THE CHINESE UNIVERSITY OF HONG KONG
Department of Mathematics
Exercises on Polynomial Regression

Learning Outcomes 學習目標

- To understand the concepts of underfitting and overfitting problems.
  了解欠擬合和過擬合問題的概念。
- To choose the right polynomial of degree k for fitting the sample points.
  選擇適當的多項式階數k 來擬合樣本點。
- To split the sample points into training and test datasets for examining the mathematical model in terms of linear and nonlinear regressions.
  將樣本點分為訓練集和測試集，以檢驗數學模型在線性和非線性回歸中的表現。
- To validate the mathematical model by examining different metrics, e.g., Root Mean Square Error, Normalized Mean Square Error, Mean Absolute Error, and Average Relative Error.
  通過檢查不同的誤差指標來驗證數學模型，例如均方根誤差、歸一化均方誤差、平均絕對誤差和平均相對誤差。

Objectives 目標

Our aim is to use the training datasets to obtain the coefficients of a polynomial of degree k, and then predict the trend of the pattern based on the test datasets. We split the entire sample into training and test datasets and labeled them for Fold 1 (Columns 4 - 5), Fold 2 (Columns 6 - 7), and Fold 3 (Columns 8 - 9).
我們的目標是使用訓練數據集來獲取階數為k 的多項式係數，然後根據測試數據集預測模式的趨勢。我們將整個樣本分為訓練集和測試集，並標記為第1折（第4-5列）、第2折（第6-7列）和第3折（第8-9列）。

Problem Description 問題描述

Let us create 21 points following the equation:

$$Y = -X^2$$

with $-1 \leq X \leq 1$ with the space 0.1 by adding the noises drawn from a normally distributed error with a mean of 0 and a standard deviation of 0.05, as shown in the following table:
我們按照方程$Y = -X^2$ 創建21 個點，其中$-1 \leq X \leq 1$，間距為0.1，並加入從均值為0、標準差為0.05 的正態分佈誤差中抽取的雜訊，如下表所示：

	Y	X	Y	X	Y	X	Y	X
1	−0.9369	−1.0000	−0.9369	−1.0000	−0.9369	−1.0000	−0.9369	−1.0000
2	−0.8263	−0.9000	−0.8263	−0.9000	−0.8263	−0.9000	−0.8263	−0.9000
3	−0.5735	−0.8000	−0.5735	−0.8000	−0.5735	−0.8000	−0.5735	−0.8000
4	−0.4264	−0.7000	−0.4264	−0.7000	−0.4264	−0.7000	−0.4264	−0.7000
5	−0.3393	−0.6000	−0.3393	−0.6000	−0.3393	−0.6000	−0.3393	−0.6000
6	−0.3270	−0.5000	−0.3270	−0.5000	−0.3270	−0.5000	−0.3270	−0.5000
7	−0.2064	−0.4000	−0.2064	−0.4000	−0.2064	−0.4000	−0.2064	−0.4000
8	−0.1047	−0.3000	−0.1047	−0.3000	−0.1047	−0.3000	−0.1047	−0.3000
9	−0.0403	−0.2000	−0.0403	−0.2000	−0.0403	−0.2000	−0.0403	−0.2000
10	0.1102	−0.1000	0.1102	−0.1000	0.1102	−0.1000	0.1102	−0.1000
11	0.0382	0	0.0382	0	0.0382	0	0.0382	0
12	−0.0500	0.1000	−0.0500	0.1000	−0.0500	0.1000	−0.0500	0.1000
13	−0.0974	0.2000	−0.0974	0.2000	−0.0974	0.2000	−0.0974	0.2000
14	−0.1045	0.3000	−0.1045	0.3000	−0.1045	0.3000	−0.1045	0.3000
15	−0.1750	0.4000	−0.1750	0.4000	−0.1750	0.4000	−0.1750	0.4000
16	−0.2706	0.5000	−0.2706	0.5000	−0.2706	0.5000	−0.2706	0.5000
17	−0.3474	0.6000	−0.3474	0.6000	−0.3474	0.6000	−0.3474	0.6000
18	−0.5346	0.7000	−0.5346	0.7000	−0.5346	0.7000	−0.5346	0.7000
19	−0.6182	0.8000	−0.6182	0.8000	−0.6182	0.8000	−0.6182	0.8000
20	−0.8719	0.9000	−0.8719	0.9000	−0.8719	0.9000	−0.8719	0.9000
21	−1.0112	1.0000	−1.0112	1.0000	−1.0112	1.0000	−1.0112	1.0000

