Supervised Learning using KNN Classification algorithm with a5a dataset

TBRUEJ

ALHAYAJ Mohammed Jamal Hadi

Eötvös Loránd Tudományegyetem, Budapest, Hungary

jamalmohamad@caesar.elte.hu

**Abstract** In this paper I am illustrating the way of using supervised learning K-nearest neighbor classification algorithm to train/test a5a (adult) data-sets , the goal is to predict whether a person makes over 50K a year.

**Keywords:** KNN *· Supervised Learning* *· Adult data-sets*.

# Introduction

# According to given [1] data-set I found that it is related to adult datasets and the problem to solve is predicting whether a person income exceed 50k/yr.

# This is a classification supervised learning problem.

# In order to solve this problem, I will use K-Nearest Neighbors classification algorithm to train/test the data-set where KNN can be used for both classification and regression predictive industry problems.

**2 K-Nearest Neighbors Algorithm**

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems.

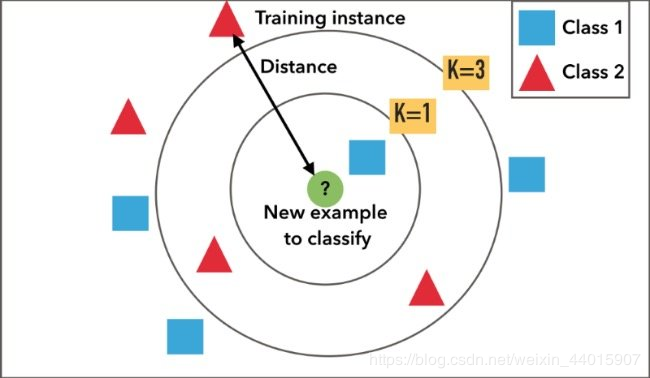
Classification is a type of supervised learning. It specifies the class to which data elements belong to and is best used when the output has finite and discrete values. It predicts a class for an input variable as well.

## Classification: Use Cases

Some of the key areas where classification cases are being used:

* To find whether an email received is a spam or ham
* To find if a person makes over 50K a year
* To find if a bank loan is granted
* To identify if a kid will pass or fail in an examination

**2.1 How KNN Algorithm works?**

****

**Fig(1) KNN with different K values**

**Fig (2)**

**2.2 Choosing the right value for K?**

To select the K that’s right for your data, we run the KNN algorithm several times with different values of K and choose the K that reduces the number of errors we encounter while maintaining the algorithm’s ability to accurately make predictions when it’s given data it hasn’t seen before.

1. **As we decrease the value of K to 1, our predictions become less stable.** Just think for a minute, imagine K=1 and we have a query point surrounded by several reds and one green (I’m thinking about the top left corner of the colored plot above), but the green is the single nearest neighbor. Reasonably, we would think the query point is most likely red, but because K=1, KNN incorrectly predicts that the query point is green.
2. **Inversely, as we increase the value of K, our predictions become more stable due to majority voting / averaging, and thus, more likely to make more accurate predictions (up to a certain point).**

Eventually, we begin to witness an increasing number of errors. It is at this point we know we have pushed the value of K too far.

1. In cases where we are taking a majority vote (e.g. picking the mode in a classification problem) among labels, we usually make K an odd number to have a tiebreaker.

**2.3 KNN Pros and Cons**

## **Pros:**

## - the algorithm is simple and easy to implement

## - the algorithm is versatile. It can be used for both classification and regression

## **Cons:** the algorithm gets significantly slower as the number of examples and/or predictors variables increase.

**3 Experiment**

In this experiment, I assigned a5a datasets, I try to investigate to which datasets belongs and what to predict as an X and Y, so I found that a5a is a binary classification with

- 2 class

- 6414 training size

- 26147 testing size

- 123 features and it is belongs to adults dataset so, I visit UCI and tried to search for adult dataset then I found that the goal is to predict whether a person makes 50k/year.

