MIT 6.C01/6.C51 Modeling with Machine Learning
Spring 2022
Problem Set 1

Release date. **Friday, February 04 2022, 5:00 PM ET**
Due date. **Friday, February 18 2022, 5:00 PM ET**

Instructions

- This problem set contains three questions, each containing sub-questions.
- Each of these three questions has one sub-question marked as **Graduate version.**This question is mandatory for those who have registered for the graduate version of this class (`6.C51`). It is optional for others.
- Submissions should be made via Gradescope. You are required to prepare one PDF containing all your answers–both your written answers and code snippets. Please answer each question in a separate page, since Gradescope will require you to select pages pertaining to each sub-question.
- You have been provided `ps1_p1.py`, a Python file containing starter code for Problem 1. Use it to write and test your implementations of the different functions we ask you to provide code for. When submitting your solutions, paste code that the question asks you into the final PDF you will submit.
- Please adhere to the collaboration policy described on our course page on Canvas. Clearly mention the names of your collaborators on the top of your first page in your submission.

Problem 1: Data Featurization

You have been asked to help a local car rental company build machine learning models. For each of the following types of data, how should the company featurize it? In other words, define $\phi(x)$ so that the company best uses this information in creating linear models to predict the desired property.

1. The company wants to analyze if a customer's age and their feedback can predict how likely they will recommend the company to their friends. For each customer, you know their age. Based on routine surveys, the company also maintains feedback provided by a customer as {"happy", "satisfied", "unhappy"}. How should the company encode these two pieces of information? Answer this by providing a Python script (fill in `phi_1(age, feedback)` in `ps1_p1.py`) which converts a customer's age (an integer) and feedback information (a string) into a feature vector which will help in predicting referrals.

2. The company wants to predict the *driver safety* of their customers. They have access to telemetry information (speed, GPS, time information) from devices placed in each of their cars, which logs driving information for each of its customers every five minutes when the car is in motion. For each customer, you have access to telemetry information for the last two years, which is logged in the following format, one line per entry:

   `date | speed | location_id`

   APIs to two other tables in their database provide you the following information—

   - `get_speed_limit(location_id)`
   - `get_location_name(location_id)`

   (a) How should the company encode this information for a given customer? Please describe at least 4 features which may be valuable for the company when evaluating driver safety. Briefly explain (1-2 sentences) why these features are useful.

   (b) Provide a Python script (fill in `phi_2(logs)` in `ps1_p1.py`) to implement your suggestions in question 1.2.a, and provide a single feature vector which will help predict the safety of one customer. You can use the APIs described in the problem statement to prepare your feature vector.

3. **Graduate version.**

   (a) Consider *linear transformations* of the independent variables x used in linear regression. Let z be a linear transformation of x, where $z = px + q\mathbf{1}$, for $x, z \in \mathbb{R}^d$ and $p, q \in \mathbb{R}$.

   You train a linear regression model on x

   $$y = \theta^\intercal x + \theta_0$$

   and you train another linear regression model on z

   $$y' = \theta'^\intercal z + \theta'_0$$

   Will the learned parameters θ', θ'_0 be linear transformations of θ, θ_0? Provide an expression relating the two sets of parameters.

(b) Let x be the monthly expense of a car in the rental company's fleet. For a particular car, you notice that the expense for a given month is roughly 1.2 times the previous month's expense. You want to train a model where you use the monthly expenses of the car, over a period of five years, as the independent variable in linear regression. You want to predict the resale value of the car. What is likely a good choice for a transformation ϕ on x?

(c) Is your strategy to model $\phi(x)$ instead of x, for the choice of ϕ you made above, reasonable in linear regression? Should you expect it to be a good model? Justify either way.

Problem 2: Hyperparameters

In linear regression problems, the choice of the featurizing function ϕ and regularization parameter λ affects the resulting linear regression coefficients θ given some dataset. To determine the optimal hyperparameters, let's assume that we have a dataset with samples X and targets y randomly divided into 5 equal size non-overlapping parts: $X = [X_0, ..., X_4]$ and $y = [y_0, ..., y_4]$. In this problem, you will devise a procedure to select the optimal ϕ^* and λ^* and evaluate the performance of the resulting linear regression model. You may find the following functions helpful for implementing your procedure:

- `fit`(X, y, ϕ, λ) fits a linear regression model using a subset of the samples X and targets y, parameters ϕ and λ, and returns the learned θ.
- `predict`(X, ϕ, θ) returns the predictions $\hat{y}$ for a given model θ.
- `mse`$(y, \hat{y})$ returns the mean squared error between the target subset y and predicted target subset $\hat{y}$.

