Report

1. Code structure

Data preprocessing function:

Used to read the data, and the corresponding format conversion

Text

Description automatically generated

Classification function:

To classify several categories for the data sets

Text

Description automatically generated

Mean and variance functions:

Calculate the mean and variance of features

Graphical user interface, text, application

Description automatically generated

Functions of categories and features:

Calculate the mean and variance values ​​for the features corresponding to each category

Graphical user interface, text, application

Description automatically generated

Probability function:

Calculate the probability corresponding to each feature value, and the predicted result probability of an input data

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Calculate the maximum value of the probability in the feature as the predicted value of the input data

2. Preprocessing:

Use numpy to read the txt file, however, every data contains a comma and is not a float data type.

Therefore, first turn these data into the form of a dataframe:

Graphical user interface

Description automatically generated with medium confidence

Then use the function in the dataframe to remove the comma and turn the corresponding feature into a float type.

3. Build the model

(1) Obtain the types of data

(2) Obtain the data corresponding to each feature

(3) Calculate the mean and variance of the data in each to calculate the probability of each feature

(4) Forecast data, select the value with the highest probability in the data of each feature as the predicted probability

4. Results

Text

Description automatically generated with medium confidence

The training set is 80%, the test set is 20%, and the accuracy is 77%

5. Challenge

The problem is that the data structure is not standardized and contains illegal characters. Solution: use dataframe function for data preprocessing

6. Weaknesses

The model only considers the probability distribution of continuous variables, not the change of discrete features. Therefore, subsequent discrete features and continuous features can be combined, which can improve accuracy.