Machine Learning and Pattern Recognition MC886/MO444

University of Campinas (UNICAMP), Institute of Computing (IC)
Assignment #1, 2018s2, Prof. Sandra Avila

Objective

Explore **linear regression** alternatives and come up with the best possible model to the problems, avoiding overfitting. In particular, predict the price of diamonds from their attributes (e.g., depth, clarity, color) using the Diamonds dataset (`https://www.kaggle.com/shivam2503/diamonds`).

Activities

1. Perform Linear Regression. You should implement your solution and compare it with `sklearn.linear_model.SGDRegressor` ("linear model fitted by minimizing a regularized empirical loss with SGD"[1]). What are the conclusions?
2. Use the specified training/test data for providing your results and avoid overfitting.
   Keep in mind that friends don't let friends use testing data for training.
3. Plot the cost function vs. number of iterations in the training set and analyze the model complexity. What are the conclusions? What are the actions after such analyses?
4. Use different Gradient Descent (GD) learning rates when optimizing. Compare the GD-based solutions with Normal Equation. You should implement your solutions. What are the conclusions?
5. Prepare a 4-page (max.) report with all your findings. It is UP TO YOU to convince the reader that you are proficient on linear regression and the choices it entails.

Dataset

The Diamonds dataset contains the prices and attributes of almost 54,000 diamonds.

Dataset Information:

- You should respect the following traininig/test split: $45,849$ training examples, and $8,091$ test examples.
- There are 9 attributes as follows:

  1: **carat**: weight of the diamond (0.2–5.01)

  2: **cut**: quality of the cut (Fair, Good, Very Good, Premium, Ideal)

  3: **color**: diamond color, from J (worst) to D (best)

  4: **clarity**: a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))

  5: **x**: length in mm (0–10.74)

  6: **y**: width in mm (0–58.9)

  7: **z**: depth in mm (0–31.8)

  8: **depth**: total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43–79)

  9: **table**: width of top of diamond relative to widest point (43–95)

  target price: price in US dollars

[1] `http://scikit-learn.org`

- The data is available at `https://www.dropbox.com/s/hm2kim0j9gwr0vk/diamonds-dataset.zip`

Deadline

Tuesday, **September 4th** in the beginning of the class, 7 pm.
Penalty policy for late submission: You are not encouraged to submit your assignment after due date. However, in case you did, your grade will be penalized as follows:

- September 5th 7pm : grade * 0.75
- September 6th 7pm : grade * 0.5
- September 7th 7pm : grade * 0.25

Submission

On the deadline day, bring your 4-page printed report. The template for report is available at `https://www.dropbox.com/s/nc6d89otr8ekvjd/report-model.zip`. Please, print on both sides of the page. The report should be written in Portuguese or English.
Submit a zip file, with the code and the report (PDF file), via Moodle.
This activity is NOT individual, it must be done in pairs (two-person group).