Questions 問題

Your tasks are:
你的任務是：

1. Enter the Y values in the second column and the X values in the third column in the Linear Regression RShiny Tool app to examine different regression models. Describe the trends of the linear and nonlinear regression lines and curves. Does a high-degree polynomial show a better predicted curve?
在Linear Regression RShiny Tool 應用程式中，將第二列的Y 值和第三列的X

值輸入，以檢查不同的回歸模型。描述線性和非線性回歸線和曲線的趨勢。高階多項式是否顯示出更好的預測曲線？

2. Do four different error metrics yield the same findings in determining the optimal polynomial when three folds are used? 當使用三折交叉驗證時，四種不同的誤差指標是否得出相同的結果來確定最佳多項式？

What have I learned? 我學到了什麼？

Repeat the above procedures for analyzing the following case:
重複上述步驟分析以下案例：

1. Create 100 points following the equation:

$$Y = -X^2$$

with $-1 \leq X \leq 1$ with the space 0.05 by adding the noises drawn from a normally distributed error with a mean of 0 and a standard deviation of 0.05.
遵循方程$Y = -X^2$ 創建100 個點，其中$-1 \leq X \leq 1$，間距為0.05，並加入從均值為0、標準差為0.05 的正態分佈誤差中抽取的雜訊。

2. Split the sample points into five folds:
將樣本點分為五折：

Training Data	Test Data
1 : 20	21 : 100

Test Data	Training Data	Test Data
1 : 20	21 : 40	41 : 100

Test Data	Training Data	Test Data
1 : 40	41 : 60	61 : 100

Test Data	Training Data	Test Data
1 : 60	61 : 80	81 : 100

Test Data	Training Data
1 : 80	81 : 100

3. Find the optimal polynomial of degree k using any error metric.
使用任何誤差指標找到最佳多項式階數k。

4. Interpret your observations.
解釋你的觀察結果。

How can I practice? 我如何練習？

Consider a real-life example of polynomial regression in finance. You're analyzing the relationship between an employee's years of experience and their salary. You suspect the relationship isn't linear and that a higher-degree polynomial might better capture the salary progression over time.
考慮一個財務中多項式回歸的實例。你正在分析員工的工作年限與薪水之間的關係。你懷疑這種關係不是線性的，高階多項式可能更好地捕捉薪水隨時間的變化。

Years of Experience	Salary (in dollars)
1	50000
2	55000
3	65000
4	80000
5	110000
6	150000
7	200000

Answer the following questions:
回答以下問題：

1. Let X = Years of Experience be the independent variable and Y = Salary be the dependent variable. Use a quadratic polynomial (degree 2) to model the relationship between years of experience and salary. In other words, write down the quadratic polynomial regression equation using β_0, β_1, and β_2 as coefficients of the polynomial.
   設X = 工作年限 為自變量，Y = 薪水 為因變量。使用二次多項式（階數為2）來建模工作年限與薪水之間的關係。換句話說，使用β_0、β_1 和β_2 作為多項式的係數，寫出二次多項式回歸方程。

2. Find the coefficients that minimize the difference between the predicted salaries and the actual salaries in the dataset using the method of least squares.
   使用最小二乘法找到使預測薪水與數據集中實際薪水之間差異最小的係數。

3. Discuss whether a higher-degree polynomial would provide any benefit for predicting the salary for 10 years of experience.
   討論高階多項式是否會對預測10年工作經驗的薪水有任何幫助。