Machine Learning (Homework 1)

Due date : 10/28

1 Bayesian Linear Regression (15%)

For a given input value x, the corresponding target value t is assumed as a Gaussian distribution $p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1})$ and the prior distribution of $\mathbf{w}$ is also assumed as a Gaussian distribution $p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|0, \alpha^{-1}\mathbf{I})$.

A linear regression function is expressed by $y(x, \mathbf{w}) = \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(x)$ where $\boldsymbol{\phi}(x)$ is a basis function. We are not only interested in the value $\mathbf{w}$ but also in making prediction of t for new test data x. We multiply the likelihood function of new data $p(t|x, \mathbf{w}, \beta)$ and the posterior distribution of the training data $p(\mathbf{w}|\mathbf{x}, \mathbf{t})$ and take the integral over $\mathbf{w}$ to find the predictive distribution

$$\int_{-\infty}^{\infty} p(t|x, \mathbf{w}, \beta) p(\mathbf{w}|\mathbf{x}, \mathbf{t}) d\mathbf{w}.$$

Please derive this predictive distribution which is a Gaussian distribution of the form $p(t|x, \mathbf{x}, \mathbf{t}) = \mathcal{N}(t|m(x), s^2(x))$ where

$$m(x) = \beta\boldsymbol{\phi}(x)^{\mathrm{T}}\mathbf{S}\sum_{n=1}^{N}\boldsymbol{\phi}(x_n)t_n$$

$$s^2(x) = \beta^{-1} + \boldsymbol{\phi}(x)^{\mathrm{T}}\mathbf{S}\boldsymbol{\phi}(x).$$

Here, the matrix $\mathbf{S^{-1}}$ is given by $\mathbf{S^{-1}} = \alpha\mathbf{I} + \beta\sum_{n=1}^{N}\boldsymbol{\phi}(x_n)\boldsymbol{\phi}(x_n)^{\mathrm{T}}$. You may use the formulas shown in page 93.

2 Jensens Inequality (10%)

Convexity implies

$$f(\lambda a + (1-\lambda)b) \leq \lambda f(a) + (1-\lambda)f(b) \tag{1}$$

A convex function satisfies

$$f\left(\sum_{i=1}^{M}\lambda_i x_i\right) \leq \sum_{i=1}^{M}\lambda_i f(x_i) \tag{2}$$

where $\lambda_i \geq 0$ and $\sum_i \lambda_i = 1$. Please use the technique of proof by induction to derive equation (2) from equation (1).

3 Polynomial Regression (75%)

In real-world applications, the dimension of data is usually more than one. Here, the California Housing Prices data set is given in (housing.csv). The data set pertains to the houses found in a given California district and some summary stats about them based on the 1990 census data. Please build a regression model for estimation of the values given in item median_house_value (i.e. house price) by applying a polynomial function

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

and minimizing the error function

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

A general polynomial with coefficients for data dimension up to two $\mathbf{x} = [x_1 \; x_2]^\top$ is formed by

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{i=1}^{D} w_i x_i + \sum_{i=1}^{D} \sum_{j=1}^{D} w_{ij} x_i x_j.$$

The data set contains 3 columns of different features that a certain house has. In this exercise, the first 90% samples are used as the Training Set and the last 10% samples are used as the Testing Set.

Data Description

Number of Instances: 20640

Number of Attributes: 4 (3-dimensional input + 1-dimensional target)

Attribute Information:

- Total Rooms: Total number of rooms within a block
- Population: Total number of people residing within a block
- Median Income: Median income for households within a block of houses (measured in tens of thousands of US Dollars)
- Median House Value: Median house value for households within a block (measured in US Dollars)

1. In the training stage, please apply the polynomials of order $M = 1$ to $M = 3$ over the 3-dimensional input data, evaluate the corresponding Root-Mean-Square error ($E_{\text{RMS}} = \sqrt{2E(\mathbf{w})/N}$) on the Training Set and Test Set and plot their RMS error versus order M. Describe in details about what you see in the plot.

2. Please apply the polynomials of order $M = 3$ and select the most contributive attribute or dimension which has the lowest RMS error on the Training Set.

3. Considering the regularized error function

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

where $\|\mathbf{w}\|^2 = \mathbf{w}^T \mathbf{w} = w_0^2 + w_1^2 + w_2^2 + \ldots + w_M^2$. Please set two values for regularization parameter as $\lambda = 0.1$ and $\lambda = 0.001$ and repeat **part 1**. (note: $E_{\text{RMS}} = \sqrt{2E(\mathbf{w})/N}$ is calculated using $E(\mathbf{w})$ not $\widetilde{E}(\mathbf{w})$) Also, plot the regularized regression result on Training Set and Testing Set for various order M from 1 to 3. Compare the result with different λ and describe the difference between **part 1** and **part 3**.