You have seven different feature mappings $\phi^{(1)}; \ldots; \phi^{(7)}$ (possibly of different dimensions) and four possible regularization parameters $\lambda_1; \ldots; \lambda_4$ to consider. Your task is to (1) find the best combination of hyperparameter and (2) evaluate the resulting method so that the end user gets a fair estimate of the performance of your method on new test cases (not included). Please explicitly identify which data is used for which purpose.

1. Write a procedure to select the best combination of ϕ and λ hyperparameters.
2. Provide an expression for the parameters, θ, expressed only using the functions above, that you would return for your best estimator.
3. Provide an expression that gives an unbiased estimate of the test error for your resulting regression function.
4. **Graduate version.**

(a) You can train other models on X by varying feature functions, training schedules, and other hyperparameters. How can you use these other models to learn more of the confidence you can attribute to the predictions made by model θ^* when released in "production". In production, you have access to X_{new}–new data (without their corresponding targets) that your models have not seen before.

(b) When working with real-world datasets with many features, the number of $\phi >> 7$. There are many ways to featurize the inputs, and different combinations of featurization functions produce varying results. Generally, evaluating every combination of hyperparameters may be computationally infeasible since the number of combinations grows exponentially with the number of hyperparameters. Please suggest one possible computationally feasible approach to explore reasonable hyperparameters. Justify your proposed procedure and describe any potential drawbacks.

Problem 3: Linear Regression

1. In this question, we will consider the linear regression (with regularization) loss function, $J_{n,\lambda}(\theta) = R_n(\theta) + \frac{\lambda}{2}\sum_i^d \theta_i^2$. In class, we considered **stochastic gradient descent (SGD)** as the optimization algorithm of choice to minimize the loss function on the training set. Alternatively, we could have solved the function exactly (as in normal equations for ordinary least squares) or used gradient descent (GD). Explain briefly why SGD is an appropriate algorithm in this setting?

2. We saw in class that the overall objective function on adding regularization to the mean squared error loss is as follows:

$$J_{n,\lambda}(\theta) = \frac{1}{2n}\sum_{t=1}^{n}(y^{(t)} - \theta^\intercal x^{(t)})^2 + \frac{\lambda}{2}\sum_i^d \theta_i^2.$$

Note that $\theta^\intercal x^{(t)}$ is the dot product of θ and $x^{(t)}$, which are both in $\mathbb{R}^d$. First, derive the gradient of the the total loss, $\nabla_\theta J_{n,\lambda}$, with respect to the parameters. Then show that the closed-form solution for the updated parameter θ^{k+1} in stochastic gradient descent over the revised objective function $J_{n,\lambda}(\theta)$ is

$$\theta^{k+1} = \theta^k(1 - \eta_k\lambda) + \eta_k(y^{(t)} - \theta^{k\intercal} x^{(t)})x^{(t)}$$

3. Consider the following four variations of simple 1-dimensional linear regression prob-

lems:

$$\text{i. } \min_{\theta_1} \ \frac{\lambda}{2}\theta_1^2 + \sum_{i=1}^{n}(y^{(i)} - \theta_1 x^{(i)})^2$$

$$\text{ii. } \min_{\theta_0,\theta_1} \ \frac{\lambda}{2}\theta_1^2 + \sum_{i=1}^{n}(y^{(i)} - \theta_1 x^{(i)} - \theta_0)^2$$

$$\text{iii. } \min_{\theta_0,\theta_1} \ \frac{\lambda}{2}\theta_0^2 + \sum_{i=1}^{n}(y^{(i)} - \theta_1 x^{(i)} - \theta_0)^2$$

$$\text{iv. } \min_{\theta} \ \frac{\lambda}{2}\|\theta\|^2 + \sum_{i=1}^{n}(y^{(i)} - \theta^\top \phi(x^{(i)}))^2, \phi(x) = [x, 1]^\top$$

Assign each method to at least one plot in Figure 1. Base the assignment on whether the method could have plausibly produced it with some value of $\lambda \geq 0$. Each method may be assigned to more than one figure, and each figure may correspond to more than one method. Briefly explain your answers.

4. **Graduate version.**

   (a) We looked at the mean squared loss function in class, and discussed how it handles outliers poorly. Can we do better?

   Given a random variable y, and an estimate of the variable $\hat{y}$, the mean squared error as a function of $\hat{y}$ can be expressed as:

   $$\text{mse}(\hat{y}) = \mathbb{E}[(y - \hat{y})^2].$$

   Show that the minimum of the mean squared error is attained by setting $\hat{y} = \mathbb{E}[y]$.

