95-828 Machine Learning for Problem Solving: Homework 1

Heinz College
Carnegie Mellon University

- There are 6 questions on this assignment: 4 conceptual (short answers) and 2 applied (implementation) questions that involve coding. Assignemt write-up should be submitted as **.pdf** using *Gradescope*. Submissions can be hand-written, but should be labeled and clearly legible—if the graders cannot read your solutions, you will not be given credit for them. Submissions can also be written in Word or LaTeX. For the applied questions, export your Jupyter notebook as **.pdf** and upload the exported **.pdf** to *Gradescope*. Upon submission, label each question using the template provided by Gradescope. We will share a separate file with instructions on how to use *Gradescope*.
- The assignment is due at 9:00 AM (before class) on **Feb 12, 2019**.
- If you have any questions, please use *Piazza* and visit the course staff during office hours.
- You may *discuss* the questions with fellow students, *however* you must always write your own solutions and must acknowledge whom you discussed the question with. Do not copy from other sources, share your work with others, or search for solutions on the web. Plagiarism will be penalized according to the university rules.

1 Conceptual: Learning Tasks [10 points]

Given each of the tasks below, specify which kind of learning would be involved (supervised or unsupervised). Further specify what kind of a supervised task (classification, regression) or unsupervised task (clustering, density estimation, dimensionality reduction) would be involved. Give a short description (1-2 sentences) of what features and (if required) what labels you might use for the corresponding ML task.

Note that some tasks may belong to more than one type; it is sufficient to provide justification (1-2 sentences) for one of the types.

a. (2 pts) Identifying underlying topics from a large collection of blogs

b. (2 pts) A credit card company determining the credit limit for a customer

c. (2 pts) Estimating if a student will pass or fail a course

d. (2 pts) Segmenting grocery store customers by purchase behavior

e. (2 pts) Estimating the probability of word 'lottery' in spam emails

2 Conceptual: Data Preparation [15 points]

a. (6 pts) One of the crucial steps in Data Science is to prepare the input data such that the data is amenable to be consumed by machine learning algorithms. Given data preparation concepts in column **A** of Table 1 below, match related concepts from column **B** against each concept in column **A** (Note: the matching must be one-to-one).

b. (6 pts) Give an example (other than the ones shown in the lecture slides) for each of the following types of attributes: (*i*) nominal, (*ii*) ordinal, and (*iii*) numerical. Which of the following quantities

A	**B**
histogram	feature skewness
one-hot encoding	outliers
radar plot	values in $[0, 1]$
box-cox transform	categorical feature
min-max	binning
robust z-score	$> 2d$ data

Table 1: Q.2.a

	(1) frequency distribution	(2) median	(3) mean
(i) nominal			
(ii) ordinal			
(iii) numerical			

Table 2: Q.2.b

would it be suitable to calculate for these types of attributes: (1) frequency distribution, (2) median, and (3) mean? Mark the suitable ones with ✗ in Table 2.

c. (3 pts) You are given a dataset with 2 attributes. Attribute `school` can take one of 5 categorical values {CMU, Stanford, MIT, UCB, Yale} and attribute `major` takes values from {Engineering, Policy}. Specify the one-hot-encoding (OHE) of the values and use the encoding to transform the following input data set with 6 observations (students). Provide the encoded/transformed data. How many features (i.e., columns) does the new data matrix contain?

$$\begin{bmatrix} \text{Yale} & \text{Policy} \\ \text{Stanford} & \text{Policy} \\ \text{Stanford} & \text{Engineering} \\ \text{CMU} & \text{Policy} \\ \text{MIT} & \text{Engineering} \\ \text{CMU} & \text{Engineering} \end{bmatrix}$$

3 Conceptual: Linear Regression [20 points]

a. (5 pts) You are given the `sales` prediction task for a portfolio of products A, B, C and D. You are given the below plots (Fig. 1) exhibiting relation between each product's `revenue` (y) and `utility` (some feature x) of the product.

   i. (1 pts) Which of the relationships in Fig. 1 is described by $y = w_0 + w_1 x$? Mark your answer on the figure.

   ii. (4 pts) Given a dataset $D = (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, provide the formulas for estimating the parameters w_0 (intercept) and w_1 (slope) for a simple linear regression model $y = w_0 + w_1 x$ using least squares as the loss function (without regularization).

   [*Hint*: Take derivative of sum of squared residuals ($\sum_i [y_i - (w_1 x_i + w_0)]^2$) first w.r.t. w_0, and then w.r.t. w_1. You'll get 2 equations. Solve for w_0 and w_1 by equating derivatives to 0. You'll get two closed-form solutions, for w_0 and w_1, respectively.]

