STA 243: Homework 2

- Homework due in Canvas: 05/17/2023 at 11:59PM. Please follow the instructions provided in Canvas about homeworks, carefully.

1. (**5 Points**) Prove that a differentiable function $f(\theta) : \mathbb{R}^d \to \mathbb{R}$ is convex if and only if

$$f(\theta_2) \geq f(\theta_1) + \nabla f(\theta_1)^\top (\theta_2 - \theta_1)$$

Hint: Think of 1-dimensional case and extend the intuition to d-dimensional case.

2. (**20 Points**) The origin of the dataset `housingprice.csv` we will use in this question is from the Coursera open course Machine Learning Foundations: A Case Study Approach by Prof. Carlos Guestrin and Prof. Emily Fox. Load the training data `train.data.csv` and testing data `test.data.csv`. We'll build our regression model on the training data and evaluate the model on the testing data.

   (a) Build a linear model (you are free to use any Python package for this) on the training data by regressing the housing price on these variables: `bedrooms`, `bathrooms`, `sqft_living`, and `sqft_lot`. What's the R^2 of the model on training data? What's the R^2 on testing data?

   (b) The image below is Bill Gates' house. Load the file `fancyhouse.csv` to obtain the features of the house. Guess the price of his house using your linear model. Do you think the predicted price is reasonable?

Figure 1: Image from Wikipedia Commons

   (c) Let's continue to improve the linear model we have. Instead of throwing only the raw data into the statistical model, we might want to use our intuition and domain expertise to extract more meaningful features from the raw data. This step is called feature engineering. Using meaningful features in the model is often crucial for successful data analysis. Add another variable by multiplying the number of bedrooms by the number of bathrooms, which describes the combined benefit of having more bedrooms and bathrooms. Add this variable to the linear model we have in Part (a). What's the R^2 of the new model on the training data and testing data respectively?

   (d) Perform parts (a), (b) and (c) above without using any in-built function in Python (i.e., any packages that are related directly to linear regression), but by using `gradient descent algorithm` on the sample-based least-squares objective function, to estimate the OLS regression parameter vector. How does your result compare to the result from previous part ? Note that you have to set the step-size parameter appropriately for this method.

(e) Perform arts (a), (b) and (c) above now using `stochastic gradient descent` (with one sample in each iteration). How does your result compare to the result from previous parts ? Note: while running `stochastic gradient descent`, you can sample without replacement and when you run out of samples, just start over. Note that you have to set the step-size parameter appropriately for this method.

3. (**15 Points**) Prove the Fact in Page 91 of `Opt.pdf` and solve the recursion to Page 92 to obtain the final result of the Theorem in Page 86. (**Hint:** You can use induction)

4. (**10 Points**) In class, we defined the notion of a sub-gradient and worked out the example of the absolute function ($f(\theta) = |\theta|$).

   Consider the function from $\mathbb{R}^d \to \mathbb{R}$ which is the Euclidean norm for a vector, i.e., $f(\theta) := \|\theta\|_2$. Compute the sub-gradient of this function.

Pledge:

Please sign below (print full name) after checking (✓) the following. If you cannot honestly check each of these responses, please email me at kbala@ucdavis.edu to explain your situation.

- I pledge that I am a honest student with academic integrity and I have not cheated on this homework.
- These answers are my own work.
- I did not give any other students assistance on this homework (beyond what is allowed as per syllabus).
- I understand that to submit work that is not my own and pretend that it is mine is a violation of the UC Davis code of conduct and will be reported to Student Judicial Affairs.
- I understand that suspected misconduct on this homework will be reported to the Office of Student Support and Judicial Affairs and, if established, will result in disciplinary sanctions up through Dismissal from the University and a grade penalty up to a grade of "F" for the course.

Signature