CS559 Hw 1

I pledge my honor that I have abided by the Stevens Honor System

Dimitrios Haralampopoulos

2023-02-22

Please follow the below instructions when you submit the assignment.

- You are NOT allowed to use packages for implementing the code required in this assignment (question 5). You can use packages for data processing and data split (k-fold cross validation).
- Your submission should consist of a zip file named Assignment1_LastName_FirstName.zip which contains:
  - a jupyter notebook file (.ipynb) or a python file (.py). The file should contain the code and the output after execution (in comments if you use python). You should also include detailed comments.
  - a pdf file to show (1) the derivation steps of for questions 1 to 4 and (2) experiment design and results (plots, tables, etc) for question 5.

(10 points) Assuming data points are independent and identically distributed (i.i.d.), the probability of the data set given parameters: μ and σ^2 (the likelihood function):

$$P(\mathbf{x}|\mu,\sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu,\sigma^2)$$

Please calculate the solution for μ and σ^2 using Maximum Likelihood (ML) estimator.

$$\mathcal{N}(x_n|\mu,\sigma^2) = -\frac{1}{(2\pi\sigma^2)^{1/2}} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$$

More convenient to maximize the log likelihood rather than the standard likelihood because of its monotonically increasing nature.
Log likelihood given by:

$$\ln p(\mathbf{x}|\mu,\sigma^2) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)$$

$$\max_{\mu,\sigma^2} \ln p(\mathbf{x}|\mu,\sigma^2)$$

requires $\frac{\partial}{\partial\mu}\ln p(\mathbf{x}|\mu,\sigma^2)$ and $\frac{\partial}{\partial\sigma^2}\ln p(\mathbf{x}|\mu,\sigma^2)$

$$\frac{\partial}{\partial \mu}\left(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2-\frac{N}{2}\ln\sigma^2-\frac{N}{2}\ln(2\pi)\right)=-\frac{1}{2\sigma^2}\cdot -2\left(\sum_{n=1}^{N}(x_n-\mu)\right)=\frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n-\mu)=\frac{1}{\sigma^2}\left(\sum_{n=1}^{N}x_n-N\mu\right)$$

which is equal to zero only when

$$\sum_{n=1}^{N}x_n-N\mu=0$$

which implies

$$\mu=\frac{1}{N}\sum_{n=1}^{N}x_n$$

We apply a similar approach for σ^2

$$\frac{\partial}{\partial \sigma^2}\left(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2-\frac{N}{2}\ln\sigma^2-\frac{N}{2}\ln(2\pi)\right)=-\frac{n}{2\sigma^2}-\left[\frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right]\frac{d}{d\sigma^2}\left(\frac{1}{\sigma^2}\right)=$$

$$-\frac{n}{2\sigma^2}-\left[\frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right]\left(-\frac{1}{(\sigma^2)^2}\right)=-\frac{n}{2\sigma^2}+\left[\frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right]\left(\frac{1}{(\sigma^2)^2}\right)=\frac{1}{2\sigma^2}\left[\frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2-N\right]$$

which gives us

$$\sigma^2=\frac{1}{N}\sum_{n+1}^{N}(x_n-\mu)^2$$

when set equal to zero.

Therefore,

$$\mu_{ML}=\frac{1}{N}\sum_{n=1}^{N}x_n,\quad \sigma^2_{ML}=\frac{1}{N}\sum_{n+1}^{N}(x_n-\mu)^2$$

(10 points) We assume there is a true function $f(\mathbf{x})$ and the target value is given by $y=f(x)+\epsilon$ where ϵ is a Gaussian distribution with mean 0 and variance σ^2. Thus,

$$p(y|x,w,\beta)=\mathcal{N}(y|f(x),\beta^{-1})$$

where $\beta^{-1}=\sigma^2$.

Assuming the data points are drawn independently from the distribution, we obtain the likelihood function:

$$p(\mathbf{y}|\mathbf{x},w,\beta)=\prod_{n=1}^{N}\mathcal{N}(y_n|f(x),\beta^{-1})$$

Please show that maximizing the likelihood function is equivalent to minimizing the sum-of-squares error function.

