Machine Learning (Homework 1)

Due date : 2019/11/1 23:59:59

1 Bayesian Linear Regression (30%)

We are given the training data $\mathbf{x}$ and $\mathbf{t}$, along with a new test point x, and our goal is to predict target value t. We therefore need to evaluate the predictive distribution $p(t|x, \mathbf{x}, \mathbf{t})$.

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 $p(t|x, \mathbf{w})$ 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

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

For a given input value x, the corresponding target value t is assumed as a Gaussian distribution

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

and the prior distribution of $\mathbf{w}$ is also assumed as a Gaussian distribution

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|0, \alpha^{-1}\mathbf{I})$$

where the parameters α and β are fixed and known. 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 $\mathbf{S}$ is given by $\mathbf{S}^{-1} = \alpha\mathbf{I} + \beta\sum_{n=1}^{N}\boldsymbol{\phi}(x_n)\boldsymbol{\phi}(x)^{\mathrm{T}}$.

2 Linear Regression (70%)

In this homework, you are asked to predict the value of PM2.5 in Hsinchu based on the relevant measurement data. The following two approaches need to be realized respectively:

- Maximum likelihood (ML)
- Maximum *a posteriori* (MAP)

You are given a dataset (dataset_X.csv, dataset_T.csv). The data set contains air measurement data from Hsinchu in 2016, 2017, and 2018. There is a total of 1096 days of historical data. Please use these data samples to build linear regression model, and analyze the relationship between measurement types and PM2.5 value.

Training Dataset

- dataset_X.csv contains 17 different air quality measurement types as Input.
  AMB_TEMP(Celsius), CH4(ppm), CO(ppm), NMHC(ppm), NO(ppb), NO2(ppb), NOx(ppb), O3(ppb), PM10($\mu g/m^3$), RAINFALL(mm/hr), RH(%), SO2(ppb), THC(ppm), WD_HR(degree), WIND_DIREC(degree), WIND_SPEED(m/sec), WS_HR(m/sec)
- dataset_T.csv contains the average PM2.5 measurement for the day as Target.
  PM2.5($\mu g/m^3$)

You can divide dataset into training and validation samples.

1. Feature selection
   In real-world applications, the dimension D of data $\mathbf{x}$ is usually more than one. In the training stage, please fit the data by applying a polynomial function of the form

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{i=1}^{D} w_i x_i + \sum_{i=1}^{D} \sum_{j=1}^{D} w_{ij} x_i x_j \quad (M = 2)$$

   and minimizing the error function

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

   (a) In the feature selection stage, please apply polynomials of order $M = 1$ and $M = 2$ over the dimension $D = 17$ of input data. Please evaluate the corresponding **RMS error** on the training set and validation set. (Hint: $M = 2$ has an overfitting phenomenon.)

   (b) Please analyze the weights of polynomial models for $M = 1$ and select the most contributive attribute which has the lowest RMS error on the Training Dataset.

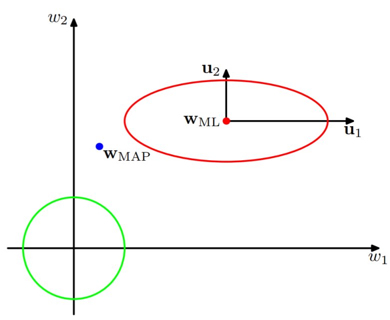
2. Maximum likelihood approach

   (a) Choose some of air quality measurement in dataset_X.csv and design your model. You can choose any basis functions you like and implemented the feature vector.(Hint: Overfitting may happen when the model is too complex. You can do some discussion.)

$$\boldsymbol{\phi}(\mathbf{x}) = [\phi_0(\mathbf{x}), \phi_1(\mathbf{x}), \ldots, \phi_{M-1}(\mathbf{x})]$$

   (b) Apply N-fold cross-validation in your training stage to select at least one hyper-parameter (order, parameter number, ...) for model and do some discussion (under-fitting, overfitting).

3. Maximum *a posteriori* approach



   (a) Use maximum *a posteriori* approach method and repeat 2.(a) and 2.(b). You could choose Gaussian distribution as a prior.

   (b) Compare the result between maximum likelihood approach and maximum *a posteriori* approach.

3 Rules

- Please name the assignment as HW1_StudentID_Name.zip (e.g. HW1_0123456_XXX.zip). The archive file contains source code and report.
- Implementation will be graded by
  - Completeness
  - Algorithm Correctness
  - Model description
  - Discussion
- Using Python and NumPy is encouraged for you in the machine learning area, and matlab is also acceptable.
- Don't use high level toolbox/module functions (e.g. sklearn, polyfit).
- DO NOT PLAGIARISM. (We will check program similarity score.)