Problem Set - 2

Please read all of the guidelines carefully before submitting the problem set. Each question is ***20 points*** and there are ***100 points*** in total.

Due date: Friday, February 4, 11:59 PM. Late submissions will be accepted with a penalty! (10% reduction per day – no submissions accepted two days after the deadline.)

Guidelines – Before You Start

1) **You should complete the problem set on your own.** Discussing ideas is fine; but, sharing answers and sharing code will be considered as plagiarism.
2) You will be using the **Python** programming language. You need to write your codes in an empty **.ipynb** file.
3) Make sure that you provide many comments to describe your code and the variables that you created.
4) Please use **LaTeX** or **MS Word** to submit your written responses (hand-written responses will not be graded).
5) For some of the coding exercises, you may need to do a little bit of "**Googling**" or review the documentation.

Deliverables:

1) The code of the problem set in **.ipynb** format (one file)
2) Short answers written with ***LaTeX*** or **MS Word** and exported in **.pdf** format (one file)

Questions

1) Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice? (Note: Please show your solution step-by-step by using what you know about marginal probability, conditional probability, joint probability, and the Bayes' theorem)

2) Suppose we have two NBA teams – for simplicity **team A** and **team B** – who have made it to NBA Playoffs. In each game between these two teams, **team A** has a winning probability of **0.55**, and **team B** has a winning probability of **0.45**. What is the probability that these two teams will play the 7th game in NBA Playoffs? (Notes: There cannot be a tie in any game (i). Please check this link for more information about NBA Playoffs: https://en.wikipedia.org/wiki/NBA_playoffs and to think about possible combinations (ii). Also, please show your solution step-by-step by using what you know about marginal probability, conditional probability, joint probability, and the Bayes' theorem (iii)).

3) From scratch (not using any pre-packaged tools for direct calculation), implement the ***gradient descent algorithm*** for linear regression and test your results on the California Housing Dataset:

https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_california_housing.html#sklearn.datasets.fetch_california_housing

Here is what you need to do step by step:

a. Implement the ***gradient descent algorithm*** from scratch
b. Choose the following features from the dataset as your ***X*** matrix: ***MedInc, HouseAge, AveRooms, AveBedrms, Population, AveOccup, Latitude, Longitude***
c. Choose the following feature from the dataset as your ***Y*** matrix: ***MedHouseVal***
d. Randomly split your data into training (70% of total) and test sets (30% of total) by using sklearn's **`train_test_split`** function. Set **`random_state = 265`**: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html.
e. Set the `number_of_steps = 1000` and `learning_rate = 0.01`.
f. By running your code, determine the best set of parameters (=weights) for the constant and your features listed in ***b***). Your cost function will be **MSE** (=you should pick the set of parameters that give you the lowest **MSE**).
g. Report and interpret the results. What are the factors that explain the house prices the most?

4) Now, try using a pre-packaged tool and comparing the results. Do the following:

a. Use **`SGDRegressor`** provided by **`scikit`**: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDRegressor.html
b. Step b), c), and d) are the same as in Question 3.
c. Set the `max_iter = 1000`, `alpha = 0.01`, `random_state = 265`, and `loss = 'squared_error'`. Other parameters should be set to 'default'.
d. By running your code, determine the best set of parameters (=weights) for the constant and your features listed in b).
e. Report and interpret the results. What are the factors that explain the house prices the most? **Are the results different from the previous question? If different, explain why the results might be different.**

5) Finally, write a function from scratch that computes a ***variance-covariance matrix*** by transforming the following formula into code:
Variance-covariance matrix: $cov(\boldsymbol{X}) = \mathrm{E}[(\mathbf{X} - \mathrm{E}[\mathbf{X}])(\boldsymbol{X} - \mathrm{E}[\boldsymbol{X}])^T]$

Your function/code should work for matrices of any size. Test that your function is running (=successfully computing the variances and covariances of the variables and variable pairs in the dataset) by using the California Housing Dataset that you have used in previous questions.