CS 412 - Machine Learning

Assignment 2 — Spring 2020-2021

Linear Regression, MLE, MAP and Naive Bayes

Due Date: Sunday, April 18, 2021, 23:55

Late Submission: till Tuesday, April 20, 2021, 23:55

(-10pts penalty for each late submission day)

1. In this part of the assignment, you will predict real estate prices using Machine Learning approaches. You are asked to implement a Simple Regression and Polynomial Regression models to perform this task using the scikit-learn library. You will write your findings, results, and interpretations into a report and submit that as well.

   Dataset Description

   Download the dataset from the link. $X_1, X_2, ..., X_6$ are the predictor features and Y is the response. For Polynomial Regression, the degree of the polynomial, p is our hyperparameter.

   (a) Print the shape of the dataset, and show the last five rows of it. Define the predictor features as X and response as Y. Split them as train-validation-test sets with percentages 70-15-15 respectively.

   (b) Draw on a figure the mean squared training and validation error curves as a function of p for $p = 1, 2, 3, 4, 5$. What is the optimal value of p? Justify your reasoning.

   (c) Compute the covariance matrix of the dataset, and observing that, discard d columns from the predictor features. Notice here that d is another hyperparameter. Use the optimal value of p you found in part a, and draw on a figure the mean squared training and validation error curves as a function of d for $d = 0, 1, 2, 3$. What is the optimal value of d? Which features are discarded? Justify your reasoning.

   Implementation

   In this assignment, you are expected to use Google Colab. To start working on your homework, create a Google Colab notebook on your own google drive. You will do your implementations on this file and submit it with the expected outputs. We may just look at your notebook results; so make sure each cell is run and outputs are there.

   Report

   Write an at most 1-2 page summary of your approach to this problem at the end of your notebook; this should be like an abstract of a paper or the executive summary. You should write the report under the cell in the Colab notebook. Your report must include statements such as:

   - Include the problem definition: 1-2 lines
   - Talk about any preprocessing you did, how you handle missing values, and explain your reasoning
   - Talk about train/val/test sets, size, and how to split
   - State what your test results are with the chosen method, parameters: e.g. ”We have obtained the best results with the ….. classifier (parameters=....), giving

classification accuracy of ... % on test data. . . .”

- Comment on the speed of the algorithms and anything else that you deem important/interesting.

You will get full points from here as long as you have a good (enough) summary of your work, regardless of your best performance or what you have decided to talk about in the last few lines. You are expected to write this report on your own.

2. Consider a linear regression problem, with two predictor variables X_1 and X_2 and one response variable Y . The training dataset is given by $\{(xi, yi)\}_{i=1}^{3}$ where the feature matrix is given by

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 3 \\ 1 & 0 & 2 \end{bmatrix}$$

and the response vector is given by $Y = [-2\ 1\ 0]^T$

(a) Write the formula of the linear regression model and the residual sum of squares using the weight vector $\beta = [\beta_0\ \beta_1\ \beta_2]^T$

(b) Recall that out update rule is

$$\beta_j = \beta_j - \alpha \frac{\partial E}{\partial \beta_j}$$

where α is the learning rate hyperparameter and E is the mean squared error. $(MSE = \frac{1}{N} RSS)$ Calculate the partial derivates $\frac{\partial E}{\partial \beta_j}$ using the formula for RSS in part a.

(c) Assume that $\beta = [-1\ 0\ 1]^T$. Perform one update using gradient descent with learning rate $\alpha = 0.5$ What is the resulting weight vector and RSS corresponding to the new weight vector?

3. Let $Y_1, Y_2, ..., Y_6$ be a sequence of i.i.d. random variables where $P(Y_i = 1) = \frac{2\theta}{3}$, $P(Y_i = 2) = \frac{\theta}{3}$ and $P(Y_i = 3) = 1 - \theta$ for $i = 1, 2, ..., 6$. Assume that we observe a realization $(y_1, y_2, ..., y_6) = (1, 2, 3, 3, 1, 1)$ of $(Y_1, Y_2, ..., Y_6)$. Compute the maximum likelihood estimate of θ.

4. (a) 20 pencils and 21 books are randomly distributed among 20 boy and 21 girl students such that each student gets one item. Find the probability that at least one boy gets a book.

   (b) A box contains 2 white and 7 black balls numbered **2012**, 3 white and 5 black balls numbered **2016** and 5 white and 2 black balls numbered **2020**. Let A be the event that the randomly drawn ball is white and B the event that the randomly drawn ball is numbered **2012**. Are A and B independent?

5. The table below provides a training data set containing for risk rate of stock prices with six observations, three categorical predictors, and one qualitative response variable.

Observations	Volatile	Dividend	MACD	Risk
1	Yes	No	Converge	High
2	Yes	Yes	Diverge	High
3	No	No	Diverge	High
4	No	Yes	Converge	Low
5	No	Yes	Linear	Low
6	Yes	No	Linear	High

(a) Train a Naive Bayes Classifier by constructing a one-hot encoding of the features. List each of the parameters of the model and estimate them using Maximum Likelihood Estimation.

(b) We make a new new observation $[Yes, No, Linear]$. What is the Naive Bayes model's class prediction for this observation? Show your calculations.

Office Hours

There will be dense office hours between 17:00 and 20:00 (three hours) every day between 12-17 April. You can join the office hours using the link below:
https://sabanciuniv.zoom.us/j/93721115949?pwd=SW1PNlF2V041Y2RlZVF5REtVRUtFZz09
Meeting ID: 937 2111 5949
Passcode: cs412

Submission Instructions

- You will submit this homework via SUCourse.
- Please read this document again before submitting it.
- Please submit your ”share link” INLINE in Sucourse submissions. That is we should be able to click on the link and go there and run (and possibly also modify) your code. For us to be able to modify, in case of errors, etc, you should get your ”share link” as share with anyone in edit mode.
- Download the .ipynb and the .html file and upload both of them to Sucourse.
- Upload your handwritten answers for the questions 2-5 as a .pdf file. Write your solutions step-by-step and clearly. You can use the blank spaces between the questions to write your answers if you would like to print out the questions and work on it, or you can write your answers on a blank paper. You don’t need to write anything about the question 1 in your .pdf file.
- Please do your assignment individually, do not copy from a friend or the Internet. Plagiarized assignments will receive -100.