Assignment 2: Practical Machine Learning Project

Yujun Liang 12786277

Liangzhu Jiang 13064593

Introduction

In this assignment, we decide to use K nearest neighbor to design an algorithm which can predict a person whether his bone is normal. K nearest neighbor is a lazy learning and non-parametric learning algorithm which can be used for both classification and regression predictive problems. There are many advantages of K nearest neighbor: 1) Easy to understand 2) Less calculating time 3) Easy to train 4) High accuracy 5) suitable for proper capability dataset. Although it has some inevitable problem like it is easily influenced by noise data, we can solve it through preprocessing like deleting meaningless column and only reserve useful column. The inputs of our algorithm are ‘pelvic_ incidence’, ‘pelvic_ tilt numeric’, ‘lumbar_ lordosis_ angle’, ‘sacral_ slope’, ‘pelvic_ radius’ and ‘degree_ spondylolisthesis’ and output is ‘class’ which is the prediction result.

Exploration

During the process of building K nearest neighbor model, we encounter some challenges because we are novices in Python. We conclude five challenges for us: 1) We are not familiar with the parameters and return values of many functions of the sklearn library, such as sklearn's GridSearchCV and logistic regression. 2) The process of modeling general data modeling is not enough. In the process of modeling with sklearn, some data preprocessing processes we don't know how to do it. 3) We are unfamiliar with the algorithm of K nearest neighbor and logistic regression. The meaning of some parameters is not deep enough, which makes it impossible to adjust parameters accurately when modeling. 4) We are unfamiliar with the library handlers of

pandas and matplotlib, which leads to a very slow writing process. 5) We want to display some images and edit text which should use markdown model, but we are not very skillful about it, so we can only learn online step by step.

As we all know, data structure is very important because a data structure is used to store data for the purpose of working on it with various algorithm. In our algorithm, we store our data through saving the csv file in memory as a dataframe. The structure of our data is two-dimensional data table which is also a primitive data structure. We test it through loading the csv file into the algorithm and train it. After our model completed, we will test these data from data structure.

Methodology

We use K nearest neighbor algorithm and do several steps to implement our algorithm. First, after we decide to use K nearest neighbor algorithm, we start looking for a dataset so we define this step as prepare data. Second, after we get dataset, we load dataset and do some preprocessing like separate data into training set and testing set. Then we build our K nearest neighbor model. Next we train our K nearest neighbor model and use different K parameter to test our accuracy until we find the highest accuracy. After that, we also make a K nearest neighbor visualization which shows the range of accuracy of our testing. Last, we train logistic regression to compare with K nearest neighbor so that we can get a comparison result.

Evaluation

As we learned from the introduction to data analysis, the accuracy does not worth that much for evaluating an algorithm. We used the confusion matrix to describe the performance of the model, the confusion matrix is useful for addressing the below mentioned ethical issue which is to detect normal born as abnormal states.

We used LogisticRegression in the comparative study, we import it from sklearn linear model, using Limited-memory BFGS as the solver. By comparison between KNN and LR, the favor is a non-parametric model. KNN

having to keep tracking all training data and find the neighbors, whereas LR can easily extract output from tuned θ coefficients.

Reflections

One member of our group is not major in data analysis, the second member of our group is also first-time using python to implement an algorithm. Due to unfamiliar with coding in python, Our project stuck in the implementation of the project, we do a lot of research, asking assistance from classmate and tutor. Finally, we do have the ability to implement a simple algorithm to address a practical problem. At the very beginning, we planned to design a model to help NBA coaches to decide what kind of role that the player is most suitable, according to their different physical body data, such as body fat, height, weight, standing reach, to predict their position of point guard, shooting guard, small forward etc.. However, we found that the accuracy of the output can only reach 65-70%, we tried changing different k value, preprocessing the data, the output still unsatisfied. We believe the causes may be dataset is not suitable to this algorithm and the prediction label contains 5 classes. This reflects us before we do a machine learning project, we are required to outline the tasks, define the input/output, evaluate the algorithm whether they suit for the data.

Possible improvements

According to our study and consult to the tutor, the possible improvements are listed following.

In the aspect of data processing

1. Explore the feasibility of data grouping, divide the original data into different groups according to certain criteria.
2. Analyze data distribution by using matplotlib and seaborn on python for

advanced information displaying.

In the aspect of modeling.

1. Use sklearn, GirdSearchCV to search for larger parameter spaces for improving the accuracy.
2. Add additional constraints or penalty terms to the existing model to prevent overfitting and improve generalization.

Ethical issue

For discussing the ethical related to our project. Our project aims to support the doctor to judge a patient whether their bone is normal. The prediction is convincing, reasonable and has high accuracy since 6 stats of bone condition is used to evaluate. However, the doctor should not relies on the output, giving hasty judgment without any other consideration, for example, body condition may change under varying circumstances. For example, after strenuous exercise, some of the bone stats would have changed, this may result in the prediction is not feasible. Besides, due to the limitation of the algorithm, around 300 simples are trained, the result may not present the real-world situation.

The project can only be a support method, the output is not recommended to read by the patient, because the effect from a wrong prediction can be negative. When healthy people are detected as abnormal, further detection will be applied. What if the abnormal bone is detected as normal, ignoring the symptoms may lead to delay or miss the treatment.

Link:

Youtube link: https://youtu.be/n0fSgPOaHbs

Github link: https://github.com/Yujun-Liang/UTS_ML2019_ID12786277