gradient descent. Although the ridge regression does have a closed form solution, gradient descent form lets you avoid explicit matrix inversion and scale to larger data. We will see this in our subsequent lectures.

For this problem, we assume we are given the labeled data pair:
$\{(\mathbf{x}^{(i)} \in \mathbb{R}^d, y^{(i)} \in \mathbb{R})\}_{i=1}^N$. The regularized objective function in this case is given by:

$$\min_{b\in\mathbb{R},\mathbf{w}\in\mathbb{R}^d} L(b, \mathbf{w}; \{(\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^N) := \frac{1}{N}\sum_{i=1}^N (y^{(i)} - (b + \mathbf{w}^T \cdot \mathbf{x}^{(i)}))^2 + \lambda \cdot ||\mathbf{w}||_2^2$$

The pseudocode for performing gradient descent is given by the following algorithm. The main structure consists of a loop which continues for a given number of epochs T. η is the learning rate that controls the amount you want to step into the direction of the negative gradient, and λ is the regularization parameter.

```
function GDRIDGE($S_{\mathsf{train}} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^N$, $T$, $\eta$, $\lambda$)
    Initialize the bias term $b \leftarrow 0$ and the slope $\mathbf{w} \leftarrow \mathbf{0}$
    for $t = 1, \cdots, T$ do
        $b_{\mathsf{new}} \leftarrow b - \eta \cdot \frac{\partial L}{\partial b}$,    $\mathbf{w}_{\mathsf{new}} \leftarrow \mathbf{w} - \eta \cdot \frac{\partial L}{\partial \mathbf{w}}$
        $b \leftarrow b_{\mathsf{new}}$,    $\mathbf{w} \leftarrow \mathbf{w}_{\mathsf{new}}$
    end for
end function
```

(a) Write the expression for the gradient update for b and $\mathbf{w}$.

Solution:

The gradient update for w is

$$\begin{aligned}\mathbf{w}_{\mathsf{new}} &= \mathbf{w} - \eta \cdot \frac{\partial L}{\partial \mathbf{w}} \\ &= \mathbf{w} - \eta \left[\frac{2}{N}\sum_{i=1}^N (y^{(i)} - (b + \mathbf{w}^T \cdot \mathbf{x}^{(i)}))(-\mathbf{x}^{(i)}) + 2\lambda\mathbf{w}\right]\end{aligned}$$

The gradient update for b is

$$
\begin{aligned}
b_{\textsf{new}} &= b - \eta \cdot \frac{\partial L}{\partial b} \\
&= b - \eta \frac{2}{N} \sum_{i=1}^{N} (y^{(i)} - (b + \mathbf{w}^T \cdot \mathbf{x}^{(i)}))(-1) \\
&= b + \eta \left[\frac{2}{N} \sum_{i=1}^{N} (y^{(i)} - (b + \mathbf{w}^T \cdot \mathbf{x}^{(i)})) \right]
\end{aligned}
$$

(b) Implement the function GDRidge described above. Please include your source code in the writeup.

Solution:

```
def gradient_descent_ridge_reg(X, y, num_epochs,learning_rate,lambda)
    w = np.zeros((X.shape[1],1)) #initialize the slope to 0
    b = 0 #initialize the bias term to 0
    N = y.shape[0] #extract N from input

    for i in range(num_epochs):

        temp = y-(b+np.dot(X, w)) #temporary variable to hold value f
        w_partial = -2*np.matmul(X.T,temp)/N + 2*lambda*w
        b_partial = 2*np.sum(temp)/N

        b_new = b + learning_rate * b_partial
        w_new = w - learning_rate * w_partial
        b = b_new #update bias term
        w = w_new #update slope
```

(c) Use the Boston housing data (`https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html`) and scale the data appropriately: standardize the feature matrix and [0, 1] scale the y values. Choose $T = 100$ and $\eta = 0.01$.

For $\lambda = 0.1$, provide a x-y plot with the epoch number as x-axis and plot the following quantities on the y-axis:

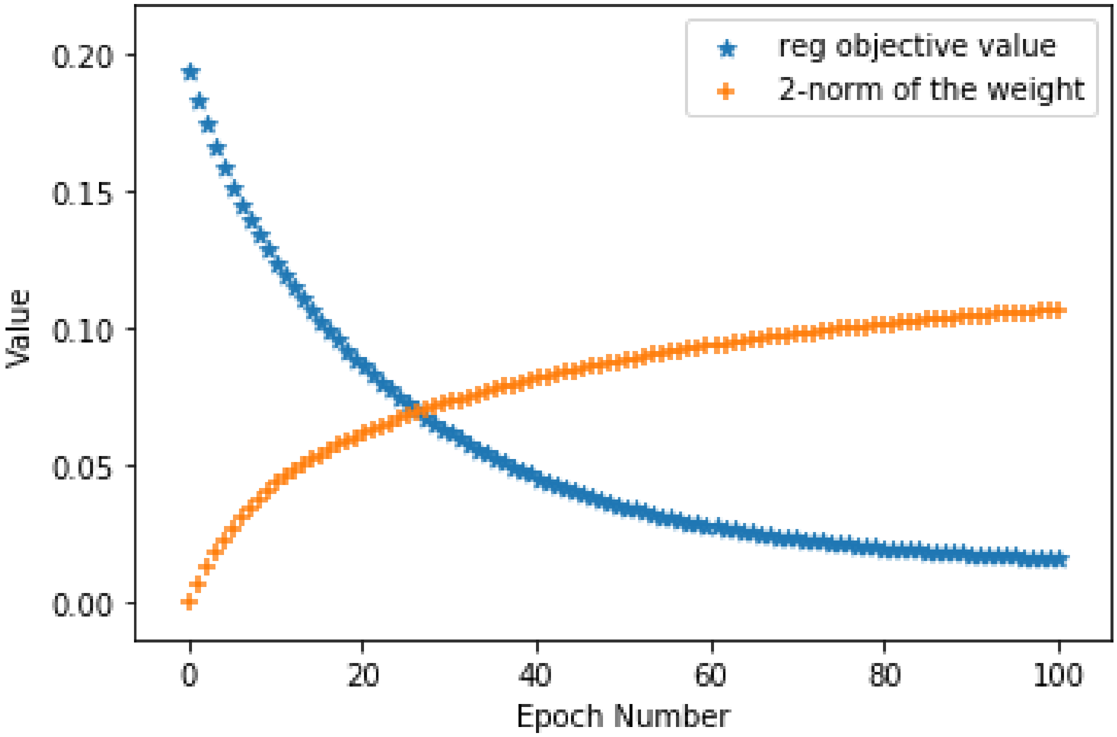
- The value of regularized objective L at the start of each epoch.
- The 2-norm of the weight vector $\mathbf{w}$ at the start of each epoch.

Here is an example plot:

`convergence.png`

This will require you to modify the function written in Part (b) to compute the required quantities.

Solution:



reg objective value
2-norm of the weight
Value
0.20
0.15
0.10
0.05
0.00
0
20
40
60
80
100
Epoch Number

(d) The next plot will examine how the coefficients for each of the features change as the regularization parameter is varied.

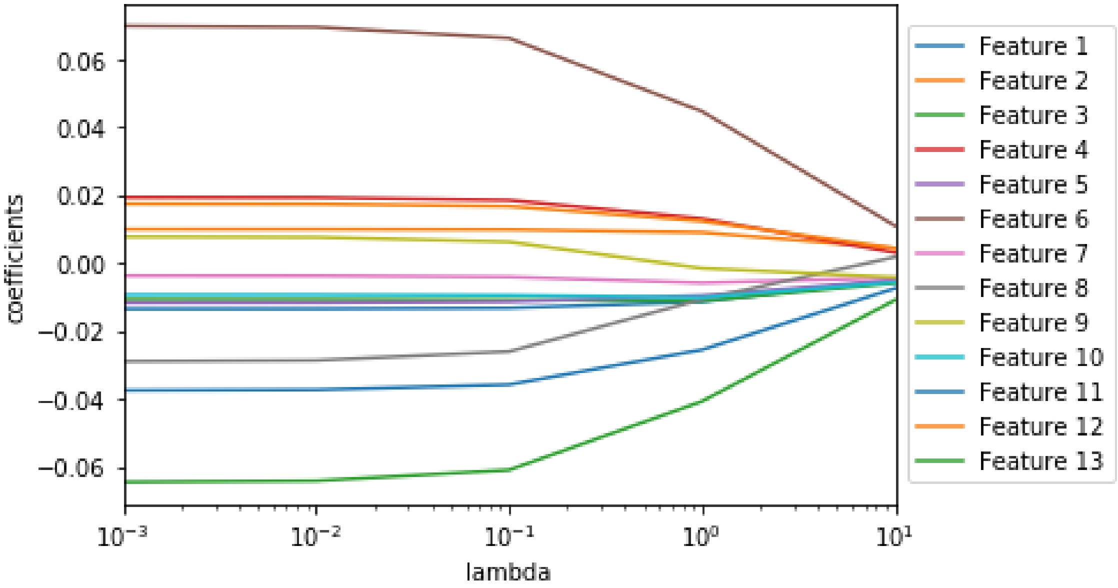
Provide a coefficient path plot for the Boston housing data for $\lambda = 10, 1, 0.1, 0.01, 0.001$. x-axis is for the regularization value and y-axis for the coefficient of the final converged iterate of your gradient descent algorithm. Make sure to scale the data appropriately. Choose $T = 100$ and $\eta = 0.01$.

Here is an example plot for a dataset with 7 features: there is a connected path for each of the 7 features as the regularization parameter is varied.

coef_path.png

What do you notice about the behavior of the coefficients as the regularization is varied? Include the coefficient path plot and your analysis.

Solution:



From the coefficient path plot, we notice that the ridge estimate coefficients converge to zero as the regularization parameter tends to infinity.

Ridge regression is a regularization method which tries to avoid overfitting of data by penalizing large coefficients. Ridge regression has an additional factor λ which is called the penalty factor that is added while estimating the coefficients. This penalty factor penalizes high value of coefficients which in turn shrinks coefficients, thereby reducing the mean squared error and predicted error.

Therefore, the higher the value of λ, the greater will be the shrinkage of the coefficients. As λ approach infinity, the coefficients would converge to 0 (but never become 0).