COMP3020 : Machine Learning

Assignment 03

2023-11-11

Instructions

- Submit your assignment via Canvas.
- Your submission should include the zip file only (code + written part)
- You may use LaTex or Word to create the written part of your assignment, but you must submit it in PDF format.
- You should be able to finish this assignment within 10 hours

1 Writing Problems

1.1 Join Distribution

Exercise 1.1b (10 points). Let X be a random bit such that

$$X = \begin{cases} +1, & \text{with prob } \frac{1}{2}, \\ -1, & \text{with prob } \frac{1}{2}. \end{cases}$$

Suppose that X is transmitted over a noisy channel so that the observed signal is

$$Y = X + N,$$

Where $N \in$ Gaussian(0,1) is the noise that is independent of the signal X. Find the probabilities $\mathbb{P}[X = +1|Y > 0]$ and $\mathbb{P}[X = -1|Y > 0]$.

1.2 Bayes's Theorem

Exercise 1.2b (10 points).

Imagine you hear a description of a new friend: "He was born in 1996, studied Mechanics at Bach Khoa University, is motivated to earn more money, has an interest in software technology, and is very detail-oriented." Based on this, you might assume he's likely to switch to an IT career, since 70% of your friends who made this transition share these characteristics, while only 10% of those who remain in Mechanics do. However, later you also know that only 1 in 20 people who studied Mechanics actually switch to IT (assuming "Mechanic" and "IT" are the only career options). What is the probability that your friend will switch to IT given such a description?

1.3 Maximum Likelihood Estimation

In this problem we study a single-photon image sensor. Given that photons arrive according to a Poisson distribution, i.e., the probability of observing x_n photons is:

$$\mathbb{P}(X_n = x_n) = \frac{\lambda^{x_n} e^{-\lambda}}{x_n!}$$

Where λ is the (unknown) underlying photon arrival rate. A single-photon image sensor is slightly more complicated in the sense that it does not report X_n but instead reports a truncated version of X_n. Let Y be the random variable denoting the response of the single-photon detector. When photons arrive , the detector generates a binary response "1" when one or more photons are detected, and "0" when no photon is detected.

$$Y_n = \begin{cases} 1, & X_n >= 1, \\ 0, & X_n = 0 \end{cases}$$

Exercise 1.3b (5 points). Find the PMF of Y_n.

Exercise 1.3a (5 points). Suppose we have obtained T independent measurements with realizations $Y_1 = y_1, Y_2 = y_2, \ldots, Y_N = y_n$. Show that the underlying photon arrival rate λ can be estimated by

$$\lambda = -log\Big(1 - \frac{\sum_{n=1}^{N} y_n}{T}\Big)$$

1.4 Bayesian Linear Regression

Exercise 1.4(10 points). We learned about the MLE way to update parameters $\boldsymbol{\theta}$ based on Gradient Descent. Now we will try to update $\boldsymbol{\theta}$ using Bayes Theorem. Consider the Linear Regression again where $y = \boldsymbol{\theta}^T\boldsymbol{x} + \boldsymbol{\varepsilon}$, y is the target variable, $\boldsymbol{x}$ is input variables, $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{I})$ and $\boldsymbol{\theta}$ is the parameters. we consider the prior distribution of parameters $\boldsymbol{\theta} \sim p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta} \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$. Prove that the posterior $p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \mathcal{N}(\boldsymbol{\theta} \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ where

