Machine Learning Assignment

Members

815108 N. Nonjoli

805494 D. Khumalo

1126619 O.N. Mekgwe

Contents

1 Introduction

1.1 What is Supervised Learning?

2 Dataset

2.1 Description

The aim of this dataset is to predict whether a person earns $50,000 per annum. This dataset has 14 variables, is multivariate and the area of focus is social.

Adult Data Set	
Attribute	Values
age	Age of person
workclass	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
fnlwg	continuous
education	Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool
education-num	continuous
marital-status	Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
occupation	Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces
relationship	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
race	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
sex	Female, Male
capital-gain	continuous
capital-loss	continuous
hours-per-week	continuous
native-country	United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad and Tobago, Peru, Hong, Holand-Netherland
48842 Datapoints	

2.2 Terminology

Age	:	Age of person
Work Class	:	Class of work
Final Weight	:	Final weight of how much of the population it represents
Education	:	Education level
Education Number	:	Numeric education level
Occupation	:	Occupation of the person
Relationship	:	Type of relationship
Sex	:	Gender of the person
Capital Gain	:	Rise in value of an investment or real estate that gives it a higher worth than the purchase price
Capital Loss	:	Loss incured when an investment or real estate decreases in value
Hours	:	Average number of working hours per week
Native Country	:	Country of origin

2.3 Targets

2.4 Sample

2.5 What are we predicting?

3 Algorithms

3.1 Decision Tree

3.1.1 Description

Decision Trees are used to classify data, the classification can either be categorical or continuous. They are a type of Supervised Machine Learning. The tree can be described by decision nodes and leaves. The leaves describe the final outcomes, and the decision nodes are where the data is split[2].

3.1.2 How data was handled

The following was done to prepare the data:

- Headers were added and saved to a new file adult.csv
- Rows that had missing variables were removed from the data set.
- Redundant attributes/columns were removed, i.e: education-num
- Irrelevant columns were removed
  - **Final Weight** (**fnlwgt**) weight does not necessarily correlate with income[1].
  - **Capital Gains/Losses** (**capital-gain/capital-loss**) including assets and investments wastes computation resources[3].
  - **Native Country** (**native-country**) obvious, and would waste computation resources.
- Continuous variables were made categorial
  - **Age** (**age**) 3 groups namely, see reference[4].
    - **Baby Boomers** (**baby-boomers**) 50 years old and upwards
    - **Generation X** (**generation-x**) between 35 and 50 years old
    - **Millennials** (**millennials**) between 17 and 34 years old
  - **Work Hours** (**hours-per-week**) 3 groups, see reference[5]
    - **Part-Time** (**part-time**) less than 37 hours
    - **Full-Time** (**full-time**) 37-40 hours
    - **Overtime** (**overtime**) more than 40 hours
- Reduced number of education levels to 7 groups using the United State's education level system, see references
  - **Preschool** (**preschool**) values Preschool[6]
  - **Elementary School** (**elementary**) values 1st-4th[6]
  - **Middle School** (**middle**) values 5th-6th, 7th-8th[6]
  - **High School** (**middle**) values 9th, 10th, 11th, 12th, HS-grad[6]
  - **Undergraduate** (**undergrad**) values Bachelors[7], Assoc-acdm[11]
  - **College** (**college**) values Some-college[9], Assoc-voc[10]
  - **Postgraduate** (**postgrad**) values Prof-school, Masters, Doctorate[8]

3.1.3 Reason

3.1.4 Performance

3.2 Gaussian Naïve Bayes

3.2.1 Description

Gaussian Naïve Bayes is one of Naïve Bayes modelling algorithms used for classification with an assumption of normal distribution of the data features.

3.2.2 How data was handled

Considering that we had some categorical data and given the fact that Gaussian Naïve Bayes works with continuous input data, a label encoder method had to be imported from the sklearn python library in order to convert the categorical data into continuous data. In other words strings matching the category where encoded into numbers.This was needed to allow the GaussianNB() method from the sklearn.naive bayes library to fit the training data into the model. Also the data had to include non-nulls thus a dropna() method was used to remove any missing values in the dataset. Column headers also had to be added into the dataset which mirrored the attributes given to us from the dataset repository (`https://archive.ics.uci.edu/ml/datasets/Adult`). Afterwards training and testing data were split with a test size ratio of 0.2 produced at random, this allowed us to fit the non-biased training data into the Gaussian Naive Bayes model.

3.2.3 Reasons

Since we are given a classification problem, Gaussian Naïve Bayes is a simple and efficient model to implement given that it is one-dimensional. Also from a coding perspective, the algorithm is quick to implement as there is support from the sklearn python library. The Naïve Bayes algorithm works well with large datasets giving an almost accurate and fast way of prediction

3.2.4 Performance

As stated above the algorithm models large data quickly and efficiently. With our dataset the model produced an accuracy level of 0.804514742014742.

3.3 Logistic Regression

3.3.1 Description

3.3.2 How data was handled

3.3.3 Reason

3.3.4 Performance

4 Results

4.1 Findings

4.1.1 Best Algorithm

4.1.2 Worst Algorithm

4.2 Recommendations

References

[1] Does thinking you look fat affect how much money you earn?
https://www.timeslive.co.za/sunday-times/lifestyle/health-and-sex/2018-07-23-does-thinking-you-look-fat-affect-how-much-money-you-earn/

[2] Decision Trees for Classification: A Machine Learning Algorithm
https://www.xoriant.com/blog/product-engineering/decision-trees-machine-learning-algorithm.html

[3] Income tax vs. Capital Gains Tax: What's the Difference?
https://www.investopedia.com/ask/answers/052015/what-difference-between-income-tax-and-capital-gains-tax.asp

[4] What Marketing Content Do Different Age Groups like to Consume?
https://www.digitaldoughnut.com/articles/2016/september/what-marketing-content-different-age-groups-like

[5] Types of Work Schedules
https://www.thebalancecareers.com/types-of-work-schedules-2064278

[6] Educational stage
https://en.wikipedia.org/wiki/Educational_stage#United_States

[7] Undergraduate education
https://en.wikipedia.org/wiki/Undergraduate_education#United_States_system

[8] Postgraduate education
https://en.wikipedia.org/wiki/Postgraduate_education

[9] What Is the Difference Between a Certificate, Diploma and Degree?
https://study.com/articles/What_is_the_Difference_Between_a_Certificate_Diploma_and_Degree.html

[10] Vocational School
https://en.wikipedia.org/wiki/Vocational_school

[11] Associate degree
https://en.wikipedia.org/wiki/Associate_degree