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
Solution:
The probability density of observing a single data point x, that is generated from a Gaussian distribution is given by,

$P(x|\mu,\sigma^2) = \frac{1}{(\sigma\sqrt{2\pi}))} exp^{-\frac{(x-\mu)^2}{(2\sigma^2)}}$

The joint probability of observing n data points is given by,

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

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

Taking log on both sides, we get

$LnP(x_n|\mu,\sigma^2) = Ln[\sum_{n=1}^{N} \frac{1}{\sigma\sqrt{2\pi})} exp^{-\frac{(x_n-\mu)^2}{2\sigma^2}}]$

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

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

Differentiating equation(1) partially w.r.t μ

$\frac{\partial[LnP(x_n|\mu,\sigma^2)]}{\partial\mu} = \sum_{n=1}^{N} \{(\frac{1}{2\sigma^2} \times 2 \times -(x_n - \mu) \times -1) + 0 + 0\}$

In order to maximise the likelihood the partial derivative, $\frac{\partial[LnP(x_n|\mu,\sigma^2)]}{\partial\mu} = 0$

Therefore we have,

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

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

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

$\mu = \frac{1}{N}\{\sum_{n=1}^{N} x_n \qquad ...(2)$

Differentiating equation(1) partially w.r.t σ
$\frac{\partial[LnP(x_n|\mu,\sigma^2)]}{\partial\sigma} = \sum_{n=1}^{N}\{-\frac{(x_n-\mu)^2}{2} \times -2\sigma^3 - \frac{1}{\sigma}\}$

In order to maximise the likelihood the partial derivative, $\frac{\partial[LnP(x_n|\mu,\sigma^2)]}{\partial\sigma} = 0$

$\sum_{n=1}^{N}\{-\frac{(x_n-\mu)^2}{\sigma^3} - \frac{1}{\sigma}\} = 0$

$\sum_{n=1}^{N}\{\frac{(x_n-\mu)^2-\sigma^2}{\sigma^3}\} = 0$

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

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

$\sum_{n=1}^{N} x_n^2 - 2\mu \times \sum_{n=1}^{N} x_n + \mu^2\sum_{n=1}^{N} 1 = N\sigma^2$
Dividing the whole equation by N and from eq(2), we get
$\frac{1}{N} \times \sum_{n=1}^{N} x_n^2 - 2 \times \frac{1}{N^2} \times \sum_{n=1}^{N} x_n^2 \times \sum_{n=1}^{N} x_n^2 + (\frac{1}{N} \times \sum_{n=1}^{N} x_n)^2 = \sigma^2$

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

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

Now we have,
$E(x^2) = \frac{1}{N}\sum_{n=1}^{N} x_n^2$
$E(x) = \frac{1}{N}\sum_{n=1}^{N} x_n$
Therefore by substituting in above equation we get,

$\sigma^2 = E(x^2) - [E(x)]^2$

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.

Solution:

The target value is given by,

$y = f_w(x) + \epsilon$

where $\epsilon isa Gaussian distribution with mean 0 and variance \sigma^2$

The residual of the above function is given by the formula,

$Residual = f_w(x) - y$

The loss is calculated as a sum of square errors, as all the terms in the sum are no-negative and error above the line is same as error below the line, thus, ignoring the sign of the value and only considering its magnitude. The loss function is defined as,

$L(x, y, w) = [f_w(x) - y]^2$

The error associated with the above line is,

$E = \sum_{i=1}^{N} (w^T x_i - y_i)^2 \qquad ...(1)$

The goal is to minimize this sum of squared prediction error(least squared error or LEER).

