Homework 1

Gareth Sun

Course Name: CS5785 Fall 2023
Professor's Name: Prof. Kyra Gan

---

1 Problem 1

For each of the problem, identify whether it's more naturally characterized as a binary classification, multiclass classification, multilabel classification, regression, clustering, density modeling, or RL problem.

1. Given a stream of customers each characterized by some attributes, learn which ads to show them given that you can only show each customer one ad.

   Answer: Multiclass Classification

2. Classify emails as spam or not spam.

   Answer: Binary Classification

3. Given a news article, predict which topics it covers.

   Answer: Multilabel Classification

4. Given a pair of images of faces, identify whether they depict the same person.

   Answer: Multiclass Classification

5. Learn to play a board game against randomly matched opponents on the internet.

   Answer: RL problem

6. Identify whether a new data point is expected given those you have seen before or extremely unlikely.

   Answer: Density Modeling

7. Figure out whether a group of patients happens to naturally break down into some number of subgroups.

   Answer: Clustering

8. Recognize celebrities based on photographs scraped from Twitter.

   Answer: Multiclass Classification

9. Predict the starting salaries of new graduates based on their academic record.

   Answer: Regression

10. Identify the best (personalized) treatment among a set of drugs for a given chronic condition.

    Answer: Multiclass Classification

2 Problem 2

Based on the materials covered so far, in supervised machine learning, why must we make assumptions? Why can't we just learn from data alone?

Answer: In supervised machine learning, we need to make assumptions for following reasons:

1. Overfitting: Without any assumptions or constraints, models can become overly complex and might fit the noise in the training data rather than the underlying distribution.
2. Theoretical Stability: Some algorithms are actually built on the foundations of some key assumptions. If we don't make such assumptions, we may not be able to make it converged or find it optimal point.
3. Simplifying Complexity: The real-life data could be complex and high-dimensional. It may also contain a lot of noise. We may need some assumptions to reduce the complexity and make it computationally feasible.

If we just learn from the data alone, we will not be able to apply it to the problem beyond the training data set. And cause a lot of problems such as the overfitting.

3 Problem 3

Analytical solution of the Ordinary Least Squares Estimation. Consider we have a simple dataset of n labeled data $(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), ..., (x^{(n)}, y^{(n)})$, where data $x \in \mathbb{R}$ and $y \in \mathbb{R}$ is its corresponding label. We use a simple estimated regression function of:

$$\widehat{y}^{(i)} = \theta_o + \theta_1 x^{(i)}$$

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_1 , θ_2 :

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

1. Calculate the partial derivatives $\frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1)$ and $\frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1)$

$$\begin{aligned} J(\theta_0, \theta_1) &= \sum_{i=1}^{n} (y_i^2 + \hat{y}^2 - 2 y_i \hat{y}_i) \\ &= \sum_{i}^{n} y_i^2 + n\theta_0^2 + \theta_1^2 \times \sum_{i}^{n} x_i^2 + 2\theta_0\theta_1 \times \sum_{i}^{n} x_i - 2\theta_0 \times \sum_{i}^{n} y_i - 2\theta_1 \times \sum_{i}^{n} x_i y_i \end{aligned}$$

Answer:

$\frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1) = 2n\theta_0 + 2 \sum_{i}^{n} x_i \theta_i - 2 \sum_{i}^{n} y_i$

$\frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1) = 2 \sum_{i}^{n} x_i^2 \theta_1 + 2 \sum_{i}^{n} x_i \theta_0 - 2 \sum_{i}^{n} x_i y_i$

2. Consider the fact that $J(\theta_0, \theta_1)$ has an unique optimum, which we denote as θ_1^* , θ_2^* .The analytical solution for minimizing θ_1^* , θ_2^* can be obtained 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:

$$\theta_0^* = \overline{y} - \theta_1 \overline{x}$$

and

$$\theta_1^* = \frac{\sum_i^n x_i(y_i - \overline{y})}{\sum_i^n x_i(x_i - \overline{x})}$$

Proof1:

$$\frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1) = 0$$
$$2n\theta_0^* + 2\sum_i^n x_i \theta_1^* - 2\sum_i^n y_i = 0$$
$$n\theta_0^* + \sum_i^n x_i \theta_1^* = \sum_i^n y_i$$
$$\theta_0^* = \bar{y} - \theta_1^* \bar{x}$$

Proof2: Substitute the parameters into the equation $\frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1) = 0$, we can get:

$$2\sum_i^n x_i^2 \theta_1^* + 2\sum_i^n x_i(\bar{y} - \bar{x}\theta_1^*) = 2\sum_i^n x_i y_i$$
$$\theta_1^*(\sum_i^n x_i^2 - \bar{x} \times \sum_i^n x_i) = \sum_i^n x_i y_i - \bar{y} \times \sum_i^n x_i$$
$$\theta_1^* = \frac{\sum_i^n x_i y_i - \bar{y} \times \sum_i^n x_i}{\sum_i^n x_i^2 - \bar{x} \times \sum_i^n x_i} = \frac{\sum_i^n x_i(y_i - \bar{y})}{\sum_i^n x_i(x_i - \bar{x})}$$

3. For the optimalθ_1^* , θ_2^*,calculate the sum of the residuals $\sum_i^n e_i = \sum_i^n (y_i - (\theta_0^* + \theta_1^* x_i))$. What can you learn from the value of $\sum_i^n e_i$?

$$\sum_i^n e_i = \sum_i^n y_i - \sum_i^n \theta_0^* + \sum_i^n \theta_1^* x_i$$
$$= \sum_i^n y_i - n\theta_0^* - \theta_1^* \times \sum_i^n x$$

From Question2, we can know that $\theta_0^* = \overline{y} - \theta_1 \overline{x}$: Then:

$$\sum_i^n e_i = \sum_i^n y_i + \sum_i^n \theta_1^* x_i - \sum_i^n y_i - \sum_i^n \theta_1^* x_i$$

$$\sum_i^n e_i = 0$$

We can learn that: At the optimal situation, the prediction data is the same as the real data. The line produced by regression is the same as the line we want to predict.