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.

Sol: $\because f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

$$\therefore P(x|\mu,\sigma^2) = \prod_{n=1}^{N} f(x)$$

$$= \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

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

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

$\therefore$ we have

$$\begin{cases} \dfrac{\partial \ln P(x|\mu,\sigma^2)}{\partial \mu} = \dfrac{1}{\sigma^2}\left(\sum_{n=1}^{N} x_n - N\mu\right) = 0 \\ \dfrac{\partial \ln P(x|\mu,\sigma^2)}{\partial \sigma^2} = -\dfrac{N}{2\sigma^2} + \dfrac{\sum_{n=1}^{N}(x_n-\mu)^2}{2\sigma^4} = 0 \end{cases}$$

$\therefore$ we get

$$\begin{cases} \mu = \dfrac{1}{N}\sum_{n=1}^{N} x_1 = \bar{x} \\ \sigma^2 = \dfrac{\sum_{n=1}^{N}(x_1-\bar{x})}{N} \end{cases}$$

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.

Sol: $P(y|x,w,\beta) = \prod_{n=1}^{N} N(y_n|f(x), \beta^{-1})$

$\therefore \ln P(y|x,w,\beta) = -\frac{N}{2}\ln 2\pi + \frac{N}{2}\ln\beta - \beta \cdot \frac{1}{2}\sum_{n=1}^{N}(y_n - f(x))^2$

$\because E_D(w) = \frac{1}{2}\sum_{n=1}^{N}(y_n - w^T x_n)^2 = \frac{1}{2}\sum_{n=1}^{N}(y_n - f(x_n))^2$

$\therefore$ ML function equal to minimizing the sum of squares error 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.**

Sol: Bayes' function : $P(B|A) = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{N} P(A|B_j)P(B_j)}$

$\therefore$ MAP = $\arg\max P(x_i|y)$

$= \arg\max\left(\frac{P(y|x_i)P(x_i)}{\sum_{j=1}^{N} P(y|x_j)P(x_j)}\right)$

$= \arg\max \prod_{i=1}^{N} P(y|x_i)P(x_i)$

$\because P(y|X, w, \beta) = \frac{1}{\sqrt{2\pi\beta}} e^{-\frac{\beta}{2}(y_n - f(x_n; w))^2}$

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

$MAP = (2\pi)^{-\frac{N}{2}} \cdot \beta^{\frac{N}{2}} \cdot e^{-\frac{1}{2}\beta \sum_{i=1}^{N}(y_i - f(x_i; w))^2} \cdot (\frac{\alpha}{2\pi})^{\frac{M+1}{2}} e^{-\frac{\alpha}{2}w^T \cdot w}$

$\ln MAP = -\frac{N}{2}\ln 2\pi + \frac{N}{2}\ln\beta - \frac{1}{2}\beta \sum_{i=1}^{N}[y_i - f(x_i, w)]^2 + (\frac{M+1}{2}) \cdot \ln(\frac{\alpha}{2\pi}) + -\frac{\alpha}{2}w^T \cdot w$

$\because -\frac{N}{2}\ln 2\pi, \frac{N}{2}\ln\beta, \frac{M+1}{2}\ln\frac{\alpha}{2\pi}, -\frac{\alpha}{2}w^T w$ are constant

$\therefore$ we need $\frac{1}{2}\beta \sum [y - f(x, w)]^2$ minimizing to make MAP max.

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.

Sol: $\tilde{L}_D(w) = \frac{1}{2}\sum_{n=1}^{N}\{w_0 + \sum_{i=1}^{D} w_i x_i - t_n\}^2$

$= \frac{1}{2}\sum_{n=1}^{N}\{w_0 + \sum_{i=1}^{D} w_i(x_i+\epsilon_i) - t_n\}^2$

$= \frac{1}{2}\sum_{n=1}^{N}\{w_0 + \sum_{i=1}^{D} w_i x_i - t_n + \sum_{i=1}^{D} w_i \epsilon_i\}^2$

$= \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, w) - t_n + \sum_{i=1}^{D} w_i \epsilon_i\}^2$

$= \frac{1}{2}\sum_{n=1}^{N}\{[y(x_n, w) - t_n]^2 + (\sum_{i=1}^{D} w_i \epsilon_i)^2 + 2(\sum_{i=1}^{D} w_i \epsilon_i)(y(x_n w) - t_n)\}$

$E[(\sum_{i=1}^{D} w_i \epsilon_i)^2] = E[\sum_{i=1}^{D}\sum_{j=1}^{D} w_i \epsilon_i w_j \epsilon_j] = \sum_{i=1}^{D}\sum_{j=1}^{D} w_i w_j E[\epsilon_i \epsilon_j]$

$= \sum_{i=1}^{D}\sum_{j=1}^{D} w_i w_j \delta_{ij} \sigma^2$

$E[2(\sum_{i=1}^{D} w_i \epsilon_i)(y(x_n w) - t_n)] = E[\epsilon_i] \cdot 2\sum_{i=0}^{D} w_i \cdot E[y(x_n w - t_n)] = 0.$ $= \sigma^2 \sum_{i=1}^{D} w_i^2$

~~that~~ minimizing L_D over

$\because E[(\sum_{i=1}^{D} w_i \epsilon_i)^2] = \sigma^2 \sum_{i=1}^{D} w_i^2$ & is a constant, $E[2(\sum_{i=1}^{D} w_i \epsilon_i)(y(x_n w) - t_n)] = 0$

$\therefore$ minimizing L_D averaged over the noise distribution is equivalent to minimizing the sum of squares error for

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.

b) step 1: Download UCI Machine Learning: Bike sharing Data Set and create a new python file

step 2: ~~Get feature from~~ Get feature from .csv file as X and set a label as y.

step 3: Using train_test_split get X_train, X_test, y_train and y_test.

step4 Create ridge and lasso rlinear regression (Using sklearn import these two model)

step5: Create "cross_val_score()" this method to archieve k-fold cross validation generate y_predict.

step 6: Create "mean_squared_error" this method to archieve MSE on testing data.