CS5785 Homework 1

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

This homework is due on **September 13, 2022 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 Canvas 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. One common method of pre-processing categorical features is to use a one-hot encoding (OHE).

   Suppose that we start with a categorical feature x_j, taking three possible values: $x_j \in \{R, G, B\}$. A one-hot encoding of this feature replaces x_j with three new features: x_{jR}, x_{jG}, x_{jB}. Each feature contains a binary value of 0 or 1, depending on the value taken by x_j. For example, if $x_j = G$, then $x_{jG} = 1$ and $x_{jR} = x_{jB} = 0$.

   Give some examples of features that you think should use a one-hot encoding and explain why. Convert at least one feature to a one-hot encoding (you can use your own implementation, or that in pandas or scikit-learn) and visualize the results by plotting feature histograms of the original feature and its new one-hot encoding.
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 the MSE and the R^2 score. For this question, you need to implement OLS from scratch without using any external libraries or packages.
6. Train your model using all of the training data (all data points, but not necessarily all the features), 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. Implement logistic regression (it's ok to use sklearn or similar software packages), try to predict whether a passenger survived the disaster with your model. 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. In machine learning, we often need to assess the similarity between pairs of distributions. This is often done using the Kullback-Leibler (KL) divergence:

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

The KL divergence is always non-negative, and equals zero when p and q are identical. This makes it a natural tool for comparing distributions.

This question explores connections between the KL divergence and maximum likelihood learning. Suppose we want to learn a supervised probabilistic model $p_\theta(y|x)$ (e.g., logistic regression) with parameters θ over a dataset $\mathcal{D} = \{(x^{(i)}, y^{(i)}) \mid i = 1, 2, ..., n\}$. Let $\hat{p}(x, y)$ denote the *empirical* distribution of the data, which is the distribution that assigns a probability of $1/n$ to each of the data points in $\mathcal{D}$ (and zero to all the other possible (x, y)):

$$\hat{p}(x, y) = \begin{cases} \frac{1}{n} & \text{if } (x, y) \in \mathcal{D} \\ 0 & \text{otherwise.} \end{cases}$$

The empirical distribution can be seen as a guess of the true data distribution from which the dataset $\mathcal{D}$ was sampled i.i.d.: it assigns a uniform probability to every possible training instance seen so far, and does not assign any probability to unseen training instances.

Prove that selecting parameters θ by maximizing the likelihood is equivalent to selecting θ that minimize the average KL divergence between the data distribution and the model distribution:

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

Here, $\mathbb{E}_{p(x)} f(x)$ denotes $\sum_{x \in \mathcal{X}} f(x)p(x)$ if x is discrete and $\int_{x \in \mathcal{X}} f(x)p(x)dx$ if x is continuous.

2. Gradient and log-likelihood for logistic regression.

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

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

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

   where

$$\ell(\theta) = y \log \sigma(\theta^T \mathbf{x}) + (1 - y) \log(1 - \sigma(\theta^T \mathbf{x}))$$

3. Analytical solution of the Ordinary Least Squares Estimation. Consider we have a simple dataset of n labeled data $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, where data $x_i \in \mathbb{R}$ and $y_i \in \mathbb{R}$ is its corresponding label. We use a simple estimated regression function of:

$$\widehat{y}_i = \theta_0 + \theta_1 x_i$$

Instead of gradient descent which works in an iterative manner, we try to directly solve this problem. We define the cost function as the residual sum of squares, parameterized by θ_0, θ_1:

$$J(\theta_0, \theta_1) = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$$

(a) Calculate the partial derivative of $\frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1)$ and $\frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1)$.

(b) Consider the fact that $J(\theta_0, \theta_1)$ has an unique optimum, we can actually get the analytical solution of θ_0, θ_1 by the following normal equations:

$$\frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1) = 0$$
$$\frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1) = 0$$

prove the following proprieties that

$$\theta_0 = \bar{y} - \theta_1 \bar{x}$$

and

$$\theta_1 = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\sum_{i=1}^{n} x_i (x_i - \bar{x})}$$

(Note: $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$.)

(c) Calculate the sum of the residuals $\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - (\theta_0 + \theta_1 x_i))$. What can you learn from the value of $\sum_{i=1}^{n} e_i$?