Pattern Recognition
Homework 1 Announcement

Lastest update: 2023.03.08 13:30

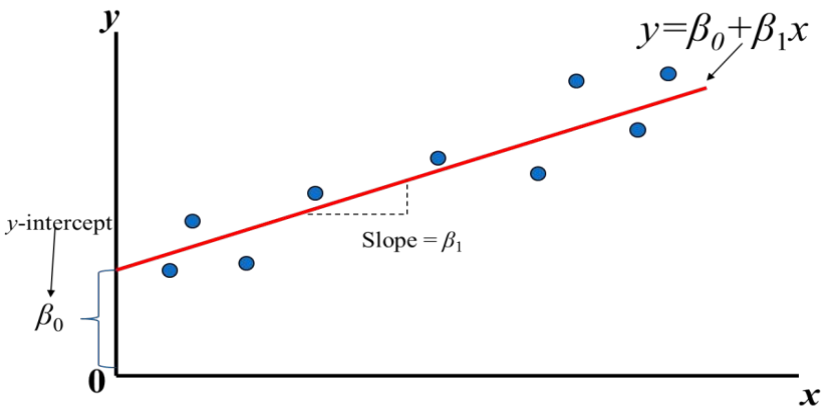
Homework 1

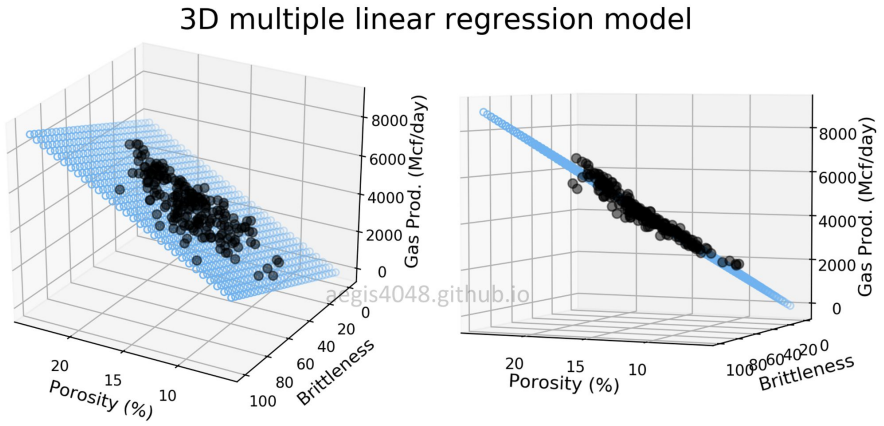
- Deadline: Mar. 22 , Wed. at 23:59
  - Code assignment (70%): Implement linear regression using only numpy
  - Questions (30%): Write your answer in detail on pdf

- Question: Link
- Sample code: Link
- Dataset: Link
- Report template: Link
- Sample prediction file: Link

Linear Regression

- Find the value of weight and intercept





https://aegis4048.github.io/mutiple_linear_regression_and_visualization_in_python

How to find β0 and β1?

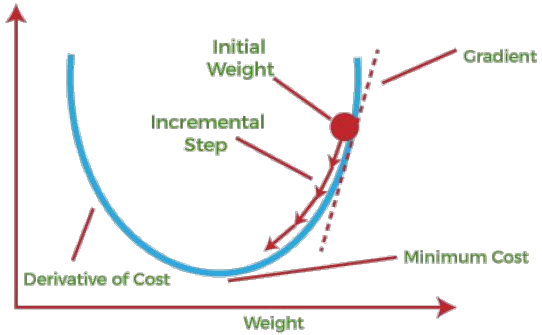
Trial and Error

Closed form solution

Gradient Descent

β0 = -2, -1, 0, 1, 2,...
β1 = 1, 2, 3, 4, 5,...

