CS5785 Homework 1

The homework is generally split into programming exercises and written exercises.

This homework is due on **September 16, 2021 at 11:59 PM EST**. Upload your homework to Gradescope. There are two assignments for this homework in Gradescope. Please note a complete submission should include:

1. A write-up as a single `.pdf` file, which should be submitted to "Homework 1 (write-up)" This file should contain your answers to the written questions **and** exported pdf file / structured write-up of your answers to the coding questions (which should include core codes, plots, outputs, and any comments / explanations).
2. Source code for all of your experiments (AND figures) zipped into a single .zip file, in `.py` files if you use Python or `.ipynb` files if you use the IPython Notebook. If you use some other language, include all build scripts necessary to build and run your project along with instructions on how to compile and run your code. **If you use the IPython Notebook to create any graphs, please make sure you also include them in your write-up.** This should be submitted to "Homework 1 (code)".

The write-up should contain a general summary of what you did, how well your solution works, any insights you found, etc. On the cover page, include the class name, homework number, and team member names. You are responsible for submitting clear, organized answers to the questions. You could use online LaTeX templates from Overleaf, under "Homework Assignment" and and "Project / Lab Report".

Please include all relevant information for a question, including text response, equations, figures, graphs, output, etc. If you include graphs, be sure to include the source code that generated them. Please pay attention to Slack for relevant information regarding updates, tips, and policy changes. You are encouraged (but not required) to work in groups of 2.

IF YOU NEED HELP

There are several strategies available to you.

- If you get stuck, we encourage you to post a question on the Discussions section of Canvas. That way, your questions/solutions will be available to other students in the class.
- Your instructor and TAs will offer office hours, which are a great way to get some one-on-one help.
- You are allowed to use well known libraries such as `scikit-learn`, `scikit-image`, `numpy`, `scipy`, etc. in this assignment. Any reference or copy of public code repositories should be properly cited in your submission (examples include Github, Wikipedia, Blogs).

PROGRAMMING EXERCISES

Please use different `.py` or `.ipynb` files for different parts

Part I. The Housing Prices

1. Join the House Prices - Advanced Regression Techniques competition on Kaggle. Download the training and test data.
2. Give 3 examples of continuous and categorical features in the dataset; choose one feature of each type and plot the histogram to illustrate the distribution.
3. Pre-process your data, explain your pre-processing steps and the reasons why you need them. (Hint: data pre-processing steps can include but are not restricted to: dealing with missing values, normalizing numerical values, dealing with categorical values etc.)
4. During data pre-processing, one common method for dealing with categorical features is to use one-hot encoding (OHE). Give some examples of features that you think should use OHE and explain why. User OHE to encode these features with scikit-learn and visualize the results.
5. Using ordinary least squares (OLS), try to predict house prices on this dataset. Choose the features (or combinations of features) you would like to use or ignore, provided you justify your choice. Evaluate your predictions on the training set using MSE and R^2 score. For this question, you need to implement OLS from scratch without using any external libraries or packages.
6. Train your classifier using all of the training data, and test it using the testing data. Submit your results to Kaggle.

Part II. The Titanic Disaster

1. Join the Titanic: Machine Learning From Disaster competition on Kaggle. Download and pre-process the data.
2. Using logistic regression, try to predict whether a passenger survived the disaster. Choose the features (or combinations of features) you would like to use or ignore, provided you justify your choice.
3. Train your classifier using all of the training data, and test it using the testing data. Submit your results to Kaggle.

WRITTEN EXERCISES

1. Maximum Likelihood and KL Divergence. Let $\hat{p}(x, y)$ denote the empirical data distribution over a space of inputs $x \in \mathcal{X}$ and outputs $y \in \mathcal{Y}$. For example, in an image recognition task, x can be an image and y can be whether the image contains a cat or not, and $\hat{p}(x, y)$ denotes how these data pairs are distributed. Let $p_\theta(y|x)$ be a probabilistic classifier parameterized by θ, e.g., a logistic regression classifier with coefficients θ. Show that the following equivalence holds:

$$\arg\max_\theta \mathbb{E}_{\hat{p}(x,y)} \left[\log p_\theta(y|x)\right] = \arg\min_\theta \mathbb{E}_{\hat{p}(x)} \left[KL(\hat{p}(y|x)||p_\theta(y|x))\right]$$

where KL denotes the KL-divergence, which is a common metric for measuring the similarity between two distributions:

$$KL(p(x)||q(x)) = \mathbb{E}_{p(x)} \left[\log p(x) - \log q(x)\right]$$

Explain in words what this equation suggests for training a logistic regression classifier.

2. Gradient and log-likelihood for logistic regression.

   (a) Let $\sigma(a) = \dfrac{1}{1+e^{-a}}$ be the sigmoid function. Show that $\dfrac{d\sigma(a)}{da} = \sigma(a)\,(1-\sigma(a))$.

   (b) Using the previous result and the chain rule of calculus, derive the expression for the gradient of the log likelihood:

   $$\nabla L(\theta) = [\sigma(\theta^T \mathbf{x}) - y]\mathbf{x}$$

   where

   $$L(\theta) = -\left[y \log \sigma(\theta^T \mathbf{x}) + (1-y)\log(1-\sigma(\theta^T \mathbf{x}))\right]$$

3. Linear regression and OLS. For 7 dots $(-1,-1), (-1,0), (0,-1), (0,0), (0,1), (1,0), (1,1)$,

   (a) Plot the dots in a graph, without computing, what do you think is the result (slope and intercept) of running linear regression by minimizing MSE?

   (b) Calculate the true results, show by running code or mathematical derivation that this answer minimizes MSE. If it is different from you answer to (a), explain why you think this happened.

   (c) What about minimizing the absolute error? Justify your answer with a plot or math derivation.