Once again most convenient to maximize the log likelihood function of the distribution:

$$\ln p(\mathbf{y}|\mathbf{x}, w, \beta) = -\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, w) - y_n\}^2 + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi)$$

let us maximize first with respect to w:

$$\frac{\partial}{\partial w}\ln p(\mathbf{y}|\mathbf{x}, w, \beta) = -\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, w) - t_n\}^2\left(\frac{\partial}{\partial w}\right)$$

We can scale the log likelihood by a positive coefficient without altering the location of the maximum with respect to w, so we can replace the coefficient $\frac{\beta}{2}$ with $\frac{1}{2}$, yielding:

$$\frac{\partial}{\partial w}\ln p(\mathbf{y}|\mathbf{x}, w, \beta) = -\frac{1}{2}\sum_{n=1}^{N}\{y(x_n, w) - t_n\}^2\left(\frac{\partial}{\partial w}\right)$$

$$E(w) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, w) - t_n\}^2,$$

where $E(w)$ is the sum-of-squares error function. From here we can see that the maximization of the likelihood function is equivalent to minimizing the sum-of-squares error function multiplied by a constant of -1.

$$\frac{\partial}{\partial w}\ln p(\mathbf{y}|\mathbf{x}, w, \beta) = -\frac{1}{2}\sum_{n=1}^{N}\{y(x_n, w) - t_n\}^2\left(\frac{\partial}{\partial w}\right) = -1\cdot\left(\frac{1}{2}\sum_{n=1}^{N}\{y(x_n, w) - t_n\}^2\right)\frac{\partial}{\partial w} = -E(w)\frac{\partial}{\partial w}$$

(15 points) Given input values $\mathbf{x} = (x_1, ..., x_N)^T$ and their corresponding target values $\mathbf{y} = (y_1, ..., y_N)^T$, we estimate the target by using function $f(x, \mathbf{w})$ which is a polynomial curve. Assuming the target variables are drawn from Gaussian distribution:

$$p(y|x, \mathbf{w}, \beta) = \mathcal{N}(y|f(x, \mathbf{w}), \beta^{-1})$$

and a prior Gaussian distribution for $\mathbf{w}$:

$$p(\mathbf{w}|\alpha) = (\frac{\alpha}{2\pi})^{(M+1)/2}\exp(-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w})$$

Please prove that maximum posterior (MAP) is equivalent to minimizing the regularized sum-of-squares error function. Note that the posterior distribution of $\mathbf{w}$ is $p(\mathbf{w}|\mathbf{x}, \mathbf{y}, \alpha, \beta)$. **Hint: use Bayes' theorem.**

Bayes' Theorem is defined as $p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})}$, where $\mathcal{D}$ is the observed data set and $\mathbf{w}$ is parameter vector. We can also view Bayes' Theorem as posterior $\propto$ likelihood $\times$ prior.

Using this definition of Bayes' Theorem, we can see that $p(\mathbf{w}|\mathbf{x}, \mathbf{y}, \alpha, \beta) \propto p(y|\mathbf{x}, \mathbf{w}, \beta) \cdot p(\mathbf{w}|\alpha)$

$$p(\mathbf{w}|\mathbf{x}, \mathbf{y}, \alpha, \beta) \propto \mathcal{N}(y|f(x, \mathbf{w}), \beta^{-1}) \cdot (\frac{\alpha}{2\pi})^{(M+1)/2} \exp(-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}) =$$

$$-ln\left(p(\mathbf{w}|\mathbf{x}, \mathbf{y}, \alpha, \beta) \propto \mathcal{N}(y|f(x, \mathbf{w}), \beta^{-1}) \cdot (\frac{\alpha}{2\pi})^{(M+1)/2} \exp(-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w})\right) \frac{\partial}{\partial \mathbf{w}} =$$

$$-ln(p(\mathbf{w}|\mathbf{x}, \mathbf{y}, \alpha, \beta) \frac{\partial}{\partial \mathbf{w}} = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - y_n\}^2 + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}$$

The regularized sum-of-squares error function is given as:

$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - y_n\}^2 + \frac{\lambda}{2}||\mathbf{w}||^2$$

As given in (1.67), $\lambda = \alpha/\beta$ and $||\mathbf{w}||^2 = \mathbf{w}^T\mathbf{w}$, then:

$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - y_n\}^2 + \frac{\alpha/\beta}{2}\mathbf{w}^T\mathbf{w}$$

Given that we will be taking the partial derivatives of both $p(\mathbf{w}|\mathbf{x}, \mathbf{y}, \alpha, \beta)$ and $\widetilde{E}(\mathbf{w})$ with respect to $\mathbf{w}$, we can say that β becomes 1 in this scenario, like in the proof for the maximization of the likelihood function. This yields:

$$p(\mathbf{w}|\mathbf{x}, \mathbf{y}, \alpha, \beta) \frac{\partial}{\partial \mathbf{w}} = \left(\frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - y_n\}^2 + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}\right) \frac{\partial}{\partial \mathbf{w}}$$

and

$$\widetilde{E}(\mathbf{w}) \frac{\partial}{\partial \mathbf{w}} = \left(\frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - y_n\}^2 + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}\right) \frac{\partial}{\partial \mathbf{w}}$$

From this, we can conclude that MAP is equivalent to minimizing the regularized sum-of-squares error function.

(20 points) Consider a linear model of the form:

$$f(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{i=1}^{D} w_i x_i$$

together with a sum-of-squares error/loss function of the form:

$$L_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{f(\mathbf{x}_n, \mathbf{w}) - y_n\}^2$$

Now suppose that Gaussian noise ϵ_i with zero mean and variance σ^2 is added independently to each of the input variables x_i. By making use of $\mathbb{E}[\epsilon_i] = 0$ and $\mathbb{E}[\epsilon_i \epsilon_j] = \delta_{ij}\sigma^2$ where $\delta_{ii} = 1$, show that minimizing L_D averaged over the noise distribution is equivalent to minimizing the sum-of-squares error for noise-free input variables with the addition of a weight-decay regularization term, in which the bias parameter w_0 is omitted from the regularizer.

$$\hat{f}(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{i=1}^{D} w_i (x_i + \epsilon_i)$$

$$L(\mathbf{w}) = L_D(\mathbf{w}) + \lambda L_W(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{f(\mathbf{x}_n, \mathbf{w}) - y_n\}^2 + \frac{\lambda}{2} \mathbf{w}^T \mathbf{w}$$

$$\widehat{L}_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{\hat{f}(\mathbf{x}_n, \mathbf{w}) - y_n\}^2 = \frac{1}{2} \sum_{n=1}^{N} \{w_0 + \sum_{i=1}^{D} w_i (x_n + \epsilon_i) - y_n\}^2 = \frac{1}{2} \sum_{n=1}^{N} \{w_0 + \sum_{i=1}^{D} (w_i x_n + w_i \epsilon_i) - y_n\}^2$$

$$\widehat{L}_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{w_0^2 + (\sum_{i=1}^{D} (w_i x_n + w_i \epsilon_i))^2 + y_n^2 + 2w_0 \sum_{i=1}^{D} (w_i x_n + w_i \epsilon_i) - 2w_0 y_n - 2y_n \sum_{i=1}^{D} (w_i x_n + w_i \epsilon_i)\}$$

$$L_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{w_0 + \sum_{i=1}^{D} w_i x_n - y_n\}^2 = \frac{1}{2} \sum_{n=1}^{N} \{w_0^2 + (\sum_{i=1}^{D} w_i x_n)^2 + y_n^2 + 2w_0 \sum_{i=1}^{D} (w_i x_n) - 2w_0 y_n - 2y_n \sum_{i=1}^{D} (w_i x_n)\}$$

$$\widehat{L}_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{w_0^2 + \sum_{i=1}^{D} (w_i x_n + w_i \epsilon_i)^2 + 2 \sum_{i<j} (w_j x_n + w_j \epsilon_j)(w_i x_n + w_i \epsilon_i) + y_n^2 + \cdots - 2y_n \sum_{i=1}^{D} (w_i x_n + w_i \epsilon_i)\}$$

$$\widehat{L}_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{w_0^2 + \sum_{i=1}^{D} (w_i^2 x_n^2 + w_i^2 \epsilon_i^2 + 2w_i^2 x_n \epsilon_i) + 2 \sum_{i<j} (w_j x_n + w_j \epsilon_j)(w_i x_n + w_i \epsilon_i) + y_n^2 + \cdots - 2y_n \sum_{i=1}^{D} (w_i x_n + w_i \epsilon_i)\}$$