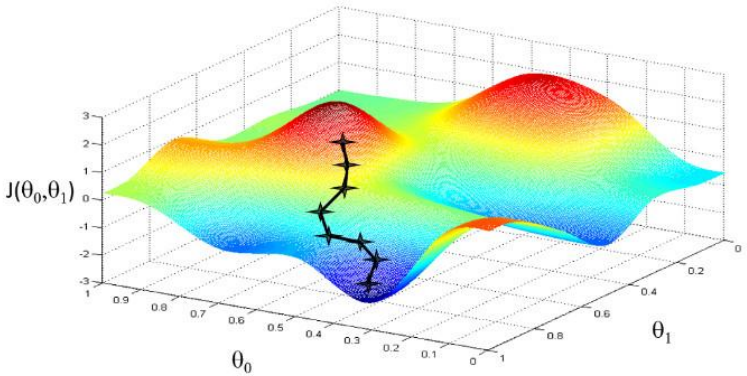
$$\hat{\beta} = (X^T.X)^{-1}X^T.Y$$



https://www.javatpoint.com/gradient-descent-in-machine-learning

Gradient Descent

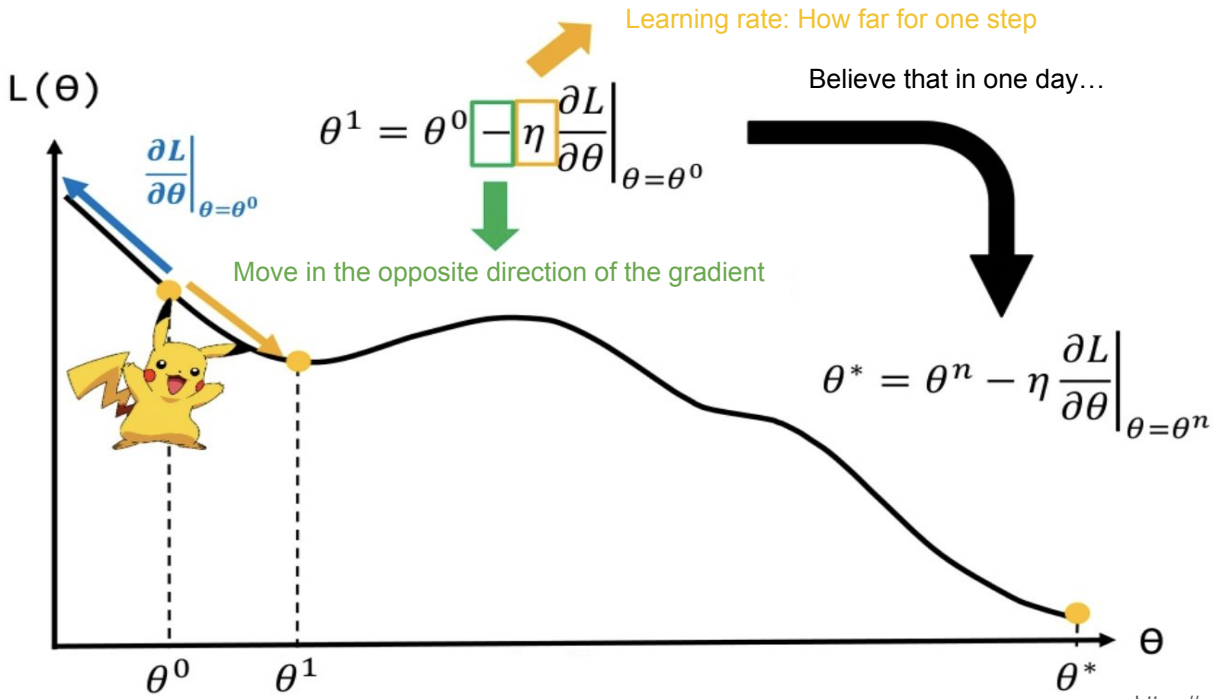
- x-axis and y-axis represent the value of weights
- z-axis represents the loss of the corresponding weights
- Goal: Find the weights that minimize the loss



https://www.linkedin.com/pulse/gradient-descent-visualization-python-ibrahim-sobh/

Gradient Descent

- Gradient tells the direction



https://www.youtube.com/watch?v=yKKNr-QKz2Q

Dataset

- Medical Insurance Charges Forecast
- Training set: 938
- Validation set: 200
- Testing set: 200

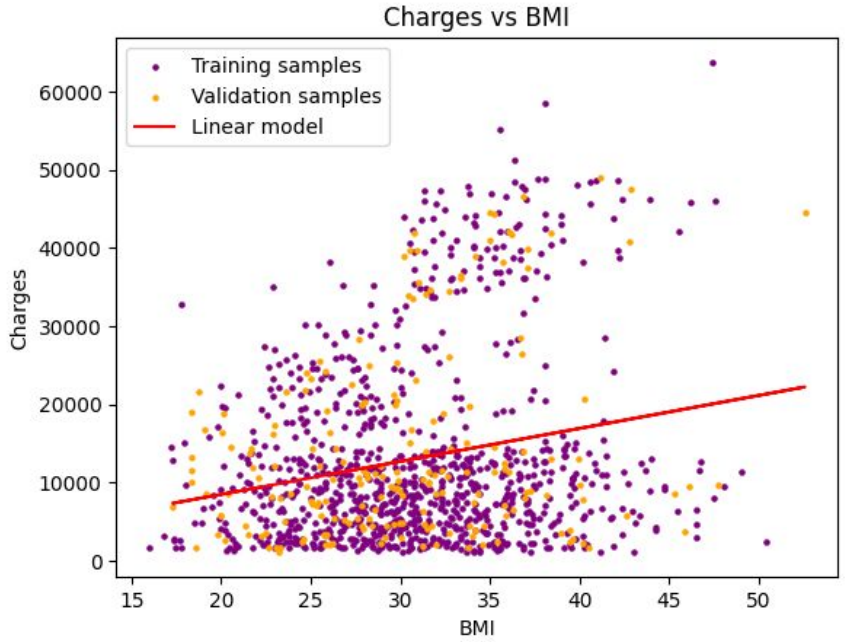
	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520
5	31	female	25.740	0	no	southeast	3756.62160
6	46	female	33.440	1	no	southeast	8240.58960
7	37	female	27.740	3	no	northwest	7281.50560
8	37	male	29.830	2	no	northeast	6406.41070
9	60	female	25.840	0	no	northwest	28923.13692

