Assignment 1

Homework assignments will be done individually: each student must hand in their own answers. Use of partial or entire solutions obtained from others or online is strictly prohibited. Electronic submission on Canvas is mandatory.

1. **Maximum Likelihood estimator** (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.

Answer:

Given,

$$\begin{aligned} P(\mathbf{x}|\mu,\sigma^2) &= \prod_{n=1}^{N} \mathcal{N}(x_n|\mu,\sigma^2) \\ &= \prod_{n=1}^{N} [\frac{1}{(2\pi\sigma^2)^{N/2}} \exp\{\frac{-1}{2\sigma^2}(x_n - \mu)^2\}] \end{aligned} \tag{1}$$

Taking log likelihood of $P(\mathbf{x}|\mu,\sigma^2)$ to simplify the equation,

$$\begin{aligned} \ln P(\mathbf{x}|\mu,\sigma^2) &= \ln(\frac{1}{(2\pi)^{N/2}}) + \ln(\frac{1}{(\sigma)^{N/2}}) + \ln(\exp\{\frac{-1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2\}) \\ &= -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2 \end{aligned} \tag{2}$$

i) To find the solution for μ, taking derivative of (2) with respect to μ and setting it to 0,

$$\frac{\delta}{\delta\mu}(\ln P(\mathbf{x}|\mu,\sigma^2)) = 0 \tag{3}$$

$$-0 - 0 - \frac{1}{2\sigma^2}(-2\sum_{n=1}^{N}(x_n - \mu)) = 0 \tag{4}$$

$$\sum_{n=1}^{N}(x_n - \mu) = 0 \tag{5}$$

$$\mu_{ML} = \sum_{n=1}^{N} x_n \qquad (as \sum_{n=1}^{N} \mu = \mu) \tag{6}$$

ii) To find the solution for σ^2, taking derivative of (2) with respect to σ^2 and setting it to 0,

$$\frac{\delta}{\delta\sigma^2}(\ln P(\mathbf{x}|\mu,\sigma^2)) = 0 \tag{7}$$

$$-0 - \frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}(\sum_{n=1}^{N}(x_{\mathrm{n}} - \mu)^2) = 0 \tag{8}$$

$$\frac{1}{2\sigma^4}(\sum_{n=1}^{N}(x_{\mathrm{n}} - \mu)^2) = \frac{N}{2\sigma^2} \tag{9}$$

$${\sigma_{\mathrm{ML}}}^2 = \frac{1}{N}\sum_{n=1}^{N}(x_{\mathrm{n}} - \mu)^2 \tag{10}$$

Hence, (6) and (10) are required solutions.

2. **Maximum Likelihood** (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.

Answer:

Given,

$$p(\mathbf{y}|\mathbf{x}, w, \beta) = \prod_{n=1}^{N} \mathcal{N}(y_n|f(x), \beta^{-1}) \tag{11}$$

$$= \prod_{n=1}^{N} (\frac{\beta}{2\pi})^{N/2} \exp\{\frac{-\beta}{2}(y_\mathrm{n} - f(\mathbf{x}))^2\} \tag{12}$$

Taking log likelihood of $p(\mathbf{y}|\mathbf{x}, w, \beta)$, to simplify the equation

$$\ln p(\mathbf{y}|\mathbf{x}, w, \beta) = \ln(\beta)^{N/2} - \ln(2\pi)^{N/2} + \ln(\exp\{\frac{-\beta}{2}\sum_{n=1}^{N}(y_\mathrm{n} - f(\mathbf{x}))^2\}) \tag{13}$$

$$= \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) - \frac{\beta}{2}\sum_{n=1}^{N}(y_\mathrm{n} - f(\mathbf{x}))^2 \tag{14}$$

$$= \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) - \beta E_\mathrm{N}(w) \tag{15}$$

where $E_\mathrm{N}(w) = \frac{1}{2}\sum_{n=1}^{N}(y_\mathrm{n} - f(\mathbf{x}))^2$ is the sum-of-sqaures loss function.
So, when we try to maximize this function, indirectly we will be minimizing sum-of-squares loss function.