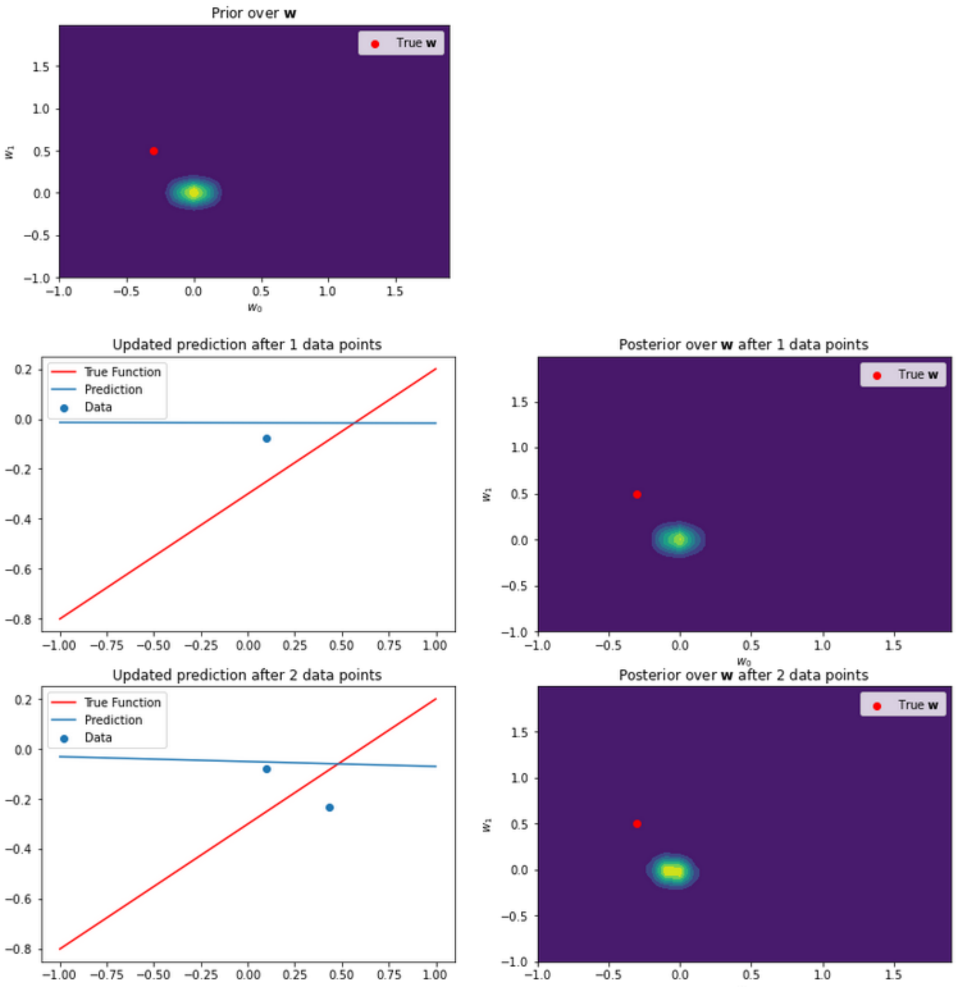
$$\boldsymbol{\mu}_1 = \boldsymbol{\Sigma}_1(\boldsymbol{\Sigma}_{\mathbf{0}}^{-1}\boldsymbol{\mu}_0 + \sigma^{-2}\boldsymbol{X}^T\boldsymbol{y})$$
$$\boldsymbol{\Sigma}_1^{-1} = \boldsymbol{\Sigma}_0^{-1} + \sigma^{-2}\boldsymbol{X}^T\boldsymbol{X}$$

2 Coding Problems

2.1 Bayesian Linear Regression

Exercise 2.1(10 points). You just learned how to update parameters using Bayes Theorem in the above exercise. Open the **bayes_lr.ipynb** to implement it in 2-D case and visualize to see how the distribution changes.

Note: For this question, you are not allowed to import any other package or change the name of the function. At the end, you should be able to see something similar like the figure below.



2.2 Semantic Classification

Exercise 2.2(40 points). Use the data from assignment 2, try to create the best text classification model and submit it to Kaggle `https://www.kaggle.com/t/0887b53b30a64b8aaa32464a335af4a8`. Please use the format ***Your_name-your_student_id*** when submitting your result so we know who you are.

Write a report on what you did and why you choose what you choose. **Important:** Your code should reflect what you write in the report, if we will not count that part in your report.

For this exercise, you will be given:

- **Maximum 20 points** for your report explaining why you did what you did that helps you achieve the result
- **Maximum 5 points** if your code reflect what you write in your report
- **Maximum 15 points** based on your ranking in the competition

You are allowed to use:

- **Model**: K-NN, Decision Tree, Logistic Regression, Neural Network
- **Machine Learning package**: numpy, scipy, scikit-learn, imbalance-learn
- **Data Analysis package**: pandas, matplotlib
- **NLP package**: NLTK, gensim

If you think there are other Python package that might be useful, please ask us before using it.

Some hints for you to get a good result

- You might want to try different data processing methods like tokenizer, stemming and lemmatization, stopwords, Term Frequency, TF-IDF, word2vec,...
- You might want to try to change model's hyperparameters to get best result

Good luck!