CS767

SPR23

Assignment 1

Yiduo Feng

01/28/2023

**Assignment 1 (20 points)**

Under the phrase “Machine Learning” in Wikipedia, there is a section named “Applications”.

(The URL to the Application section is:

<https://en.wikipedia.org/wiki/Machine_learning#Applications> )

For 10 of the applications answer following questions:

* Write a sentence about the main application(s) of Machine learning in each and every application in the list.
* **Bioinformatics:** For genomics, proteomics and so on, they use image recognition or logistic regression to check abstract features of genes.
* **Computer vision:** object detection and classification. According to the characteristics of the specified object, learn and advance the corresponding graphics from the video.
* **DNA sequence classification:** Use classification models to predict DNA sequences and learn DNA samples through data mining.
* **Economics:** Machine learning can predict multiple economic situations at the same time, reducing human consumption and errors caused by human calculations.
* **General game playing:** Machine learning can calculate rewards to evaluate player levels through reinforcement learning.
* **Handwriting recognition**: Similar to image recognition, handwriting recognition uses machine learning to read the characteristics of words and numbers to verify and classify words.
* **Information retrieval**: Information retrieval needs to find the corresponding documents among tens of thousands of documents, which requires the use of Clustering to regroup all questions and make the search faster in continuous iterative learning.
* **Knowledge graph embedding:** In knowledge graph embedding, the amount added will become the amount of change used in the prediction. In composite correlation graphs, machine learning is able to more accurately calculate the errors of errors.
* **Speech recognition**: Speech recognition uses neural networks and converts voice into text through machine learning.
* **Natural language processing:** Siri is an example of natural language processing, through speech recognition and text analysis, and finally information retrieval to deliver information to users.
* If you have to choose one, which of the learning methods (supervised, unsupervised or reinforcement learning) would you choose for each application. Explain in one sentence “why?”

**Bioinformatics:** supervised. For example, retain biometric information such as fingerprints to improve security.

**Computer vision:** supervised. In computer vision, we need to check if the item is what we want, so we need classification.

**DNA sequence classification**: supervised. Using classification to check DNA structure is supervised.

**Economics:** supervised. Using regression to get the increasing and decreasing of money.

**General game playing:** reinforcement learning. In the game, we calculate the reward every time we take a step, and finally maximize the reward.

**Handwriting recognition**: supervised. We need to use classification to recognize 26 letters.

**Information retrieval**: unsupervised. Using cluster to find the field of information.

**Knowledge graph embedding:** supervised. It learns vector representations of nodes and edges of labeled, directed multigraphs.

**Speech recognition**: supervised. Like Siri, we need to recognize person’s word by supervised learning.

**Natural language processing:** supervised. It allows us to use the unlabeled data itself to generate labels without manual data labeling

* Between two big groups of ML methods, regression and classification, which one fits the best for each application? Explain in one sentence “why?”
* **Bioinformatics:** classification. Feedback of checking fingerprint should be yes or no.
* **Computer vision:** classification. Feedback of picture should be yes or no.
* **DNA sequence classification**: classification. DNA sequence is not numerical**.**
* **Economics:** regression. Increasing or decreasing of stock should be quantitative.
* **General game playing:** regression. We need to compare the increase in rewards.
* **Handwriting recognition**: classification. 26 letters can be consider as 26 “types”.
* **Information retrieval**: classification. Create different classifications using cluster before searching.
* **Knowledge graph embedding:** regression. The feedback of vector should be number.
* **Speech recognition**: classification. Similar to Handwriting recognition.
* **Natural language processing:** classification. Text classification need to understand text, sign, and semantic.

**Assignment number 2 (20 points):**

Find two interesting websites about Machine Learning and/or Python on internet, for example following web sites:

ML: http://www.kaggle.com

ML: https://registary.opendata.aws

Python: <http://learnpython.org>/

Python: <https://docs.python.org/3/tutorial/>

* Write 3 most interesting things about each of the websites.

` <https://www.geeksforgeeks.org/python-programming-language/?ref=shm>

1. It sometimes gives gif picture for us to learn algorithms.
2. We can even find job in this website.
3. You can run python code in this website.

<https://archive.ics.uci.edu/ml/datasets.php?format=&task=&att=&area=comp&numAtt=&numIns=&type=&sort=nameUp&view=table>

1. A lot of data sources in this website.
2. It tells you which one is label class and which are features for each source.
3. It separate the sources by classification and regression.