Dataset

- Medical Cost Personal Datasets:
  - age: age of primary beneficiary
  - sex: insurance contractor gender, female, male
  - bmi: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
  - children: Number of children covered by health insurance / Number of dependents
  - smoker: Smoking
  - region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
  - charges: Individual medical costs billed by health insurance.
- 6 features, 1 target

Single Feature Linear Regression (25%)

- Train linear regression model by gradient descent using only bmi feature.
- Tune the learning rate and epoch to get the same results as sklearn
- Reference parameters from sklearn:
  - Intercept: 1382
  - Weight: 380



Multiple Features Linear Regression (25%)

- Train linear regression model by gradient descent using all six features.
- Tune the learning rate and epoch to get the same results as sklearn
- Reference parameters from sklearn:
  - Intercepts: [-11857]
  - Weights: [[259]
    [-383]
    [333]
    [442]
    [24032]
    [-416]]

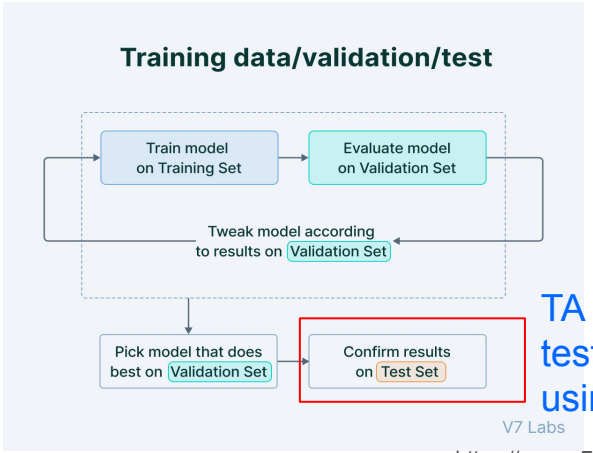
	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520
5	31	female	25.740	0	no	southeast	3756.62160
6	46	female	33.440	1	no	southeast	8240.58960
7	37	female	27.740	3	no	northwest	7281.50560
8	37	male	29.830	2	no	northeast	6406.41070
9	60	female	25.840	0	no	northwest	28923.13692

label encoding (see sample code)

Train your own model (20%)

- In this part, you can choose any learning rate, epoch, batch_size, and features you want to train your model to beat the baseline.
- Explain in detail how you choose the parameters in the report.
- Predict for the testing data and save the result into a csv file (refer to sample prediction.csv)

	age	sex	bmi	children	smoker	region	charges
0	33	male	30.250	0	no	southeast	NaN
1	19	female	32.490	0	yes	northwest	NaN
2	50	male	37.070	1	no	southeast	NaN
3	41	female	32.600	3	no	southwest	NaN
4	52	female	24.860	0	no	southeast	NaN
5	39	male	32.340	2	no	southeast	NaN
6	50	male	32.300	2	no	southwest	NaN
7	52	male	32.775	3	no	northwest	NaN
8	60	male	32.800	0	yes	southwest	NaN
9	20	female	31.920	0	no	northwest	NaN



TA will evaluate your testing performance using your csv file.

https://www.v7labs.com/blog/train-validation-test-set

Train your own model (20%)

$$\text{MSE} = \boxed{\text{Mean}\ \frac{1}{n}\sum_{i=1}^{n}\boxed{\text{Error}\ (Y_i - \hat{Y}_i)}^{\boxed{\text{Squared}\ 2}}}$$

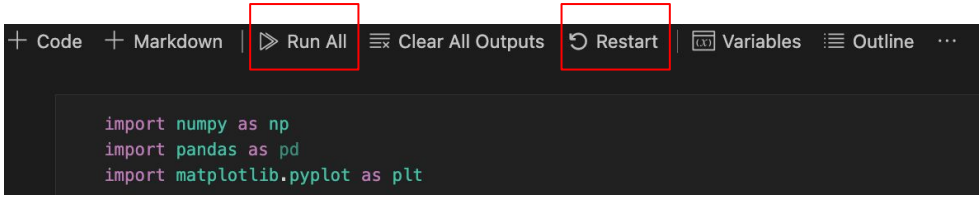
- Evaluation will be based on the testing mse loss.
- Testing data distribution is guaranteed to be similar with validation data.
- Explain your method in detail in report. Otherwise, extra penalty.