a5a dataset does not have any labels and the numbers is not pure integer, it is something like this

**-1 3:1 11:1 14:1 19:1 39:1 42:1 55:1 64:1 67:1 73:1 75:1 76:1 80:1 83:1**

**+1 3:1 6:1 18:1 20:1 37:1 40:1 51:1 63:1 71:1 73:1 74:1 76:1 82:1 83:1**

so, I started to analysis the numbers and found that

-1 is <=50k and

+1 is >50k

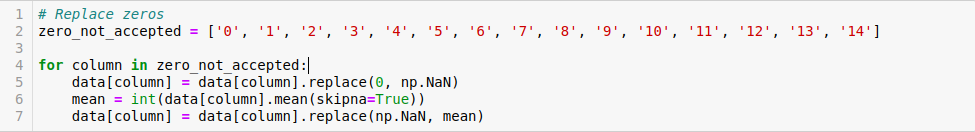
starting from 3:1 to 83:1 is original adult dataset 14 features.

Now, I have to find a way to make the numbers integer in order to start working with the data so, I found that **:1** is repeated to the whole dataset numbers so using Visual studio code I replaced all :1 with empty spaces then I use an online csv conversion tool [**2**] so I ended up with **cvDatasets.csv** file.

After reading csv file using pandas library, I found that there is some columns does not contains number so, in this case, I have to be sure that:

- zero is not accepted in every column if there is a zero

- replace NaN (which is empty numbers) with the Mean

****

In order to apply KNN algorithm, I need to find X and Y, in this case

X = 14 features

Y = -1 & +1 which is (<=50k , >50k)

by calling

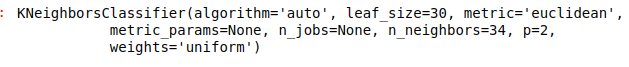
**X\_train, X\_test, y\_train, y\_test = train\_test\_split(x, y, random\_state=0, test\_size=0.2)**

to split train and test, **X\_train – X\_test / y\_train – y\_test**

two most common scenario will raise overfitting and underfitting during prediction,

**4 Results & Discussions**

the KneighborsClassifier gives a result of

**where**

* metrics is a distance calculation between points
* n\_neighbors is number of neighbors point
* p = 2 is number of a5a dataset class

**y\_pred gives a result of**



after evaluating the model as a matrix of y\_test and y\_pred I got this result

[[911 58]

[175 139]]

the result for f1\_score

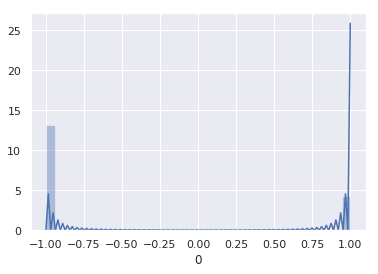
**0.5440313111545988**

while the accuracy score result is

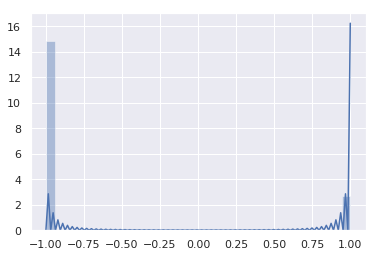
**0.818394388152767**

**the sientist view differ from industry view**

**from youtube video**

****

**Fig (3) graph showing y\_test**

****

**Fig (4) graph showing y\_pred**

**5 Conclusion**

As a conclusion, by using KNN classification algorithm, the problem to predict whether a person makes 50k/year is solved with a reasonable and accepted accuracy score

**References**

1. [https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html#a5a](https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html" \l "a5a)
2. <https://www.youtube.com/watch?v=4HKqjENq9OU>

# Problem

# Why? (describtion from datasets if u want)

**describing data → number of classes, features,**

# How solve it? Ex knn

# What knn can solve?

# Where knn can used?