Hence, maximum likelihood is equivalent to minimizing sum-of-squares loss function

3. **MAP estimator** (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.**

Answer:

From Bayes' theorem we know that,

$$posterior \propto likelihood * prior \tag{16}$$

$$p(\mathbf{w}|\mathbf{x}, \mathbf{y}, \alpha, \beta) \propto p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \beta) * p(\mathbf{w}|\alpha) \tag{17}$$

$$\propto \prod_{n=1}^{N} \mathcal{N}(y_n|f(x_n, \mathbf{w}), \beta^{-1}) * (\frac{\alpha}{2\pi})^{(M+1)/2} \exp(-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}) \tag{18}$$

$$\propto \prod_{n=1}^{N} (\frac{\beta}{2\pi})^{N/2} \exp\{\frac{-\beta}{2}(y_n - f(x_n, \mathbf{w}))^2\} * (\frac{\alpha}{2\pi})^{(M+1)/2} \exp(-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}) \tag{19}$$

To simplify equation taking negative logarithm,

$$\ln p(\mathbf{w}|\mathbf{x}, \mathbf{y}, \alpha, \beta) \equiv -\ln(\frac{\beta}{2\pi})^{N/2} - \ln(\exp\{\frac{-\beta}{2}\sum_{n=1}^{N}(y_n - f(x_n, \mathbf{w}))^2\}) - \ln(\frac{\alpha}{2\pi})^{(M+1)/2} - \ln(\exp\{\frac{-\alpha}{2}\mathbf{w}^T\mathbf{w}\}) \tag{20}$$

$$\equiv \frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\beta) + \frac{\beta}{2}\sum_{n=1}^{N}(y_n - f(x_n, \mathbf{w}))^2 - \frac{(M+1)}{2}(\ln(\alpha) - \ln(2\pi)) + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} \tag{21}$$

$$\equiv \frac{\beta}{2}\sum_{n=1}^{N}(y_n - f(x_n, \mathbf{w}))^2 + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} + \underbrace{\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\beta) - \frac{(M+1)}{2}(\ln(\alpha) - \ln(2\pi))}_{\text{constant w.r.t } \mathbf{w}} \tag{22}$$

When we try to maximize eq(22) by taking derivative w.r.t $\mathbf{w}$, we get the following:

$$\frac{\delta}{\delta\mathbf{w}}\ln p(\mathbf{w}|\mathbf{x}, \mathbf{y}, \alpha, \beta) \equiv \frac{\delta}{\delta\mathbf{w}}[\frac{\beta}{2}\sum_{n=1}^{N}(y_n - f(x_n, \mathbf{w}))^2 + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}] + 0 + 0 + 0 \tag{23}$$

which implies, trying to maximize posterior is equivalent to minimizing the regularized sum-of-squares error function.

4. **Linear model** (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 $\mathbb{E}[\epsilon_i \epsilon_j] = \delta_{ij}\sigma^2$, 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.

Answer:

Adding Gaussian noise ϵ_{i} to every input variable x_{i},

$$\hat{f} = f(\mathbf{x}_{\mathrm{i}} + \epsilon_{\mathrm{i}}, \mathbf{w}) \tag{24}$$

$$= w_0 + \sum_{i=1}^{D} w_{\mathrm{i}}(x_{\mathrm{i}} + \epsilon_{\mathrm{i}}) \tag{25}$$

$$= \underbrace{w_0 + \sum_{i=1}^{D} w_{\mathrm{i}} x_{\mathrm{i}}} + \sum_{i=1}^{D} w_{\mathrm{i}} \epsilon_{\mathrm{i}} \tag{26}$$

$$= f(\mathbf{x}, \mathbf{w}) + \sum_{i=1}^{D} w_{\mathrm{i}} \epsilon_{\mathrm{i}} \tag{27}$$

Now, new sum-of-sqaures loss function would become,

$$\hat{L_{\mathrm{D}}}(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{f(\mathbf{x}_{\mathrm{n}} + \epsilon_{\mathrm{n}}, \mathbf{w}) - y_{\mathrm{n}}\}^2 \tag{28}$$

$$= \frac{1}{2} \sum_{n=1}^{N} \{f(\mathbf{x}_{\mathrm{n}} + \epsilon_{\mathrm{n}}, \mathbf{w})^2 - 2f(\mathbf{x}_{\mathrm{n}} + \epsilon_{\mathrm{n}}, \mathbf{w}) y_{\mathrm{n}} + {y_{\mathrm{n}}}^2\} \tag{29}$$

$$= \frac{1}{2} \sum_{n=1}^{N} \{f^2(\mathbf{x}, \mathbf{w}) + 2f(\mathbf{x}, \mathbf{w}) \sum_{i=1}^{D} w_{\mathrm{i}} \epsilon_{\mathrm{ni}} + (\sum_{i=1}^{D} w_{\mathrm{i}} \epsilon_{\mathrm{ni}})^2 - 2(f(\mathbf{x}, \mathbf{w}) + \sum_{i=1}^{D} w_{\mathrm{i}} \epsilon_{\mathrm{ni}}) y_{\mathrm{n}} + {y_{\mathrm{n}}}^2\} \tag{30}$$

$$= \frac{1}{2} \sum_{n=1}^{N} \{f^2(\mathbf{x}, \mathbf{w}) + \underbrace{2f(\mathbf{x}, \mathbf{w}) \sum_{i=1}^{D} w_{\mathrm{i}} \epsilon_{\mathrm{ni}}} + (\sum_{i=1}^{D} w_{\mathrm{i}} \epsilon_{\mathrm{ni}})^2 - 2f(\mathbf{x}, \mathbf{w}) y_{\mathrm{n}} - \underbrace{2(\sum_{i=1}^{D} w_{\mathrm{i}} \epsilon_{\mathrm{ni}}) y_{\mathrm{n}}} + {y_{\mathrm{n}}}^2\} \tag{31}$$