The maximum likelihood for a Gaussian distribution is given by,

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

$p(y|x, \omega, \beta) = \sum_{n=1}^{N} (\sqrt{\frac{\beta}{2\pi)}} exp^{-\frac{(y_n - f(x))^2 \times \beta}{2}})$

Taking logarithm on both sides, we get

$Ln[p(y|x, \omega, \beta)] = \sum_{n=1}^{N} \{\frac{-1}{2} \times Ln(2\pi) + \frac{1}{2} Ln(\beta) - \frac{\beta}{2} \times (y_n - f(x))^2\}$

$= \frac{Ln(\beta)}{2} \sum_{n=1}^{N} (1) + \frac{Ln(2\pi)}{2} \sum_{n=1}^{N} (1) - \frac{\beta}{2} \sum_{n=1}^{N} (y_n - f(x))^2$

$= \frac{N \times Ln(\beta)}{2} + \frac{N \times Ln(2\pi)}{2} - \frac{\beta}{2} \sum_{n=1}^{N} (f(x) - y_n)^2$

where,

Sum of square error $= \sum_{n=1}^{N} (f(x) - y_n)^2$

Since for fixed β, i.e standard deviation $\beta > 0$, both terms are constant.

Only sum of square error vary. In order to maximise value of the likelihood function the sum of square must be minimised.

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.**
Solution:

The maximum likelihood function is given as,

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

$p(y|x, w, \beta) = \sum_{n=1}^{N} \sqrt{\frac{\beta}{(2\pi)}} \times exp^{\frac{\beta(y_n - f(x,w))^2}{2}}$

The prior Gaussian distribution for w,

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

Using Bayes theorem, the posterior distribution for w is given as follows,

$p(w|x, y, \alpha, \beta) = p(y|x, w, x, \beta) \times p(w|\alpha)$

$= \{\sum_{n=1}^{N} \sqrt{\frac{\beta}{(2\pi)}} \times exp^{\frac{\beta(y_n - f(x,w))^2}{2}}\} \times \{(\frac{\alpha}{2\pi})^{\frac{M+1}{2}} exp^{-(\frac{\alpha}{2}w^T w)}\}$

Taking logarithm on both sides,

$Ln[\text{p(w—x,y,}\alpha, \beta)] = Ln\{\sum_{n=1}^{N} \sqrt{\frac{\beta}{(2\pi)}} \times exp^{\frac{\beta(y_n - f(x,w))^2}{2}}\} + Ln\{(\frac{\alpha}{2\pi})^{\frac{M+1}{2}} exp^{-(\frac{\alpha}{2}w^T w)}\}$

$= \sum_{n=1}^{N} \{\frac{-\beta}{2}[y_n - f(x_n, w)]^2 + \frac{1}{2}Ln(\beta) - \frac{1}{2}Ln(2\pi)\} + \frac{M+1}{2}\{Ln(\alpha) - Ln(2\pi)\} - \alpha\frac{w^T w}{2}$

$= -\beta \sum_{n=1}^{N} \{\frac{1}{2}[y_n - f(x_n, w)]^2\} + \frac{1}{2}Ln(\beta) \sum_{n=1}^{N} 1 - \frac{1}{2}Ln(2\pi) \sum_{n=1}^{N} 1 + \frac{M+1}{2}\{Ln(\alpha) - Ln(2\pi)\} - \alpha\frac{w^T w}{2}$

$= -\beta \sum_{n=1}^{N} \{\frac{1}{2}[y_n - f(x_n, w)]^2\} + \frac{N}{2}Ln(\beta) - \frac{N}{2}Ln(2\pi) + \frac{M+1}{2}\{Ln(\alpha) - Ln(2\pi)\} - \alpha\frac{w^T w}{2}$

Now we have sum of square errors,

$E(y|f(x, w)) = \sum_{n=1}^{N} \{\frac{1}{2}[y_n - f(x_n, w)]^2\}$

Therefore, we get $Ln[\text{p(w—x,y,}\alpha, \beta)] = E(y|f(x, w)) - \frac{\alpha}{2}w^T w + constant$

As other terms in the above equation are constant. The term $w^T w$ is a quadratic regularization term added to the equation.

Therefore, $E(w) = E(y|f(x, w)) + \frac{\alpha}{2}w^T w$

Hence, the maximum posterior(MAP) is equivalent to minimizing the regularized sum of square 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.

Solution:

The prediction for the linear model is given by, $f(x, w) = \omega + \sum_{i=1}^{D} w_i x_i$ ...(1) The sum of squares errors/loss function is given by,

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

The Gaussian noise ϵ with zero mean and variance σ^2 is added independently to each of the input variable x_i.