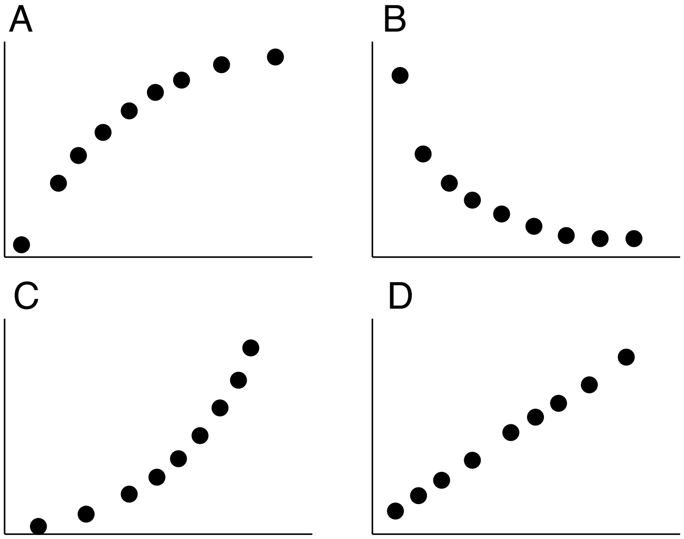

Figure 1: Product revenue vs utility for each of the products A, B, C and D

b. (5 pts) You are given a task of predicting the sales of a company. In the dataset you find that there are 16,000 features that are measured, and there are 2000 sales instances. The domain expert in the company has suggested that the sales and features should be linearly related. Given this dataset, answer the following:

i. (3 pts) Should you employ linear regression (without regularization) for this task? What potential problem would you encounter for employing linear regression on this data? Name the issue and explain briefly.

ii. (2 pts) Could you suggest an approach to address this potential problem? Justify your approach in 1-2 sentences.

c. (1 pts) Suppose you chose linear regression model for the sales prediction task. You observe that for the given sales dataset, the coefficient w_j of one particular feature j has a relatively very large negative value. What would be your assessment of this feature? (Select one or more of the following statements that apply.):

(i) This feature should be kept in the final model since it has a strong effect on the model.

(ii) This feature should be dropped from the model since a large negative value will not have a strong effect on the model.

(iii) It is not possible to determine the importance of this feature unless we know the other coefficients.

d. (2 pts) Consider the relation between `revenue` and `utility` for a product. The domain expert tells you that the relation is quadratic. Can you extend linear regression model to capture the quadratic relation? If yes, write down the model equation and explain briefly. If no, explain briefly why this should be the case.

e. (5 pts) A data engineer in the sales department recorded a feature 3 times under different names. Given the d dimensional sales data including these identical copies of that feature, answer the following:

i. (3 pts) What potential problem would we face when applying linear regression on this data? Name the issue and explain briefly.

ii. (2 pts) Is there a systematic approach to address the above problem? Name the approach and explain in 1-2 sentences.

f. (2 pts) Suppose the sales data has three features; namely `price`, `brand_perception` and `region`. The features are related to the outcome variable `revenue` as shown in Fig. 2. Is it possible to employ a linear regression for the `revenue` prediction task using this dataset? If yes, explain briefly how you would use linear regression in this case. If no, explain why linear regression would not be a suitable choice to model this data.

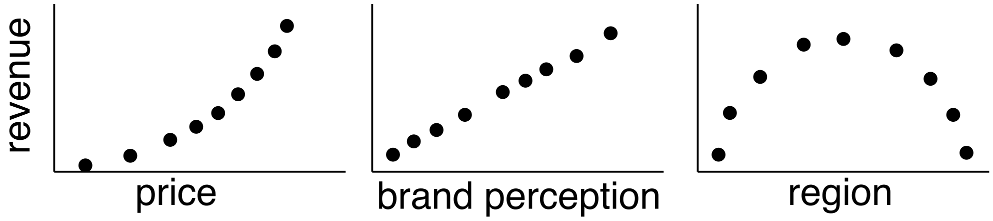

Figure 2: Relation of outcome `revenue` vs features {`price`, `brand_perception`, and `region`}

4 Conceptual: Model Selection and Cross Validation [20 points]

a. (6 pts) Suppose you have regression data generated by a polynomial of degree 3. Characterize the bias and variance of the estimates of the following models on the data with respect to the true model by circling the appropriate entry.

	Bias	Variance
Linear regression	low / high	low / high
Polynomial regression with degree 3	low / high	low / high
Polynomial regression with degree 10	low / high	low / high

b. (8 pts) Cross Validation (CV) is a common tool for model selection and a 4-fold CV is shown in the following Figure 3 for illustration purpose only.