It is given that $\mathbb{E}[\epsilon_i] = 0$ and $\mathbb{E}[\epsilon_i \epsilon_j] = \delta_{ij}\sigma^2$, hence if we take expectation of $\hat{L_{\mathrm{D}}}(\mathbf{w})$, then underlined term from eq 31 will be equal to 0 as well as,

$$\mathbb{E}[(\sum_{i=1}^{D} w_{\mathrm{i}}\epsilon_{\mathrm{ni}})^2] = \mathbb{E}[(\sum_{i=1}^{D} {w_{\mathrm{i}}}^2{\epsilon_{\mathrm{ni}}}^2)] \tag{32}$$

$$= \sum_{i=1}^{D} {w_{\mathrm{i}}}^2\mathbb{E}[{\epsilon_{\mathrm{ni}}}^2] \tag{33}$$

$$= \sum_{i=1}^{D} {w_{\mathrm{i}}}^2\sigma^2 \tag{34}$$

Hence, from eq31 and eq34,

$$\mathbb{E}[\hat{L_{\mathrm{D}}}(\mathbf{w})] = \frac{1}{2}\sum_{n=1}^{N}\{f^2(\mathbf{x},\mathbf{w}) - 2f(\mathbf{x},\mathbf{w})y_{\mathrm{n}} + {y_{\mathrm{n}}}^2\} + \frac{1}{2}\sum_{i=1}^{D} {w_{\mathrm{i}}}^2\sigma^2 \tag{35}$$

$$= \frac{1}{2}\sum_{n=1}^{N}\{f(\mathbf{x}_n,\mathbf{w}) - y_n\}^2 + \frac{1}{2}\sum_{i=1}^{D} {w_{\mathrm{i}}}^2\sigma^2 \tag{36}$$

$$= L_{\mathrm{D}}(\mathbf{w}) + \frac{1}{2}\sum_{i=1}^{D} {w_{\mathrm{i}}}^2\sigma^2 \tag{37}$$

From eq37, we can say that, to minimize L_{D} averaged over the noise distribution we need to minimize the sum-of-squares error for noise free input variables with the addition of a weight-decay regularization term.

5. **Linear regression** (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 apply both Lasso regression and Ridge regression algorithms 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 testing data. You need to write down every step in your experiment.

   b) UCI Machine Learning: Bike Sharing Data Set

   Please apply both Lasso regression and Ridge regression algorithms 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 testing data. You need to write down every step in your experiment.

   Answer: **UCI Machine Learning: Bike Sharing Data Set** : Using hourly dataset
   Steps:

   1. Importing every required python package
   2. loading datset from csv file to pandas dataframe using pandas package
   3. performing feature selection to discard unnecessary features. Also, creating two dataframes: 1)for features 2)for target labels, which is 'cnt' in this problem
   4. we can use head() function from pandas dataframe to verify selected features and few rows
   5. before we can apply asked machine learning model, we need to split the dataset in training and test sets. For this purpose, I am utilizing train_test_split() function from sklearn package.
   6. first, I am calculating mean squared error(MSE) for lasso and ridge regression without kfold cross validation method. Lasso and Ridge regression is performed using sklearn package. Using small alpha to increase generalization.
   7. Withous using kfold cross validation, we get **MSE for Lasso = 21736.526 and MSE for Ridge = 21738.389**.
   8. Now, using KFold class from sklearn package to apply kfold cross validation in both Ridge and Lasso regression.
   9. Here, I am using K=10, which is proven to be the best value for most of the datasets.
   10. MSE values for all 10 iterations of KFold cross validation are given in below table.
   11. From table, we notice that split#1 has the lowest MSE in all 10 splits for both Lasso as well Ridge regression. Also, it is very low as compared to MSE when applying Lasso and Ridge regression without using K-fold. Hence, we see huge error reduction when utilizing K-fold cross validation.

Split#	MSE for Lasso	MSE for Ridge
1	11183.715	11174.003
2	11589.548	11604.182
3	24992.733	25171.491
4	16254.678	16258.456
5	12859.541	12816.065
6	17338.029	17412.233
7	35951.613	35925.332
8	33695.769	33633.476
9	45870.249	45766.028
10	28029.165	28070.004