ϵ_i is on a Gauss Distribution, therefore:

$$\mathbb{E}[\epsilon_i] = \mu = 0, \ \mathbb{E}[\epsilon_i \epsilon_j] = \delta_{ij}\sigma^2$$

$$\delta_{ij} = 1 \iff \epsilon_i = \epsilon_j, \ | \ \delta_{ij} = 0 \iff \epsilon_i \neq \epsilon_j$$

So let us consider $L_D(\mathbf{w})$ and $\widehat{L}_D(\mathbf{w})$ with the above equations in mind:

$$\widehat{L}_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{w_0^2+\sum_{i=1}^{D}(w_i^2x_n^2+w_i^2\sigma^2+0)+2\sum_{i<j}(w_jx_n+0)(w_ix_n+0)+y_n^2+2w_0\sum_{i=1}^{D}(w_ix_n+0)-2w_0y_n-2y_n\sum_{i=1}^{D}(w_ix_n+0)\} =$$

$$\widehat{L}_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{w_0^2 + \sum_{i=1}^{D}(w_i^2x_n^2 + w_i^2\sigma^2) + 2\sum_{i<j}(w_jw_ix_n^2) + y_n^2 + 2w_0\sum_{i=1}^{D}(w_ix_n) - 2w_0y_n - 2y_n\sum_{i=1}^{D}(w_ix_n)\}$$

$$L_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{w_0^2 + \sum_{i=1}^{D}(w_i^2x_n^2) + 2\sum_{i<j}(w_jw_ix_n^2) + y_n^2 + 2w_0\sum_{i=1}^{D}(w_ix_n) - 2w_0y_n - 2y_n\sum_{i=1}^{D}(w_ix_n)\}$$

$$\widehat{L}_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{w_0^2+\sum_{i=1}^{D}(w_i^2x_n^2)+\sum_{i=1}^{D}(w_i^2\sigma^2)+2\sum_{i<j}(w_jw_ix_n^2)+y_n^2+2w_0\sum_{i=1}^{D}(w_ix_n)-2w_0y_n-2y_n\sum_{i=1}^{D}(w_ix_n)\}$$

$$\widehat{L}_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{w_0^2+\sum_{i=1}^{D}(w_i^2x_n^2)+2\sum_{i<j}(w_jw_ix_n^2)+y_n^2+2w_0\sum_{i=1}^{D}(w_ix_n)-2w_0y_n-2y_n\sum_{i=1}^{D}(w_ix_n)\}+\frac{1}{2}\sum_{n=1}^{N}\sum_{i=1}^{D}(w_i^2\sigma^2)$$

We can assume the distribution of the Gaussian Noise to be Standard Normal, hence:

$$\widehat{L}_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{w_0^2+\sum_{i=1}^{D}(w_i^2x_n^2)+2\sum_{i<j}(w_jw_ix_n^2)+y_n^2+2w_0\sum_{i=1}^{D}(w_ix_n)-2w_0y_n-2y_n\sum_{i=1}^{D}(w_ix_n)\}+\frac{1}{2}\sum_{n=1}^{N}\sum_{i=1}^{D}(w_i^2)$$

$$\widehat{L}_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{w_0^2+\sum_{i=1}^{D}(w_i^2x_n^2)+2\sum_{i<j}(w_jw_ix_n^2)+y_n^2+2w_0\sum_{i=1}^{D}(w_ix_n)-2w_0y_n-2y_n\sum_{i=1}^{D}(w_ix_n)\}+\frac{1}{2}\sum_{n=1}^{N}\{w_1^2+\cdots+w_D^2\}$$

