HW12
Machine Leaning

21600004
Kang Seok-Un

Q1.

3.4 (⋆) **WWW** Consider a linear model of the form

$$y(x, \mathbf{w}) = w_0 + \sum_{i=1}^{D} w_i x_i \tag{3.105}$$

together with a sum-of-squares error function of the form

$$E_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 . \tag{3.106}$$

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 E_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.

$$E_D(w) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n+\epsilon_n, w)-t_n\}^2 \quad \cdots \text{adding noise}$$

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

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

$$E_\epsilon\left[\left(\sum_{i=1}^{D}w_i\epsilon_i\right)^2\right] = E_\epsilon\left[\sum_{i=1}^{D}\sum_{j=1}^{D}w_i w_j \epsilon_i \epsilon_j\right] = \sigma^2\sum_{i=1}^{D}\sum_{j=1}^{D}w_i w_j \delta_{ij}$$

$$= \sigma^2\sum_{i=1}^{D}w_i^2 \quad \cdots (1)$$

$$E_\epsilon\left[2(y(x_n,w)-t_n)\left(\sum_{i=1}^{D}w_i\epsilon_i\right)\right] = 2\{y(x_n,w)-t_n\}\sum_{i=1}^{D}E_\epsilon[w_i\epsilon_i]$$

$$= 0 \quad \cdots (2)$$

By (1) and (2), $E_\epsilon[E_D(w)] = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n,w)-t_n\}^2 + \frac{\sigma^2}{2}\sum_{i=1}^{D}w_i^2$

Q2.

3.5 (⋆) **www** Using the technique of Lagrange multipliers, discussed in Appendix E, show that minimization of the regularized error function (3.29) is equivalent to minimizing the unregularized sum-of-squares error (3.12) subject to the constraint (3.30). Discuss the relationship between the parameters η and λ.

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q \tag{3.29}$$

$$\sum_{j=1}^{M}|w_j|^q \leqslant \eta \tag{3.30}$$

Let's rewrite the 3.30 as $\sum_{j=1}^{M}|w_j|^q - \eta \leq 0$ ... (1)

By Using (1), we can write Lagrange function $L(w,\lambda)$

$$L(w,\lambda) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - w^T\phi(w_n)\}^2 + \frac{\lambda}{2}\left|\sum_{j=1}^{M}|w_j|^q - \eta\right| \quad \cdots(2)$$

derivative of 3.29 and $L(w,\lambda)$, then we can know

3.29 and $L(w,\lambda)$ have same dependence on w.

Let's optimal w as w^*, then we can know the relationship between η and w^* through Eq.

$$\eta = \sum_{j=1}^{M}|w_j^*(\lambda)|^q$$

Q3.

EM algorithms are generalized algorithm for obtaining MLE from probability models with latent variables Z.

EM Algorithm consists of E-step and M-step.

E-step:
Through the given parameter θ^t, calculate the likelihood of dataset X and give the label to X by considering the likelihood of the X.

$$Q(\theta; \theta^t) = E[l(\theta; X, Y)|\ X, \theta^t]$$

M-step:
From E-step, we assign the cluster to given dataset X. so, can calculate the MLE. Then, results of the MLE, we can get the newly updated parameter θ^{t+1}.

$$\theta^{t+1} = \min_{\theta} Q(\theta; \theta^t)$$

Do E-step and M-step until the parameter theta converges.

As a result of the EM Algorithm, we can get the best parameter θ^{best} which can describe the given dataset X.

Q4.

Unsupervised Learning:

In the case of unsupervised learning, it refers to a method of training a model without label information of data. Usually, the reason for using this method is used when it is difficult to collect label information for data. For example, a case where the cost of labeling work is very expensive or there is insufficient knowledge or information about training data may be exemplified.

Ex) K-means clustering and Dimensionality reduction methods.

Supervised Learning:

In the case of Supervised-Learning, it refers to a method of training a model using information with a label of specific data. At this time, if there is incorrect information in the label, it greatly affects the performance of the model. Therefore, label work also requires a certain level of quality or higher.

Ex) Since there is label data, classification(kNN, SVM) or regression(Ridge, Lasso) may be exemplified.

Semi-supervised Learning:

Semi-supervised learning can be seen as the middle between supervised learning and unsupervised learning. Semi-supervised learning is used to improve the performance of classifiers by using unlabeled data when there is little labeled data.

Ex) Deep Belief Network

Q5.

Density estimation is to estimate the density (probability) of all values that a variable can have from the data. Since one data is only one side of a variable, a large number of data is required to understand the true nature of the variable. And it is density estimation to estimate the (probability) distribution characteristics of the original variable from the distribution of the observed data.

Therefore, if the probability of a particular data X is learned when Label Y, it is possible to probabilistic infer what the label of that data X will be when there is a particular data X through Bayesian probability. Therefore, I think Density Estimation is important.

Q6.

Parametric model:

In the parametric model, the number of parameters of the model is fixed. Therefore, since what the model needs to learn is clearly determined, it has the advantage of being fast and easy to understand model. However, it has the disadvantage of being less flexible and more suitable for solving simple problems because it has to be assumed that the distribution of data follows a specific distribution.

When using the model every time, it does not require a data set and can use the model with only a tuned parameter, which has the advantage in terms of space complexity.

Ex) Linear Regression, Logistic Regression, Neural Network

Non-parametric model:

It has the advantage of being more flexible because it does not assume that data follow a specific distribution. However, it is often slow, requires larger data, and it is not easy to give a clear explanation of why the model has become like that.

Each time a model is used, a full set of data is required. Therefore, in order to use the model, a lot of space complexity is required.

Ex) KNN, K-Means, Random Forest, Decision Tree