70-511: Statistical Programming
Programming Assignment 4 – Estimating Probabilities

Introduction

Probability is a number that indicates the likelihood of some outcome occurring, where each outcome comes from a set called the *sample space*, denoted by Ω. Probabilities are used in situations where there is uncertainty in data, either due to a lack of sufficient data or some inherent randomness associated with the data. Formally, probability of each outcome x is a value, $p(x)$, that satisfies the following properties:

1. $\forall x \in \Omega\ (p(x) \in [0,1])$ (each probability value has to be between zero and one)

and

2. $\sum_{x \in \Omega} p(x) = 1$ (sum of all probabilities needs to be one)

A set of outcomes defines an *event*. The probability of an event E is defined as

$$P(E) = \sum_{x \in E} p(x)$$

In many applications, it is necessary to estimate probabilities from data. If the data contains nominal (i.e. categorical) values, we can estimate the probability of a particle value occurring in the data by counting the number of instances in which the value occurs. In particular, assume the data consists of *N* instances, which of which is associated with a fixed number of feature values. Then the probability of a particular feature i having a particular value x can be computed as

$$P(feature_i = x) = \frac{\#(instances\ with\ feature_i = x)}{N}$$

We can also compute the *conditional probability* of a particular feature value, given some other features values as

$$P\big(feature_i = x | feature_j = f\big) = \frac{\#(instances\ with\ feature_i = x\ and\ feature_j = f)}{\#(feature_j = f)}$$

Note that the denominator is assumed to be non-zero. Such estimates can then be used for various data analysis applications, such as modeling or machine learning.

Requirements

You are to create a program in Python that performs the following:

1. Asks the user for the number of cars (i.e. data instances).
2. For each car, asks the user to enter the following fields: make, model, type (coupe, sedan, SUV), rating (A, B, C, D, or F). Save the feature values for each car in a DataFrame object.
3. Displays the resulting DataFrame
4. Computes the probability of each rating and outputs to the screen.
5. For each type t, computes the conditional probability of that type, given each of the ratings: $P(type = t | rating = r)$
6. Displays the conditional probabilities to the screen.

Additional Requirements

1. The name of your source code file should be `ProbEst.py`. All your code should be within a single file.
2. You cannot import any package except for **pandas**. You need to use the pandas DataFrame object for storing data.
3. Your code should follow good coding practices, including good use of whitespace and use of both inline and block comments.
4. You need to use meaningful identifier names that conform to standard naming conventions.
5. At the top of each file, you need to put in a block comment with the following information: your name, date, course name, semester, and assignment name.
6. The output of your program should **exactly** match the sample program output given at the end.

What to Turn In

You will turn in the single ProbEst.py file using BlackBoard.

Sample Program Output

```
70-511, [semester] [year]
NAME: [put your name here]
PROGRAMMING ASSIGNMENT #4

Enter the number of car instances: 6

Enter the make,model,type,rating: ford,mustang,coupe,A

Enter the make,model,type,rating: chevy,camaro,coupe,B

Enter the make,model,type,rating: ford,fiesta,sedan,C

Enter the make,model,type,rating: ford,focus,sedan,A

Enter the make,model,type,rating: ford,taurus,sedan,B

Enter the make,model,type,rating: toyota,camry,sedan,B

     make    model   type rating
0    ford  mustang  coupe      A
1   chevy   camaro  coupe      B
2    ford   fiesta  sedan      C
3    ford    focus  sedan      A
4    ford   taurus  sedan      B
5  toyota    camry  sedan      B

Prob(rating=A) = 0.333333
Prob(rating=B) = 0.500000
Prob(rating=C) = 0.166667

Prob(type=coupe|rating=A) = 0.500000
Prob(type=sedan|rating=A) = 0.500000
Prob(type=coupe|rating=B) = 0.333333
Prob(type=sedan|rating=B) = 0.666667
Prob(type=coupe|rating=C) = 0.000000
Prob(type=sedan|rating=C) = 1.000000
```