Assignment 2: Binary Classification

Due on Apr 17, 2018, 11:59 PM

P1: Perceptron learning algorithm

In this assignment, you will play with the PLA algorithm.

We use an artificial dataset to study PLA.

Each line of the dataset contains one $(x^n, \hat{y}^n)$ with $x^n \in \mathbb{R}^4$. The first 4 numbers of the line contains the components of x^n orderly, the last number is $\hat{y}^n$. Please initialize your algorithm with w = 0 and b = 0.

P1: Perceptron learning algorihtm (cont'd)

- Your task is to Implement a version of PLA by visiting examples in fixed, pre-determined random cycles throughout the algorithm. Run the algorithm on the data set. Please repeat your experiment for 2000 times, each with a different random seed. What is the average number of updates before the algorithm halts? Plot a histogram (https://en.wikipedia.org/wiki/Histogram) to show the number of updates versus the frequency of the number.

- Save your code as hw2p1.py and the histogram as hist.jpg.

P2: Rich or poor?

1. Task: Determine whether a person makes over 50K a year.
2. Dataset: **ADULT**

Extraction was done by Barry Becker from the 1994 Census database. A set of reasonably clean records was extracted using the following conditions: ((AGE>16) && (AGI>100) && (AFNLWGT>1) && (HRSWK>0)).

3. Reference:

https://archive.ics.uci.edu/ml/datasets/Adult

P2: Attributes of the data set

train.csv 、**test.csv** :

age, workclass, fnlwgt, education, education num, marital-status, occupation relationship, race, sex, capital-gain, capital-loss, hours-per-week, native-country, make over 50K a year or not

```
1 39, State-gov, 77516, Bachelors, 13, Never-married, Adm-clerical, Not-in-family, White, Male, 2174, 0, 40, United-States, <=50K
2 50, Self-emp-not-inc, 83311, Bachelors, 13, Married-civ-spouse, Exec-managerial, Husband, White, Male, 0, 0, 13, United-States, <=50K
3 38, Private, 215646, HS-grad, 9, Divorced, Handlers-cleaners, Not-in-family, White, Male, 0, 0, 40, United-States, <=50K
4 53, Private, 234721, 11th, 7, Married-civ-spouse, Handlers-cleaners, Husband, Black, Male, 0, 0, 40, United-States, <=50K
5 28, Private, 338409, Bachelors, 13, Married-civ-spouse, Prof-specialty, Wife, Black, Female, 0, 0, 40, Cuba, <=50K
6 37, Private, 284582, Masters, 14, Married-civ-spouse, Exec-managerial, Wife, White, Female, 0, 0, 40, United-States, <=50K
7 49, Private, 160187, 9th, 5, Married-spouse-absent, Other-service, Not-in-family, Black, Female, 0, 0, 16, Jamaica, <=50K
8 52, Self-emp-not-inc, 209642, HS-grad, 9, Married-civ-spouse, Exec-managerial, Husband, White, Male, 0, 0, 45, United-States, >50K
```

P2: Provided feature format

X_train, **Y_train**, **X_test** ：

1. discrete: one-hot encoding
2. continuous: remain the same
3. X_train, X_test: each row contains one 106-dim feature represents a sample
4. Y_train: label = 0 means “<= 50K” 、 label = 1 means “ >50K ”

```
age,fnlwgt,sex,capital_gain,capital_loss,hours_per_week, Federal-gov, Local-gov, Never-worked, Private, Self-emp-inc, Self-emp-not-inc, State-gov, Without-pay,?_workclass, 10t
h, 11th, 12th, 1st-4th, 5th-6th, 7th-8th, 9th, Assoc-acdm, Assoc-voc, Bachelors, Doctorate, HS-grad, Masters, Preschool, Prof-school, Some-college, Divorced, Married-AF-spouse
, Married-civ-spouse, Married-spouse-absent, Never-married, Separated, Widowed, Adm-clerical, Armed-Forces, Craft-repair, Exec-managerial, Farming-fishing, Handlers-cleaners,
Machine-op-inspct, Other-service, Priv-house-serv, Prof-specialty, Protective-serv, Sales, Tech-support, Transport-moving,?_occupation, Husband, Not-in-family, Other-relative,
 Own-child, Unmarried, Wife, Amer-Indian-Eskimo, Asian-Pac-Islander, Black, Other, White, Cambodia, Canada, China, Columbia, Cuba, Dominican-Republic, Ecuador, El-Salvador, En
gland, France, Germany, Greece, Guatemala, Haiti, Holand-Netherlands, Honduras, Hong, Hungary, India, Iran, Ireland, Italy, Jamaica, Japan, Laos, Mexico, Nicaragua, Outlying-U
S(Guam-USVI-etc), Peru, Philippines, Poland, Portugal, Puerto-Rico, Scotland, South, Taiwan, Thailand, Trinadad&Tobago, United-States, Vietnam, Yugoslavia,?_native_country
25,226802,1,0,0,40,0,0,0,1,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0
```

Requirements

1. Implement logistic regression with gradient descent
2. Packages for binary classification are not allowed
3. Toolkit Versions:
    a. Only Python3.5+
    b. numpy, pandas and python standard library

Submission Format

Predict the labels of 16281 samples in the test set

1. format: csv
2. The first line is “id,label”. Your predictions start from line 2. First column is id and the second the predicted label, separated by comma.
3. Evaluation metric: Accuracy
4. Save your results as predictions.csv.

Submission

- Please submit hw2p1.py, the trained model, and all other functions required to run your code.

- Zip the your code of P1 and P2 into a single filename.zip file, where filename is your ID.