Points	Test MSE
20	< 30000000
15	< 40000000
10	< 50000000
5	50000000 ~ 100000000
0	> 100000000

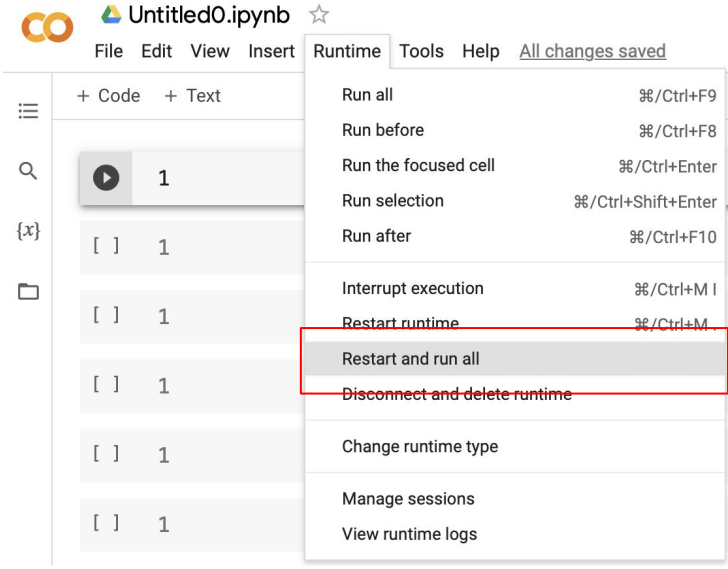
Data Analysis / Pre-processing

- Which features to be used?
- Different label encoding strategy?
- Feature Importance?
- Data distribution / normalization?
- Different feature weights?
- How to create more features?

Submission



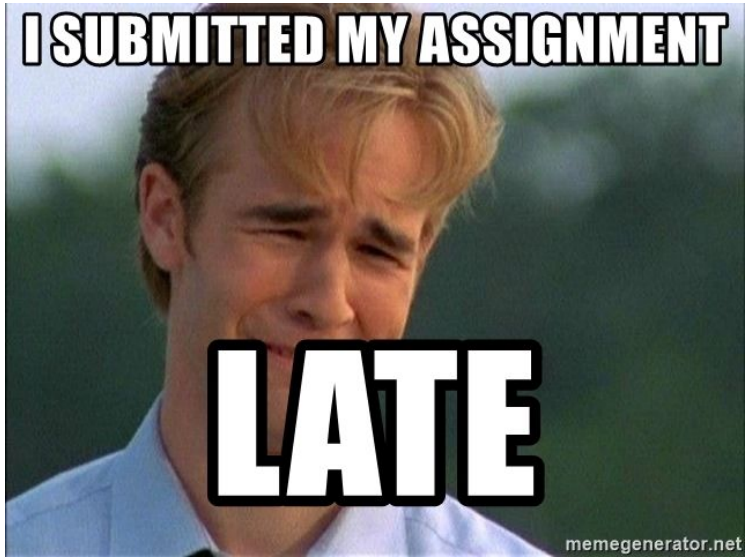
- Compress your .ipynb, .pdf, and .csv into zip file and submit on E3.
- Before submission:
  - Restart and Run All
  - Save and submit the .ipynb (keep all the cell outputs)
  - Get 0 points if you do not keep the cell outputs.
- <STUDENT ID>_HW1.zip
  - <STUDENT ID>_HW1.ipynb
  - <STUDENT ID>_HW1.pdf
  - <STUDENT ID>_prediction.csv




```
❯ zip -r 310551056_HW1.zip 310551056_HW1.ipynb 310551056_HW1.pdf 310551056_prediction.csv
  adding: 310551056_HW1.ipynb (deflated 34%)
  adding: 310551056_HW1.pdf (deflated 8%)
  adding: 310551056_prediction.csv (deflated 57%)
```

Late policy

- We will deduct a late penalty of 20 points per additional late day
- For example, If you get 90 points but delay for two days, your will get only 90-(20 x 2) = 50 points!



FAQ

- Why my loss is high and the training can not converge
  - Make sure you calculate the gradients correctly
  - Use smaller learning rate
- Can I use deep learning frameworks such as TensorFlow, PyTorch?
  - **No!** In HW1, you are request using only Numpy to implement linear regression and gradien descent. You can use matplotlib to plot the results.
- **DO NOT CHEAT!** Otherwise, you will get 0 points for the homework.
- If you have any questions, ask on E3 first! We will reply as soon as possible.

Have Fun!