**Assignment number 3 (60 points):**

Predict median house value of California houses as a linear function of other parameters using California Housing dataset.

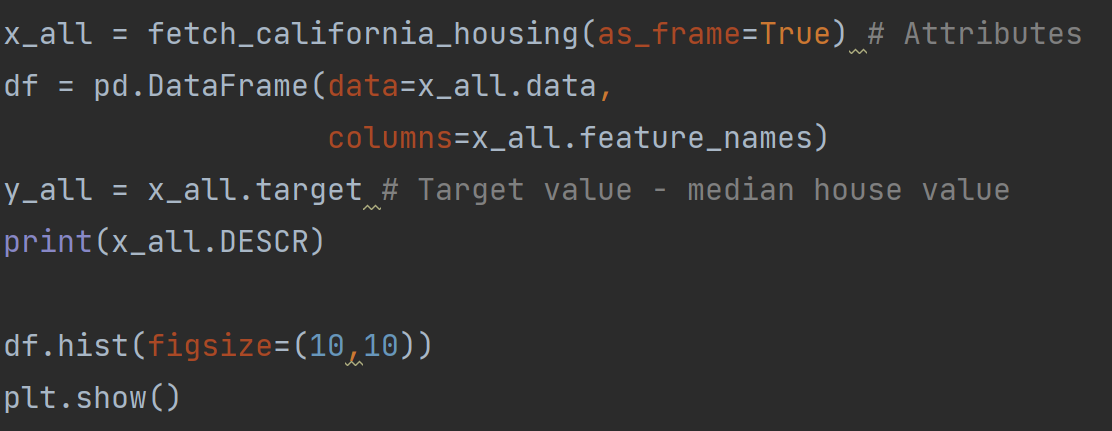
You can load California Housing dataset as follows:

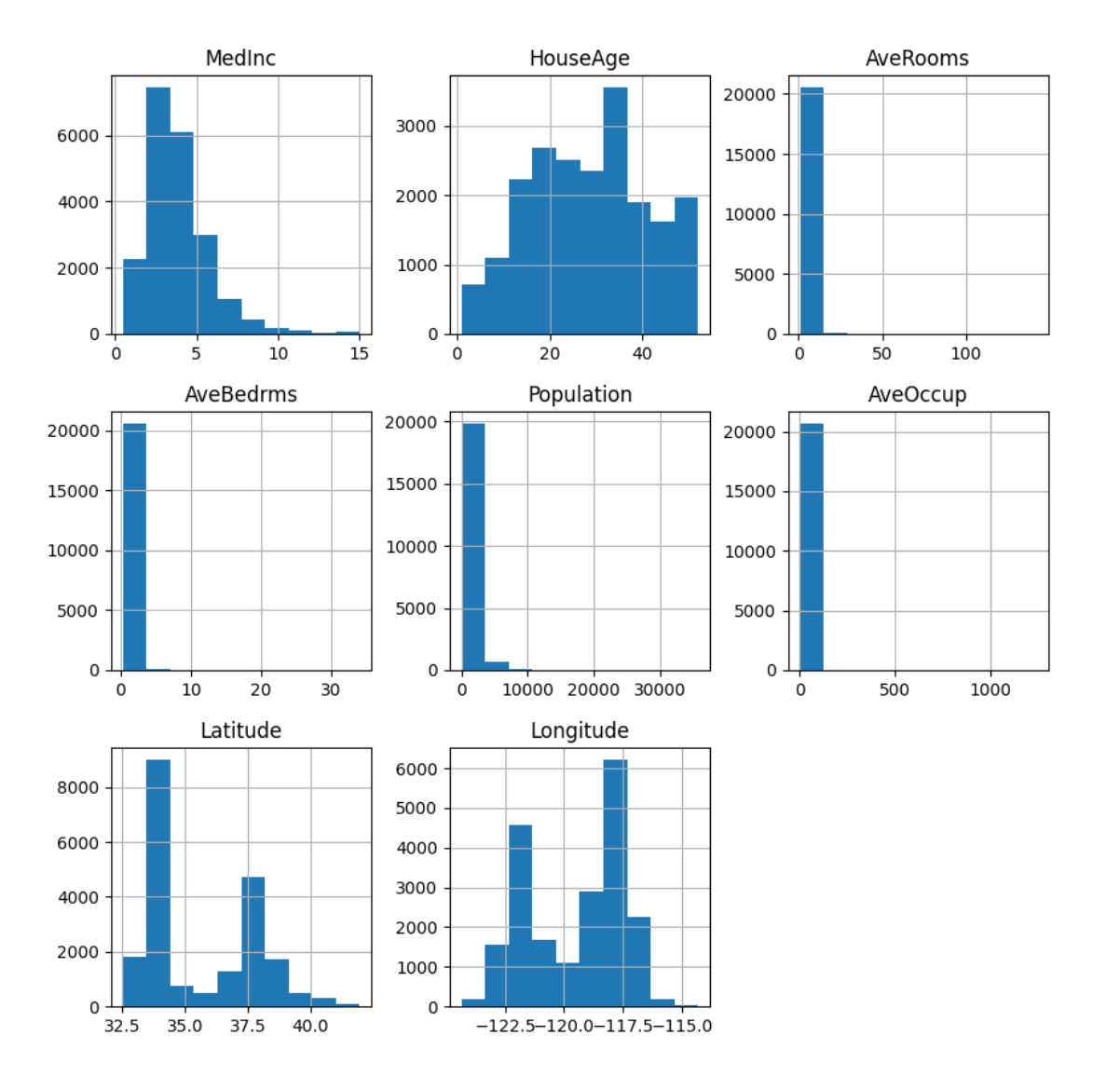


You can learn about the data by looking at the DESCR method of the object.



1. **(10 points)** Plot histogram of numerical attributes of the data.





1. **(35 points)** Predict median price of the houses as a linear function of other parameters using Huber loss.

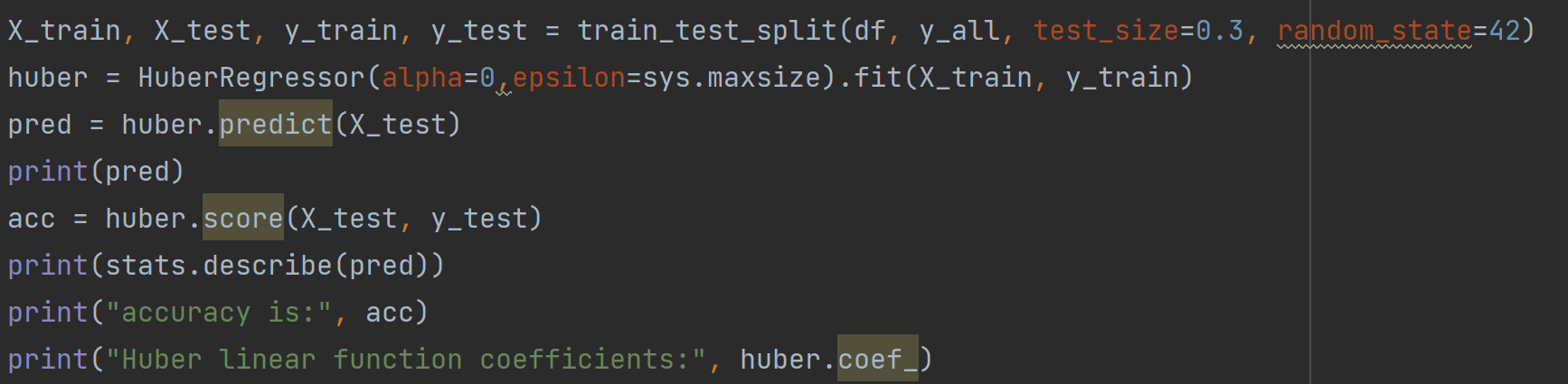
Use following Scikit-learn HuberRegressor. Keep all the default values as they are, except set the alpha to zero and set the epsilon to infinity.

