Machine Learning
LAB1: Regression

Jie Li

nijanice@163.com

LAB

- LAB01: Regression 4 weeks
- LAB02: Classification 4 weeks
- LAB03: Clustering 4 weeks

Recommendation environment

- python3.7
- sklearn
- matplotlib
- numpy

Goal

- Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables X).

Example

- sklearn.datasets.load_boston()

  波士顿房价数据集

  美国人口普查局收集的美国马萨诸塞州波士顿住房价格的有关信息

  506个样本，

  13个输入变量和1个输出变量

Example

- sklearn.datasets.load-boston()

CRIM：城镇人均犯罪率。

ZN：住宅用地超过 25000 sq.ft. 的比例。

INDUS：城镇非零售商用土地的比例。

CHAS：查理斯河空变量（如果边界是河流，则为1；否则为0）。

NOX：一氧化氮浓度。

RM：住宅平均房间数。

AGE：1940 年之前建成的自用房屋比例。

DIS：到波士顿五个中心区域的加权距离。

RAD：辐射性公路的接近指数。

TAX：每 10000 美元的全值财产税率。

PTRATIO：城镇师生比例。

B：1000（Bk-0.63）^ 2，其中 Bk 指代城镇中黑人的比例。

LSTAT：人口中地位低下者的比例。

MEDV：自住房的平均房价，以千美元计。

Example

- sklearn.datasets.load-boston()

使用常见的回归模型，对于该数据集进行机器学习回归分析

代码示例

Example

- sklearn.datasets.load-boston()

  有哪些优化模型的思路？

  去除意义较小的输入变量参数

  探究输入变量参数之间的关联性，数据升维

  数据降维？特征空间变换？启发式方法？

Process

- Data prepare
- Data clean
- Model construct
- Train & Test
- Plot result
- Optimize & Review

Dataset

- UCI数据库的“Concrete Compressive Strength Data Set”
  混凝土抗压强度数据集

1030个样本，
8个输入变量和1个输出变量

Dataset

- UCI数据库的“Concrete Compressive Strength Data Set”

每个样本包含有水泥量、高炉矿渣粉量、粉煤灰量、水量、减水剂量、粗骨料量、细骨料量、使用时间共8种输入特征值，每个样本同时也包含一个输出特征值即混凝土抗压强度数值

https://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength

Work

- 实验组成

Data prepare - Data clean -
Model construct - Train & Test -
Plot result - Optimize & Review

- 实验报告要求详见实验报告说明文件

- Extra Credit:手写实现回归算法模型

Work in teams

- Work in a team with 1-4 members
- All the team members finish the lab together
- One team only need to submit one lab report for the whole team

Lab report

- There is no requirement of number of words for the report
- There is no requirement of format for the report
- You are free to arrange the content of your report, but it should contain the workflow which described in the page 9

Presentation

- All the team need to do presentation for their lab work

- One team will have 7 minutes to introduce their work, and 5 minutes for answering questions

Submit

- Start Time: 2021/3/15
- End Time: 2021/4/11 21:00
- You will have totally five late days for all the three labs
- Submit report + code to TA's email

2030777@tongji.edu.cn

Hope you enjoy lab01!
Feel free to ask questions about lab01!

Jie Li

nijanice@163.com

Yuxuan Wei

2030777@tongji.edu.cn