The prediction function for the new linear model is as follows,

$f_1(x, w) = \omega_0 + \sum_{i=1}^{D} w_i(x_i + \epsilon_i) = \omega_0 + \sum_{i=1}^{D}(w_i x_i) + \sum_{i=1}^{D}(w_i \epsilon_i)$

From eq(1), we get $f_1(x, w) = f(x, w) + \sum_{i=1}^{D}(w_i \epsilon_i)$

where noise ϵ_i is added independently.

The new error function for the new linear model is given by,

$L_D^{'} = \frac{1}{2} \sum_{n=1}^{N} \{(f_1(x_n, w) - y_n)^2\}$

$= \frac{1}{2} \sum_{n=1}^{N} \{(f(x_n, w) + \sum_{i=1}^{D}(w_i \epsilon_{ni}) - y_n)^2\}^2$

$= \frac{1}{2} \sum_{n=1}^{N} \{(f(x_n, w) - y_n)^2 + 2(f(x_n, w) - y_n)(\sum_{i=1}^{D}(w_i \epsilon_{ni})) + (\sum_{i=1}^{D}(w_i \epsilon_{ni}))^2\}$

$= \frac{1}{2} \sum_{n=1}^{N} \{(f(x_n, w) - y_n)^2\} + \sum_{n=1}^{N}(f(x_n, w) - y_n)(\sum_{i=1}^{D}(w_i \epsilon_{ni})) + \frac{1}{2} \sum_{n=1}^{N}(\sum_{i=1}^{D}(w_i \epsilon_{ni}))^2$

From eq(2), we get

$L_D^{'} = L_D(w) + \sum_{n=1}^{N}(f(x_n, w) - y_n)(\sum_{i=1}^{D}(w_i \epsilon_{ni})) + \frac{1}{2} \sum_{n=1}^{N}(\sum_{i=1}^{D}(w_i \epsilon_{ni}))^2$

Taking the Expectation,

$E[L_D^{'}] = L_D(w) + \sum_{n=1}^{N}(f(x_n, w) - y_n)(\sum_{i=1}^{D}(w_i [epsilon_{ni}])) + E[\frac{1}{2} \sum_{n=1}^{N}(\sum_{i=1}^{D}(w_i \epsilon_{ni}))^2]$

Now, we have $\mathrm{E}[\epsilon_i] = 0$ given,

$E[L_D^{'}] = L_D(w) + \frac{1}{2} E[\sum_{n=1}^{N}(\sum_{i=1}^{D}(w_i \epsilon_{ni}))^2]$

Evaluating,

$E[\sum_{n=1}^{N}(\sum_{i=1}^{D}(w_i \epsilon_{ni}))^2] = \sum_{n=1}^{N} E[\sum_{i=1}^{D} \sum_{j=1}^{D} w_i w_j \epsilon_{ni} \epsilon_{nj}]$

$= \sum_{n=1}^{N} \{\sum_{i=1}^{D} \sum_{j=1}^{D} w_i w_j E[\epsilon_{ni} \epsilon_{nj}]\}$

$= \sum_{n=1}^{N} \{\sum_{i=1}^{D} \sum_{j=1}^{D} w_i w_j \delta_{ij}\}$

$= \sum_{n=1}^{N} \{\sum_{i=1}^{D} w_i^2\}$

By the above evaluation we get,

$E[L_D^{'}] = L_D(w) + \sum_{i=1}^{D} w_i^2 (\sum_{n=1}^{N} 1)$

$E[L_D^{'}] = L_D(w) + \frac{N}{2} \sum_{i=1}^{D} w_i^2$

We have the regularize term in which bias parameter ω_0 is omitted.

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.

   a) 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.

   Solution:

   The steps involved in the experiment are as follows:

   1) Cleaning the data:

   1.1) Removing the derived attributes from the data-set.

   1.2) Dropping the columns that has all attribute who value is equivalent to 0.

   1.3) Separating the prediction value from the data-set.

   2) Performing Regression:

   2.1) The data-set is shuffled for each regression.

   2.1) The training data is split into training and testing data.

   2.2) Based on the K-fold factor the data is split in equal k folds.

   2.3) The mean square error is calculated and the average value of mean is calculated.