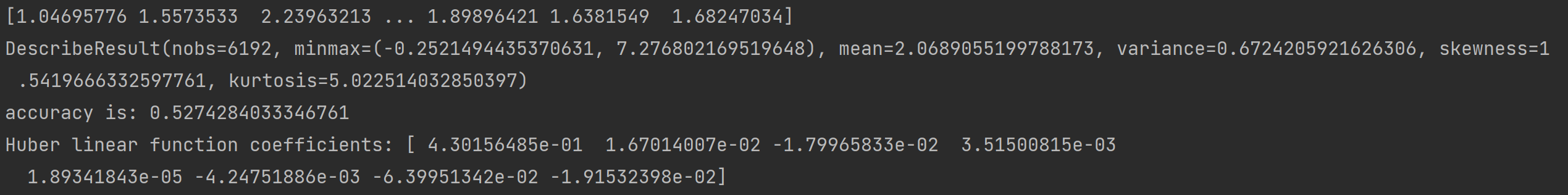
*class*sklearn.linear\_model.**HuberRegressor**

Split data to training and test set (70, 30) using following function.

from sklearn.cross\_validation import train\_test\_split

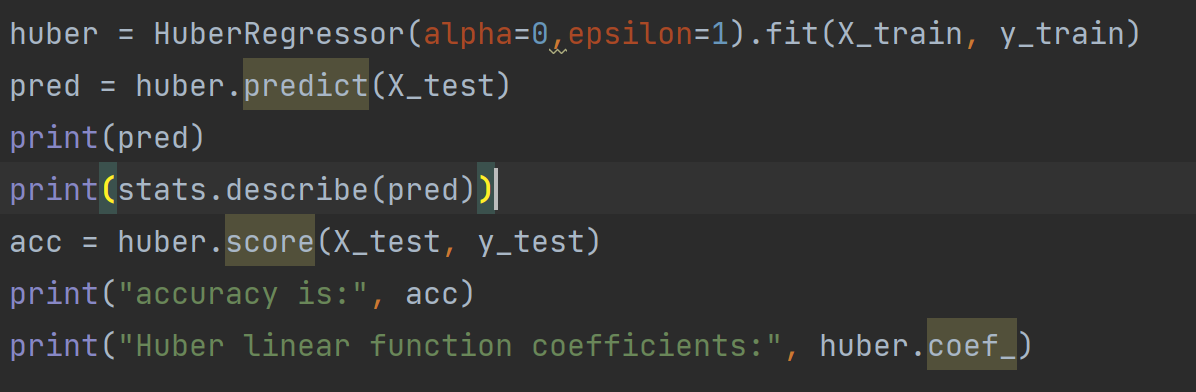
Comment on the results.

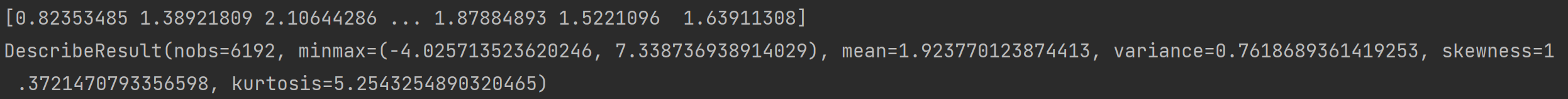


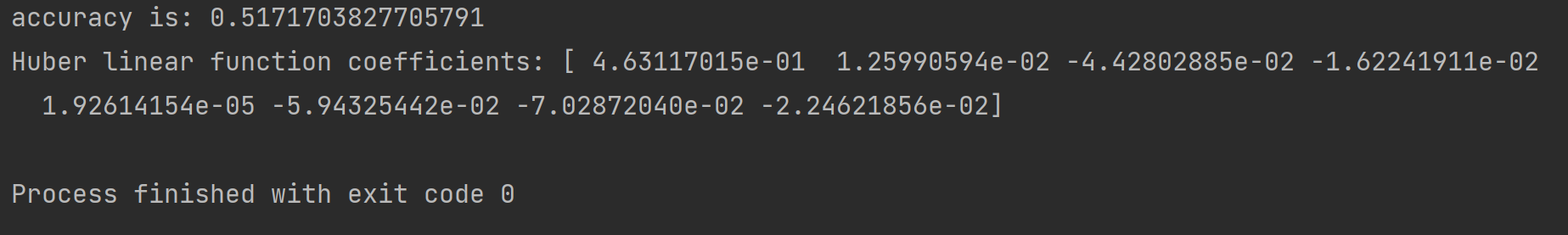


The difference between min and max is about 7, and the mean is about 2. The accuracy of the prediction is not very high.

1. **(15 points)** Repeat item 2 with epsilon set to one. Comment on the results.







After changing the epsilon to 1, the accuracy of results doesn’t improve, but the min of prediction become -4 that is much smaller than that when epsilon is infinity.