   (b) Instead of the MSE, consider the mean absolute error function (MAE) for the setting described above:

   $$\text{mae}(\hat{y}) = \mathbb{E}[|y - \hat{y}|].$$

   Now show that the minimum of the MAE function is attained by setting $\hat{y} = \text{median}(y)$.

   From this exercise, we learn that the optimal estimate of the MSE is in fact the average of y. Concretely, for any training data (X, y), using the MSE loss function will optimize the estimator/model to predict the average of y. If the dataset has outliers, then the predictions that the model learns to make are consequently affected.

   On the other hand, using the MAE loss will ensure that the model optimizes to learn the median of y, a quantity that is less affected by outliers.

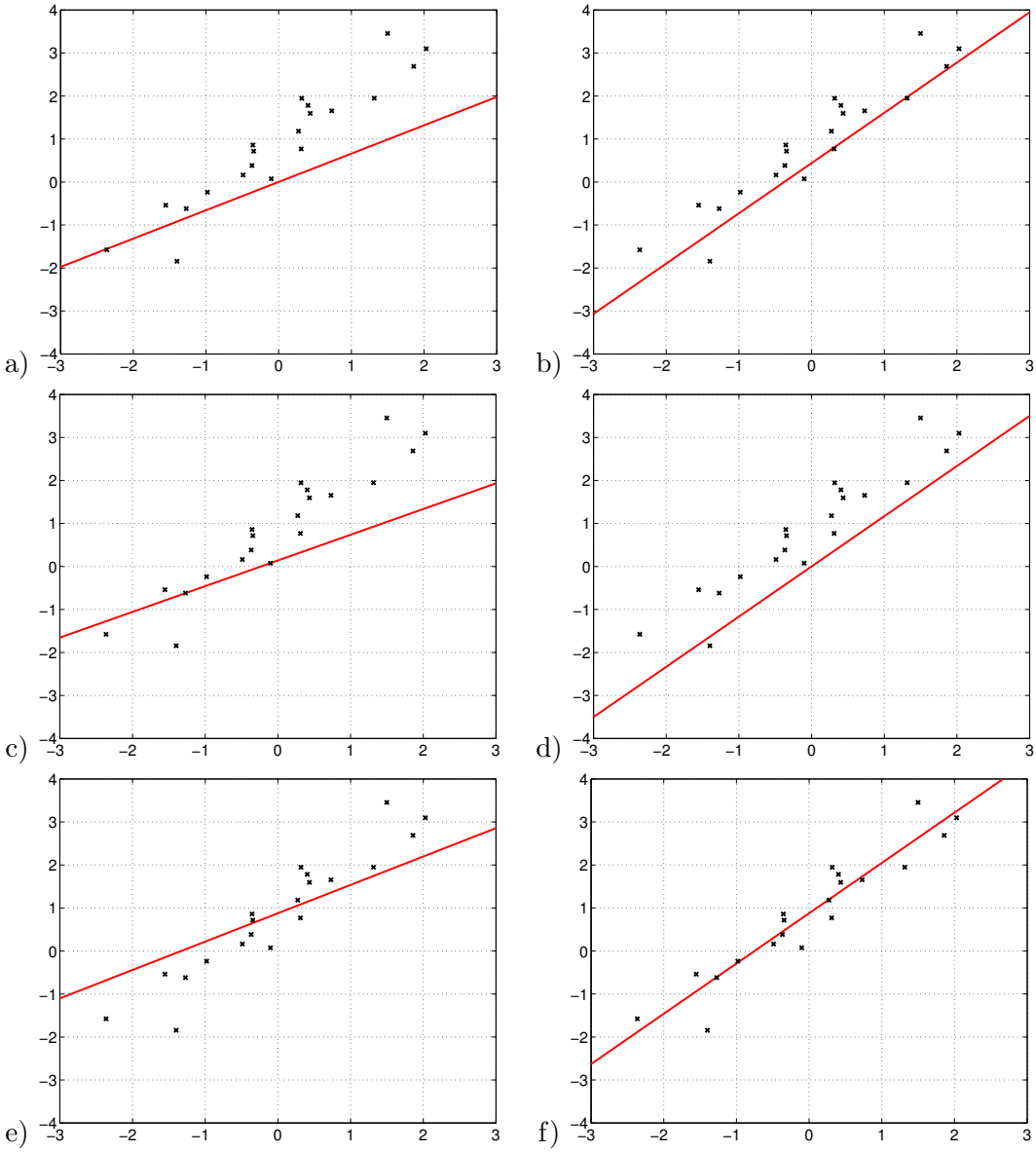

Figure 1: One dimensional regression results