# E10 Decision Tree and Naive Bayes

Suixin Ou

School of Computer Science
Sun Yat-sen University

December 14, 2021

# Background

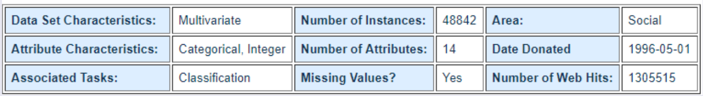
## The Adult Data Set

- The UCI dataset (http://archive.ics.uci.edu/ml/index.php) is the most widely used dataset for machine learning. If you are interested in other datasets in other areas, you can refer to https://www.zhihu.com/question/63383992/answer/222718972.
- The Adult Data Set, sourced from the 1994 U.S. Census Income, is one of many UCI datasets. In this task, you should predict whether income exceeds $50K per year based on census data.

# Task

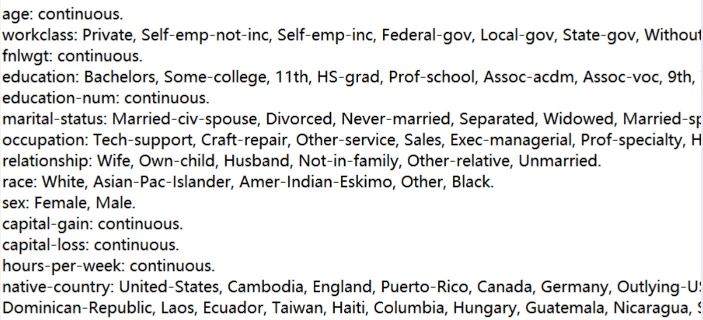
## Description

- Dataset statistics



| | | | | | |
|---|---|---|---|---|---|
| Data Set Characteristics: | Multivariate | Number of Instances: | 48842 | Area: | Social |
| Attribute Characteristics: | Categorical, Integer | Number of Attributes: | 14 | Date Donated | 1996-05-01 |
| Associated Tasks: | Classification | Missing Values? | Yes | Number of Web Hits: | 1305515 |

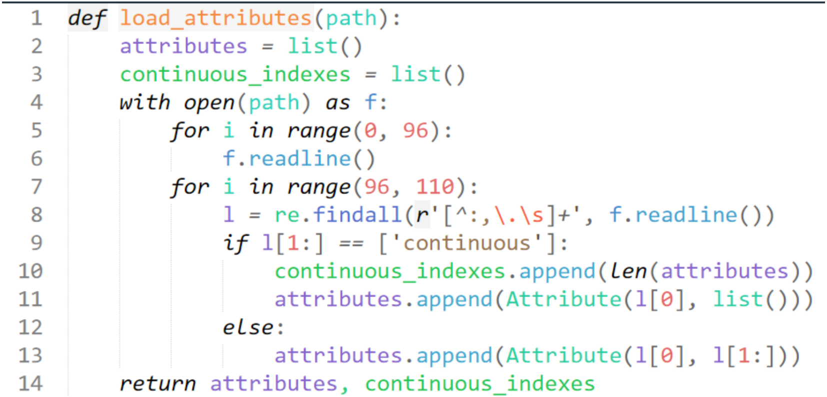
- Domain information



age: continuous.
workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without
fnlwgt: continuous.
education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th,
education-num: continuous.
marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-sp
occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, H
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-U
Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua,

# Solution

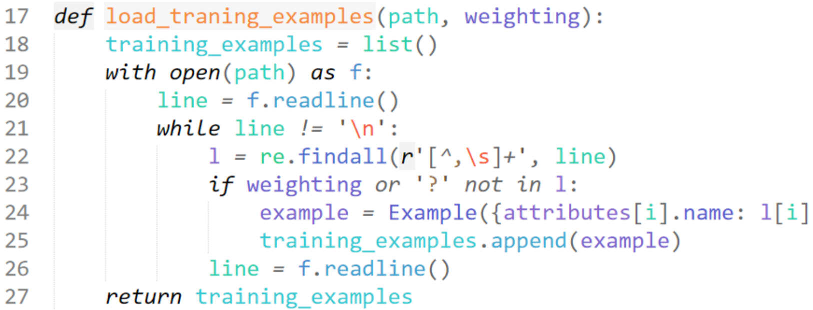
Read the file ”adult.names”



```python
def load_attributes(path):
    attributes = list()
    continuous_indexes = list()
    with open(path) as f:
        for i in range(0, 96):
            f.readline()
        for i in range(96, 110):
            l = re.findall(r'[^:,\.\s]+', f.readline())
            if l[1:] == ['continuous']:
                continuous_indexes.append(len(attributes))
                attributes.append(Attribute(l[0], list()))
            else:
                attributes.append(Attribute(l[0], l[1:]))
    return attributes, continuous_indexes
```

# Solution

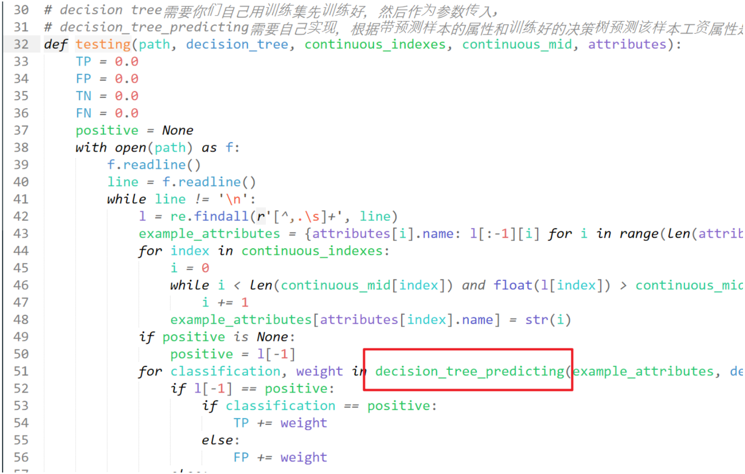
Read the file ”adult.data”



```python
def load_traning_examples(path, weighting):
    training_examples = list()
    with open(path) as f:
        line = f.readline()
        while line != '\n':
            l = re.findall(r'[^,\s]+', line)
            if weighting or '?' not in l:
                example = Example({attributes[i].name: l[i]
                training_examples.append(example)
            line = f.readline()
    return training_examples
```

# Solution

**Please Finish the DT/NB algorithm.** Read the file ”adult.test” for testing



```
# decision tree需要你们自己用训练集先训练好，然后作为参数传入，
# decision_tree_predicting需要自己实现，根据带预测样本的属性和训练好的决策树预测该样本工资属性
def testing(path, decision_tree, continuous_indexes, continuous_mid, attributes):
    TP = 0.0
    FP = 0.0
    TN = 0.0
    FN = 0.0
    positive = None
    with open(path) as f:
        f.readline()
        line = f.readline()
        while line != '\n':
            l = re.findall(r'[^,.\s]+', line)
            example_attributes = {attributes[i].name: l[:-1][i] for i in range(len(attrib
            for index in continuous_indexes:
                i = 0
                while i < len(continuous_mid[index]) and float(l[index]) > continuous_mic
                    i += 1
                example_attributes[attributes[index].name] = str(i)
            if positive is None:
                positive = l[-1]
            for classification, weight in decision_tree_predicting(example_attributes, de
                if l[-1] == positive:
                    if classification == positive:
                        TP += weight
                    else:
                        FP += weight
                else:
```

# Submission

## Submission

pack your report E10_YourNumber.pdf and source code into zip file E10_YourNumber.zip, then send it to ai_course2021@163.com.

# The End