Java Assignment

20200080168 邢峻铭

Configuration

1. JDK-16
2. IDE: Intellij IDEA
3. Third libraries for machine learning:
   1.javacsv2_1(for reading csv files)
   2.UJMP(for matrix computing)
   3.the required third libraries are packaged in "*\JavaProject\ML_Lib*" folder and loaded in the project

Introduction

1. Solution class is about machine learning
2. Reader class reads data in FireData class
3. Test class can read data in FireData class and sort them
4. FireDataMatrix class stores the data in matrix
5. LogisticRegression class has some method for logistic regression

Manipulations on the data

1. The data file is in *"\JavaProject\Data\newFireData"*.
2. Firstly read all the data in the file in a **FireData** type ArrayList.
3. To make the order of samples random, a **suffle operation** is required after read in all the data in file
4. There are totally 243 samples in the data, among them 180 will be using for trianing the **regression equation**. And the rest are for testing of the regression.

Mathematic knowledge

1. Regular equation

Matrix X contains m samples in raw. Each record has n feature value
y is a column vector which records if it is *fire*(y_i = 1) or it is *not fire*(y_i = 0)
Then the vector of theta can be calculated as the following equation
More references about regular equation

$$\theta = (X^T * X)^{-1} * X^T * y$$

2. Sigmoid
   See reference here

$$g(z) = 1/(1 + e^{-z})$$

About testing

1. First is to understand some management figure we frequently use for testing a logistic regression.
   Check here for a quick reference of *Precision, Recall, F_Beta management*
2. This training example is to estimate the probability of a fire disaster in a forest,so the costs of **differet misjudgments** are significantly different.
   It is evidently that the cost of misjudging a ***fire***(actually there won't be fire, but the result of the regression equation says there will be fire) is much less that of a misjudement of ***not fire***(will be on fire but the equation doesn't report).So the weight of **Recall** should be more significant than **Precision**. That is also why I choose the **Beta** of **F_Beta** bigger than one.