Given $||\mathbf{w}||^2 = \mathbf{w}^T\mathbf{w} = w_0^2 + w_1^2 + \cdots + w_M^2$, omitting the w_0 term from the regularizer, we can apply this to $\widehat{L}_D(\mathbf{w})$ to get:

$$\widehat{L}_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{w_0^2 + \sum_{i=1}^{D}(w_i^2x_n^2) + 2\sum_{i<j}(w_jw_ix_n^2) + y_n^2 + 2w_0\sum_{i=1}^{D}(w_ix_n) - 2w_0y_n - 2y_n\sum_{i=1}^{D}(w_ix_n)\} + \frac{N}{2}||\mathbf{w}||^2$$

We can then see that $\widehat{L}_D(\mathbf{w})$ contains the familiar structure of $L_D(\mathbf{w})$, so we see that:

$$\widehat{L}_D(\mathbf{w}) = L_D(\mathbf{w}) + \frac{N}{2}||\mathbf{w}||^2$$

Which is in a similar form to $L(\mathbf{w})$, so we can see that by minimizing both $\widehat{L}_D(\mathbf{w})$ and $L(\mathbf{w})$ with respect to $\mathbf{w}$, we get:

$$L(\mathbf{w})\frac{\partial}{\partial\mathbf{w}} = (L_D(\mathbf{w}) + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w})\frac{\partial}{\partial\mathbf{w}} = L_D(\mathbf{w})\frac{\partial}{\partial\mathbf{w}} + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w}\frac{\partial}{\partial\mathbf{w}}$$

$$\widehat{L}_D(\mathbf{w})\frac{\partial}{\partial\mathbf{w}} = (L_D(\mathbf{w}) + \frac{N}{2}\mathbf{w}^T\mathbf{w})\frac{\partial}{\partial\mathbf{w}} = L_D(\mathbf{w})\frac{\partial}{\partial\mathbf{w}} + \frac{N}{2}\mathbf{w}^T\mathbf{w}\frac{\partial}{\partial\mathbf{w}}$$

For our purposes, λ and N are sufficiently similar to prove equivalence (since N is the order of the polynomial and can also be used to control the model complexity). Therefore, minimizing L_D averaged over the noise distribution is equivalent to minimizing the sum-of-squares error for noise free input variables with the addition of a weight decay regularization term in which the bias parameter w_0 is omitted from the regularizer.

(45 points) Please choose **one** of the below problems. You will need to **submit your code**.

a) UCI Machine Learning: Facebook Comment Volume Data Set

Please implement a Ridge regression model and use mini-batch gradient descent to train the model on this dataset for predicting the number of comments in next H hrs (H is given in the feature). You do not need to use all the features. Use K-fold cross validation and report the mean squared error (MSE) on the test data. You need to write down every step in your experiment.

b) UCI Machine Learning: Bike Sharing Data Set

Please write a Ridge regression model and use mini-batch gradient descent to train the model on this dataset for predicting the count of total rental bikes including both casual and registered. You do not need to use all the features. Use K-fold cross validation and report the mean squared error (MSE) on the test data. You need to write down every step in your experiment.

For this experiment, I decided to check the correlation between the weather conditions of a given day or hour and the number of people who had used the service in the respective timeframe. For my experiment, I used a step size of 0.1, 10 epochs, 50 batches, an alpha of 1.0, and 5 generations of repetition during training for the weight vectors w and b The model produces results relatively accurately, and retain the general form of casual + registered = total with accuracy throughout predictions. The MSE produced from the predictions compared to that from the Cross-Validation is also very close, within only a few hundred for Hours but within several hundred thousand for Days:

Predicted MSE for Days: 1955387.1484762367
Predicted MSE for Hours: 9452.569950011382
Cross Validation MSE for Days: 2182835.982595852
Cross Validation MSE for Hours: 9093.387058600014

First point of prediction on training set for Days: [395.55582633 1624.33431989 2019.89014623]
First point actual for Days: [163. 3667. 3830.]
First point of prediction on training set for Hours: [31.70026545 140.23893303 171.93919848]
First point actual for Hours: [26. 363. 389.]