Foundations of Machine Learning
Assignment -1

General Instructions

Assignments has to be done as a group of 2, deadline for submission is **October 30 (strict deadline).**

Programming has to be done in python (Please write your own code and don’t use python libraries like scikit-learn or other ML packages)

All the assignments need to be submitted as a report (along with code wherever applicable) describing the solutions and observations. Code should have proper documentation, readme files, and instructions to run the code. Assignment contains 2 theory questions and 2 programming questions

Submit as a zip file, please mention name and roll number of both members in the report while only one member need to submit the assignment in the google classroom. Please use Google Classroom to upload your submission by the deadline mentioned above. Your submission should comprise a single folder, named <Your Roll No> Assignment 1, with all your solutions

Please read the department plagiarism policy in the CSE website. Do not engage in any form of cheating - strict penalties will be imposed for both givers and takers.

Question 1 [Theory] : Linear Regression

Question: Consider a linear model of the form -

$$y(x,w) = w_0 + \Sigma\, w_i x_i \quad \text{for } i = 1 \text{ to } D$$

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

$$E_D(w) = \tfrac{1}{2}\, \Sigma\, \{y_n(x_n,w) - t_n\}^2 \quad \text{for } n = 1 \text{ to } N$$

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 $E[\epsilon_i] = 0$ and $E[\epsilon_i \epsilon_j] = \delta_{ij}\sigma^2$, 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.

Question 2 [Theory] : Multi-output regression

Consider the problem where inputs are associated with multiple real valued outputs (K > 1) known as multi output regression (For e.g. predicting student score across different courses).

$$\mathbf{y}(\mathbf{x}, \mathbf{w}) = \mathbf{W}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x})$$

here y is a K-dimensional column vector, W is an M * K matrix of parameters,and φ(x) is an M-dimensional column vector with elements φj(x), with φ0(x) = 1.

1. Provide the expression for the likelihood, and derive ML and MAP estimates of W in the multi output regression case.
2. Consider a multi-output regression problem where we have multiple independent outputs in linear regression. Let's consider a 2 dimensional output vector $y_i \in R^2$. Suppose we have some binary input data, $x_i \in \{0, 1\}$. The training data is as given in the right side. Let us embed each x_i into 2d using the following basis function: φ (0) = $(1, 0)^T$, φ(1) = $(0,1)^T$. The model becomes y =W^Tφ(x) where W = [w_1, w_2] is a 2 x 2 matrix, with both w_1 and w_2 column vectors. Find the MLE for w_1 and w_2

x	y
0	$(-1,-1)^T$
0	$(-1,-2)^T$
0	$(-2,-1)^T$
1	$(1,1)^T$
1	$(1,2)^T$
1	$(2,1)^T$

Question 3 [Programming] : ML and MAP estimation of Poisson Distribution

- Poisson distribution has been introduced to model deaths of soldiers in Prussian army from 1875 to 1894 across corps. Please find the data below.

Table 1. Bortkewitsch's data table, giving numbers of deaths from horse-kicks

	Year																				
Corps	1875	1876	1877	1878	1879	1880	1881	1882	1883	1884	1885	1886	1887	1888	1889	1890	1891	1892	1893	1894	Totals
G	0	2	2	1	0	0	1	1	0	3	0	2	1	0	0	1	0	1	0	1	16
I	0	0	0	2	0	3	0	2	0	0	0	1	1	1	0	2	0	3	1	0	16
II	0	0	0	2	0	2	0	0	1	1	0	0	2	1	1	0	0	2	0	0	12
III	0	0	0	1	1	1	2	0	2	0	0	0	1	0	1	2	1	0	0	0	12
IV	0	1	0	1	1	1	1	0	0	0	0	1	0	0	0	0	1	1	0	0	8
V	0	0	0	0	2	1	0	0	1	0	0	1	0	1	1	1	1	1	1	0	11
VI	0	0	1	0	2	0	0	1	2	0	1	1	3	1	1	1	0	3	0	0	17
VII	1	0	1	0	0	0	1	0	1	1	0	0	2	0	0	2	1	0	2	0	12
VIII	1	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0	1	1	0	1	7
IX	0	0	0	0	0	2	1	1	1	0	2	1	1	0	1	2	0	1	0	0	13
X	0	0	1	1	0	1	0	2	0	2	0	0	0	0	2	1	3	0	1	1	15
XI	0	0	0	0	2	4	0	1	3	0	1	1	1	1	2	1	3	1	3	1	25
XIV	1	1	2	1	1	3	0	4	0	1	0	3	2	1	0	2	1	1	0	0	24
XV	0	1	0	0	0	0	0	1	0	1	1	0	0	0	2	2	0	0	0	0	8
Total	3	5	7	9	10	18	6	14	11	9	5	11	15	6	11	17	12	15	8	4	196

Question 3 [Programming] : ML and MAP estimation of Poisson Distribution

Model the horse kick deaths using the Poisson distribution with different parameters for each of the corps. Learn Poisson distribution parameters for each of the corps using first 13 years of data and make predictions on remaining 7 years and compute the RMSE of predictions for each of the corps.

1. Use maximum likelihood estimation to learn the parameters.
2. Use maximum aposteriori estimation to learn the parameters
    a. Assume appropriate prior distribution over parameters and justify your assumption
    b. Plot prior, likelihood and posterior and provide your observations in terms of mode of the distributions for corps 2, 4 and 6.

Question 4 [Programming]: Bike Sharing Demand

Q1) **Forecast use of a city bikeshare system**

You are provided hourly rental data spanning two years (Data (training and test) available here). The training set is comprised of the first 19 days of each month, while the test set is the 20th to the end of the month. You must predict the total count of bikes rented during each hour covered by the test set, using only information available prior to the rental period. Fit a Poisson regression model to the count data (output). Treat year, month, weekday, hour, holiday, weather, atemp, humidity, windspeed etc. as input features that are combined linearly to determine the rate parameter of the Poisson distribution. Create a 80-20 split of the train data into training, and validation.

1) Explain maximum likelihood estimation in poisson regression and derive the loss function which is used to estimate the parameters.
2) Find statistics of the dataset like mean count per year, month etc.
3) Plot count against any 5 features.
4) Apply L1 and L2 norm regularization over weight vectors, and find the best hyper-parameter settings for the mentioned problem using validation data and report the accuracy on test data for no regularization, L1 norm regularization and L2 norm regularization.
5) Determine most important features determining count of bikes rented.