Machine Learning 2021

Homework 1

Due date : 2021/10/29 23:59:59

1 Bayesian Linear Regression

Given the training data $\mathbf{x}$ and the corresponding label data $\mathbf{t}$, we want to predict the label t of new test point x. In other words, we wish to evaluate the predictive distribution $p(t|x, \mathbf{x}, \mathbf{t})$.

A linear regression function can be expressed as below where the $\boldsymbol{\phi}(x)$ is a basis function:

$$y(x, \mathbf{w}) = \mathbf{w}^\top \boldsymbol{\phi}(x)$$

In order to make prediction of t for new test data x from the learned $\mathbf{w}$, we will

- multiply the likelihood function of new data $p(t|x, \mathbf{w})$ and the posterior distribution of training set with label set.
- take the integral over $\mathbf{w}$ to find the predictive distribution

$$\begin{aligned} p(t|x, \mathbf{x}, \mathbf{t}) &= \int_{-\infty}^{\infty} p(t, \mathbf{w}|x, \mathbf{x}, \mathbf{t}) d\mathbf{w} && (1) \\ &= \int_{-\infty}^{\infty} p(t|\mathbf{w}, x, \mathbf{x}, \mathbf{t}) p(\mathbf{w}|x, \mathbf{x}, \mathbf{t}) d\mathbf{w} && (2) \\ &= \int_{-\infty}^{\infty} p(t|x, \mathbf{w}) p(\mathbf{w}|\mathbf{x}, \mathbf{t}) d\mathbf{w}. && (3) \end{aligned}$$

Now, please answer the following questions:

1. From Eq.(2) to Eq.(3), we find that $p(t|\mathbf{w}, x, \mathbf{x}, \mathbf{t})$ corresponds to $p(t|x, \mathbf{w})$ and $p(\mathbf{w}|x, \mathbf{x}, \mathbf{t})$ corresponds to $p(\mathbf{w}|\mathbf{x}, \mathbf{t})$. Please explain how this derivation is obtained? (10%) (hint: You may use the concept of **conditional independence**)
2. Prove that the predictive distribution just mentioned is the same with the form

$$p(t|x, \mathbf{x}, \mathbf{t}) = \mathcal{N}(t|m(x), s^2(x))$$

where

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

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

Here, the matrix $\mathbf{S}^{-1}$ is given by $\mathbf{S}^{-1} = \alpha \mathbf{I} + \beta \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) \boldsymbol{\phi}(x_n)^\top$. (20%) (hint: $p(\mathbf{w}|\mathbf{x}, \mathbf{t}) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}) p(\mathbf{w})$ and you may use the formulas shown in page 93.)

2 Linear Regression

In this homework, you need to predict the California housing price based on a set of features. Two learning objectives are implemented:

- Maximum likelihood approach
- Maximum *a posteriori* approach

You are given a dataset (dataset_X.csv, dataset_T.csv) to train your own linear regression model! Dataset provides total 20433 data with 8 features. Can you use these features to predict the Housing Prices of California? One might consider the following steps to start the work:

1. download and check for the dataset
2. create a new Colab or Jupyter notebook file
3. divide the dataset into training and validation

Dataset descriptions

- dataset_X.csv contains 8 different features serving as the inputs longitude, latitude, housing_median_age, total_rooms, total_bedrooms, populations, households, median_income
- dataset_T.csv contains the housing prices regarding as the targets median_house_value

Specifications

- For those problems with **Code Result** at the end, you must show your result in your .ipynb file or you will get no points.
- For those problem with **Explain** at the end, you must have a clear explanation or you will get low points.
- You are also encouraged to have some discussions on those problems which are not marked as **Explain**.

1. Feature selection
   In real-world applications, the dimension of data 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.

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

   (a) In the feature selection stage, please apply polynomials of order $M = 1$ and $M = 2$ over the dimension $D = 8$ input data. Please evaluate the corresponding RMS error on the training set and valid set. (15%) **Code Result**

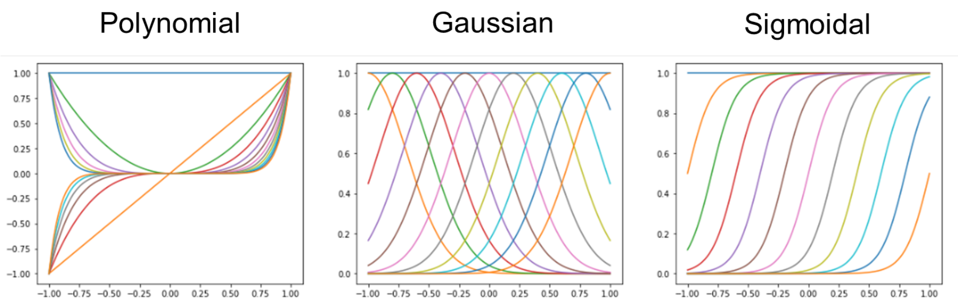
   (b) How will you analysis the weights of polynomial model $M = 1$ and select the most contributive feature? **Code Result, Explain** (10%)

2. Maximum likelihood approach

   (a) Which basis function will you use to further improve your regression model, polynomial, Gaussian, Sigmoid, or hybrid? **Explain** (5%)

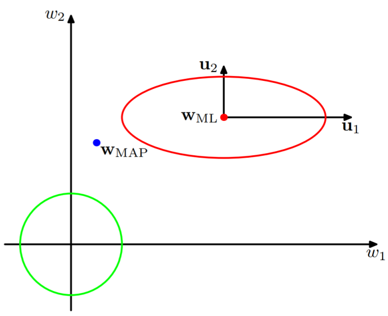
   (b) Introduce the basis function you just decided in (a) to linear regression model and analyze the result you get. (Hint: You might want to discuss about the phenomenon when model becomes too complex.) **Code Result, Explain** (10%)

$$\phi(x) = [\phi_1(x), \phi_2(x), ..., \phi_N(x), \phi_{bias}(x)]$$



   (c) 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 (underfitting, overfitting). **Code Result, Explain** (10%)

3. Maximum *a posteriori* approach



   (a) What is the key difference between maximum likelihood approach and maximum *a posteriori* approach? **Explain** (5%)

(b) Use maximum *a posteriori* approach method to retest the model in **2** you designed. You could choose Gaussian distribution as a prior. **Code Result** (10%)

(c) Compare the result between maximum likelihood approach and maximum *a posteriori* approach. Is it consistent with your conclusion in (a)? **Explain** (5%)

3 Rules

- Please name the assignment as hw1_StudentID.zip (e.g. hw1_0123456.zip).
- In your submission, it needs to contain three files.
  - **.ipynb** file which contains all the results and codes for this homework.
  - **.py** file which is downloaded from the .ipynb file
  - **.pdf** file which is the report that contains your description for this homework.
- Implementation will be graded by
  - Completeness
  - Algorithm Correctness
  - Model description
  - Discussion
- Only Python implementation is acceptable.
- Only the packages we provided is acceptable.
- DO NOT PLAGIARIZE. (We will check program similarity score.)