Homework 1 - Introduction to Data Mining (CSE 5243)

Ashwini Joshi
09/13/2027

Table of Contents

This report studies income dataset which has 520 records with a class attribute which tells whether income of a person is greater than 50K or less than 50K on the basis various attributes. First section focuses on analysing the distribution of the data and interesting patterns between the attributes. Second section highlights features of code written for finding closest 'k' matching rows for all the records in the dataset with two different methods. Last section makes an analysis of the output obtained from the code in order to compare the proximities as the value of k differs and method of execution changes.

Section 1: Exploratory Data Analysis:

This section talks about data analysis of given dataset and relationships between different attributes.

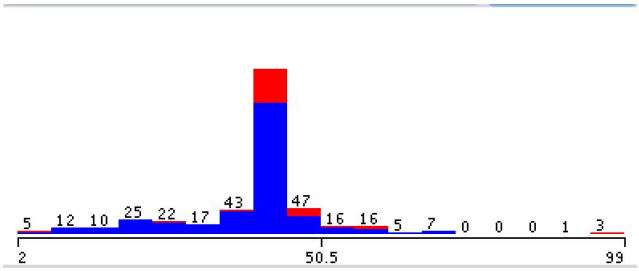
1. **Tables**: Table below summarizes continuous attributes in the given dataset. 'Capital_gain' and 'capital_loss' are sparse data columns.

Value	Minimum	Maximum	Mean	Standard Deviation
Age	17	82	38.388	13.911
fnlwgt	26880	632613	187200.004	100037.914
capital_gain	0	99999	1065.994	6822.673
capital_loss	0	4356	68.915	373.963
hours_per_week	2	99	39.57	12.02

'education' and 'education_cat' are having same content of information with different data types. 'Education_cat' is an ordinal attribute. Remaining all attributes are nominal having a specific set of values.
Also, from the dataset, it is noted that 'workclass', 'occupation' and 'native_country' have missing values.

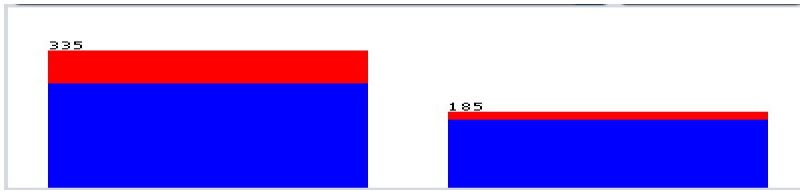
2. **Graphs and Visualisation**: Following graphs show distribution of various attributes. Blue portion indicates income <=50K and red portion indicates income greater than 50K.

Hours per week:



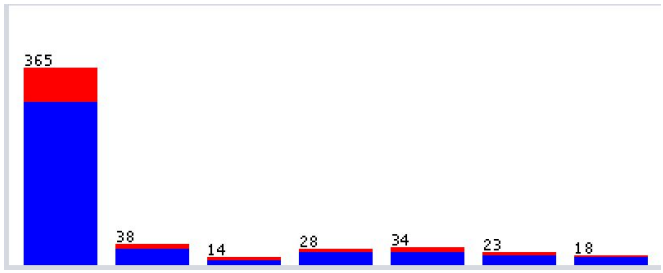
The graph shows the distribution of hours per week which is a continuous attribute with value ranging from 2 to 99. Mean hours is 39.57 and data has some outliers like 99. From the graph, it is observed that it follows somewhat normal distribution around the mean between range of around 30 to 50 hours. It is also seen that most of the people who have income greater than 50K lie in this range of hours.

Gender:



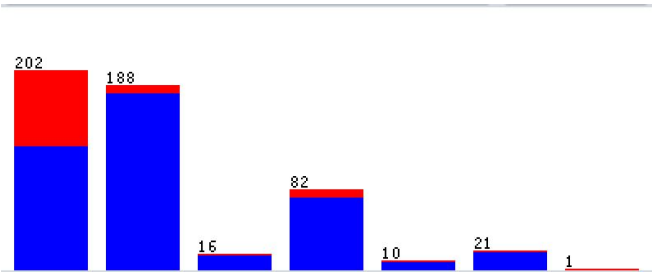
Gender is a nominal (binary) attribute with value 'male' or 'female'. Left bar indicates male while the right bar shows female. From the graph, it is seen that for most of the female workers, income is less than or equal to 50K.