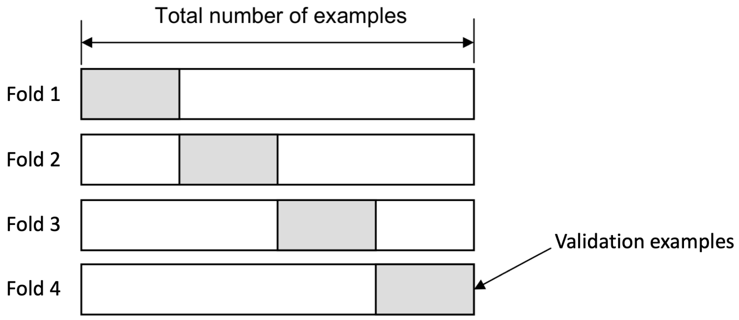

Figure 3: Illustration of data splits for 4-fold Cross Validation

i. (6 pts) Given a fixed model (for a specific choice of the hyperparameter(s)) for which we aim to compute cross-validation error, and a large number of folds (i.e., large k), what can you say about the approximation and the estimation error of the model trained during each fold as well as the overall computational time for cross-validation? (compared to smaller k) Mark with ✗ below, and provide explanation in 1-3 sentences for each property that you marked. (See *Hint* below.)

Approximation error	**Estimation error**	**Computational time**
( ) Smaller	( ) Smaller	( ) Lower
( ) Larger	( ) Larger	( ) Higher
( ) Stays the same	( ) Stays the same	( ) Stays the same

[*Hint*: Often we think of *changing the model complexity* over a *fixed* dataset. This time, we are thinking of *changing the size of training data* when the model complexity is *fixed*. Think of whether a fixed-complexity model would tend to overfit more easily on small or large data.]

ii. (2 pts) Conversely, with a small number of folds, what can you say about the above three characteristics?

c. (6 pts) Suppose you are building a machine learning model for predicting the price tag of a phone based on the phone features: `display_type`, `RAM_size`, `camera_resolution` (e.g. observation: [amoled, 4GB, 23MP]), for the purpose of predicting prices of upcoming smartphones in the market. You have collected a dataset of existing phone models with their features and prices.

i. (2 pts) In your first attempt, you build a linear regression model and observe that the model correctly predicts the training samples but performs poorly on test samples. What could cause the poor performance on test samples? Briefly justify your answer in 1 or 2 sentences.

ii. (4 pts) In your second attempt, you add regularization to your linear regression model and perform model selection to choose the hyperparameter (i.e., the regularization constant). How would you split the data for cross-validation (CV); [*Hint:* Think about the value of k in k-fold CV] and justify your answers in 1 or 2 sentences when:

(a) you have collected a large amount of data.

(b) you only have a small set of examples.

5 Applied: Exploratory Data Analysis [15 points]

Employee turnover is a key problem faced by many organizations. When good people leave, it usually costs the organization substantial time and other resources to find a replacement. Therefore, many organizations try to keep the churn rate at a low level. Imagine a company who now wants to understand its employee churn situation. Its HR (Human Resources) department gives you some data of their employees, and asks you to do exploratory data analysis and to predict employee churn.

You are free to choose any statistics library to analyze the data (e.g., `scipy`, `statsmodels`, etc. Feel free to refer to resources from the recitations). In your answer, please include both the snippets of your code as well as the outputs.

Download the dataset titled `termination.csv` and the `EDA.ipynb` Jupyter notebook template from Canvas. Use the downloaded resources to answer the questions within the notebook.

6 Applied: Sales Prediction via Linear Regression [20 points]

Suppose that we are hired by a client to provide advice on how to improve sales of a particular product. The Advertising data set consists of the `sales` of that product in 200 different markets, along with advertising budgets for the product in each of those markets for three different media: `TV, radio`, and `newspaper`.

It is not possible for our client to directly increase sales of the product. On the other hand, they can control the advertising expenditure in each of the three media. Therefore, if we determine that there is an association between advertising and sales, then we can instruct our client to adjust advertising budgets, thereby indirectly help increase the sales.

In short, our goal is to develop an accurate model that can be used to predict sales on the basis of budgets allocated for different media advertising.

In this problem, you will implement a linear regression model to predict the `sales` of a product in a given market using advertising dollars spent on three different media. Download the dataset titled `Advertising.csv` and use the provided Jupyter notebook `Linear_Regression_Model_Selection.ipynb` template to fill in the missing code snippets and answer the questions listed in the notebook.

Submission Instructions

For this assignment, you are expected to submit the following on *Gradescope*:

- 1 pdf file (e.g. handwritten & scanned or Word saved as pdf) containing answers to conceptual questions
- 1 pdf file (Jupyter notebook saved as .pdf) containing all answers and plots to applied Q5
- 1 pdf file (Jupyter notebook saved as .pdf) containing all answers and plots to applied Q6