Work Class:



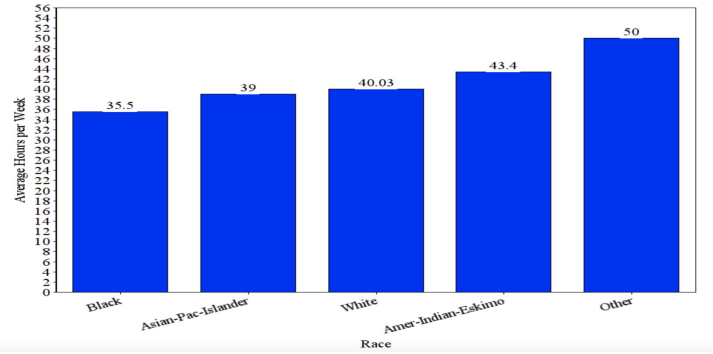
It is a nominal attribute indicating in which sector a person works in. (Private/Gov etc.). For work Class, most of the attributes have value 'Private'(Leftmost bar). It is also inferred that most of the people who have income greater than 50K are from Private class though it is not true in all cases.

Marital Status:



Leftmost bar indicates people with status 'Married-civ-spouse'. Almost all who have income >50K lie in this category. Rest of the categories have income mostly <=50K.

Race vs Average Hours per Week:



In the graph, we see that average 'Hours per Week' increases gradually as the race value changes. This is some useful information which can be used to fill out the missing values if there are any. For example, if a record has attribute hours per week as 35 hours, there is a high chance that race would be 'Black'.

3. **Correlations between attributes**: There are few correlations between various attributes of the data. Though the value is not very large, we might combine this information overall for predicting value of the class variable. Some values of correlation are as follows:
    - Age and income: 0.24.
    - Hours per Week and income: 0.20.
    - Education Category and Income: 0.34.

    From the data, we can see that capital gain and loss are sparse columns. But, most of the people who have a capital gain greater than around 5K have income greater than 50K and people incurring losses have income less than or equal to 50K, both having a few exceptions. This indicates that these attributes might play a deciding role in predicting value of the class attribute – Income.

Section 2: Program Description and Design Choices

This section explains nature of the program, approaches implemented for calculating proximities between records along with the selection preferences made suitable for the dataset. Program is implemented with two approaches - General Approach and Euclidean Distance. Reasons for choosing these methods and analysis is explained below:

1. **Data Preprocessing**:
    - Missing Values: The income dataset has some missing values in attributes - 'Workclass', 'native_country' and 'occupation'. There is no external data available from which missing values for this attributes can be figured out. Also, there is no strong correlation of these columns with any other columns to derive values from that. The simple

approach would be to delete the records. But considering proportion of missing values for the dataset (Almost 27 out of 520), these cannot be ignored because a lot of valuable information would be lost. Hence, missing values are filled with highest occurring value of that particular attribute. For example, missing values in 'native_country' are filled with most frequent value 'United-states'.

- Attribute Transformation: The income dataset contains different types of attributes with different ranges of values. If the values are taken as they are, it might result in some attribute having large weight in the calculation than the other. In order to calculate proximities between various records, normalization is done on the continuous attributes so that they fall in a range of (0, 1). The dataset has six continuous attributes namely - Age, fnlwgt, Capital Gain, Capital Loss, Hours per Week. Formula used for normalization is:

$$v' = \frac{v - min}{max - min}$$

Where v' is the new transformed value, v is the original value and min and max denote smallest and highest value of the attribute respectively.
For general approach, ordinal variable is normalized at the the time of similarity calculation. For Euclidean Distance approach, ordinal attribute 'education_cat' is also transformed using following formula:

$$new_{value} = \frac{old_{value} - 1}{len - 1}$$

Where len is the number of distinct elements of that particular attribute.

2. **Assumptions**:
   - Outliers: As seen in the graphs in section 1 of the report, there are some outliers in the continuous attributes like age. Removing them is not a good idea because a complete row would need to be neglected which results in loss of information again. So, no strategy is implemented for handling the same.

   - Sparse Data: Dataset contains sparse data columns like 'capital_gain' and 'capital_loss'. Since it is not a high dimensional sparse dataset, simpler approaches are implemented to avoid complex functionalities like binarization of nominal attributes in case of cosine similarity.

   - Attributes: 'education' and 'edu_cat' are same attributes with one having string values and the other having numerical values. Hence, only 'edu_cat' is taken into consideration as an ordinal attribute. Moreover, there is no point in taking 'ID' for calculation as it is meaningless. Also, income variable is not taken since it is class variable.

3. **Implementation**:
    - Euclidean Distance Approach: Euclidean Distance is considered as one of the standard measures to calculate proximity or establishing correlations. Also, it is simpler and faster to calculate as compared to other proximity functions. Hence, it is implemented as one of the approaches for proximity calculation.
      For continuous and ordinal attributes(age, fnlwgt, capital_gain, capital_loss, hours_per_week and education_cat), distance between values is taken and squared. For nominal attributes, distance is taken as 0 if the values are same and 1 otherwise. These values are summed up and their square root gives Euclidean distance between two rows.

    - General Approach: Income dataset has various types of attributes like continuous, ordinal and nominal. So, general approach gives a good platform for calculating proximities depending upon the type of the attribute and then combining them with their weights.
      For continuous attributes, L1 norm is taken to calculate similarity. For ordinal attribute, difference between two values is divided by length of the distinct elements in the set. For nominal attributes, approach is same as Euclidean Distance (0 if same, 1 otherwise). Each similarity is multiplied by its corresponding indicator variable delta. These values are summed up and divided by sum of all delta variables which give similarity. It is subtracted from 1 to calculate dissimilarity.

    Both of these approaches calculate dissimilarity in terms of distance between records. Hence, output shows closest matching rows with smallest distance between them.

Section 3: Analysis of the results

This section talks about analysis of the outputs from the code, distribution of the proximities, their variation as the value of parameter k increases for both methods.
Tables shows below are examined in order to establish relationships between class attribute and proximity with parameter k. They show mean and standard deviation of proximities with k=4 for income class attribute.

1. **Tables**:

- Euclidean Mean Value with Income Class:

Row Labels	Average of 1st Prox	Average of 2nd Prox	Average of 3rd Prox	Average of 4th Prox
<=50K	0.647862547	0.800932055	0.895330064	0.983899992
>50K	0.569730675	0.716160688	0.810685939	0.892766151
Grand Total	**0.632837187**	**0.784629869**	**0.879052347**	**0.966374253**

- Euclidean Standard Deviation with Income Class:

Row Labels	StdDev of 1st Prox	StdDev of 2nd Prox	StdDev of 3rd Prox	StdDev of 4th Prox
<=50K	0.455303717	0.450277154	0.428571073	0.384657346
>50K	0.447974675	0.448418434	0.428509581	0.434721066
Grand Total	**0.454520923**	**0.450731154**	**0.429446395**	**0.395971649**

- General Mean Value with Income Class:

Row Labels	Average of 1st Prox	Average of 2nd Prox	Average of 3rd Prox	Average of 4th Prox
<=50K	0.068431073	0.087532101	0.099516939	0.109325316
>50K	0.061808565	0.079183733	0.089921536	0.100459564
Grand Total	**0.067157514**	**0.085926646**	**0.097671669**	**0.107620364**

- General Standard Deviation with Income Class:

Row Labels	StdDev of 1st Prox	StdDev of 2nd Prox	StdDev of 3rd Prox	StdDev of 4th Prox
<=50K	0.053935929	0.057491655	0.057658337	0.055505482
>50K	0.051803877	0.05499779	0.055846883	0.059133425
Grand Total	**0.053547185**	**0.057063613**	**0.057386293**	**0.056271564**

2. **Observations**:
   - Mean and Standard Deviation: Mean value of the proximity increases as the the value of k increases which is obvious as the distance between rows is increasing. Also, difference between the mean values gradually decreases with increasing value of k. Standard deviation falls down with rising value of k.

   - Income class: For income attribute with value '<=50K', average distance is larger than that of '>50K'. It implies that records with value greater than 50K are more similar than value less than or equal to 50K.

   - Approach :
     - ❏ In Euclidean Distance algorithm, record number 346 is the closest for almost 5 rows. In general approach, there are three such examples which have 5 matching records - 48, 346, 356.
     - ❏ There is a slight difference in results when the proximity measure is changed. If first proximity columns for Euclidean and general are compared, almost 85% of the values are same. This percentage remains same even if the value of k changes.

   - Patterns:
     - ❏ An interesting thing to observe is that not all values for first proximity are symmetric. For example, for record number 4, the closest matching row is 184. But for record 184, record number 4 is not the closest. This is true for both the approaches. Almost half of the values are asymmetric.