ISEP
IG.3510 Machine Learning
2024-2025

Tutorial course 1 : Introduction to Machine Learning

Patricia CONDE-CESPEDES

62705 GUO Xiaofan

1. Machine Learning problems

Exercise 1:

a) Type of the scenario: Regression

Input variables: Profit - quantitative variables

Number of employees - quantitative variables

Industry – qualitative variable - retail and wholesale.

Output variables: CEO salary – quantitative variables

The number of observations n: 500

The number of predictors p: 3 (profit, employees, industry)

b) Type of the scenario: Classification

Input variables: Price charged – quantitative variable

Marketing budget – quantitative variable

Competition price - quantitative variable

Output variables: Success or failure – qualitative variable

The number of observations n: 20

The number of predictors p: 3

c) Type of the scenario: Regression

Input variables: The % change in the British market – quantitative variable

The % change in the German market – quantitative variable

The % change in the Chinese market – quantitative variable

Output variables: The % change in the American dollar – quantitative variable

The number of observations n: 52 (weekly data for the whole year 2012)

The number of predictors p: 3

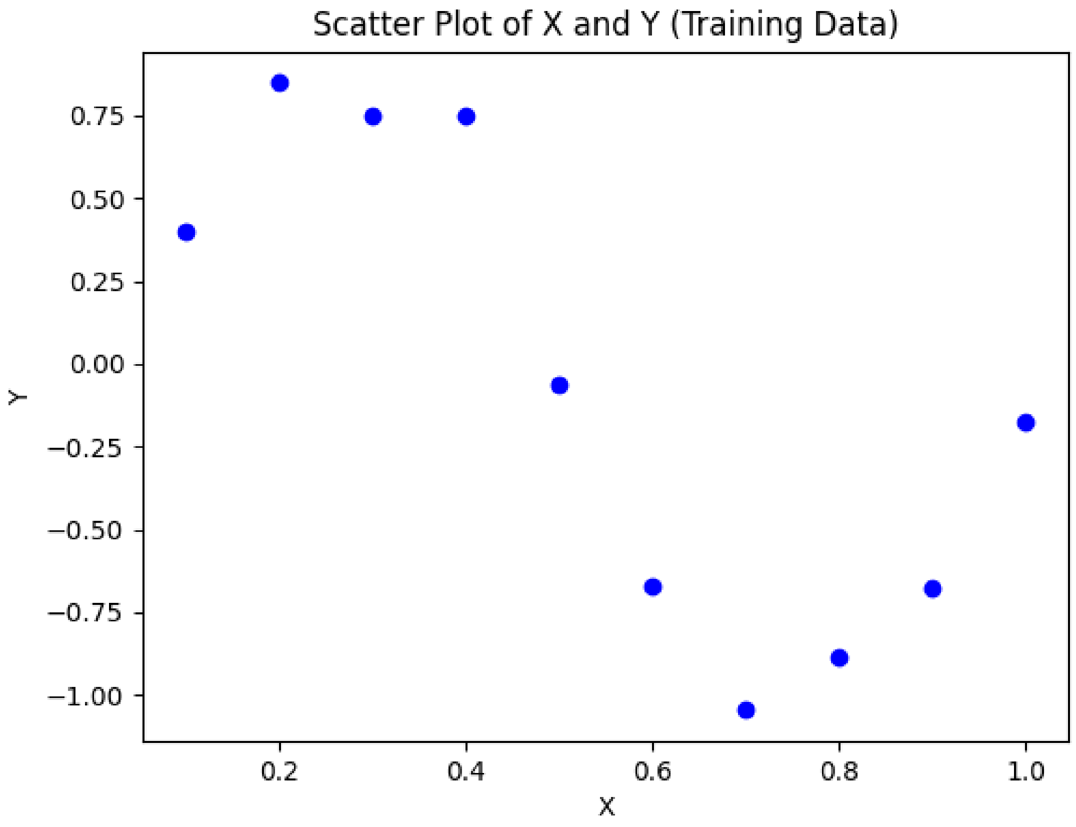
Exercise 2:

1) This is a supervised problem, the training dataset consisted of ratings between 1 and 5.
2) If we are interested in detecting groups of customers that like similar kinds of movies, then it's an unsupervised problem. Clustering algorithms can be used to group users with similar interests.

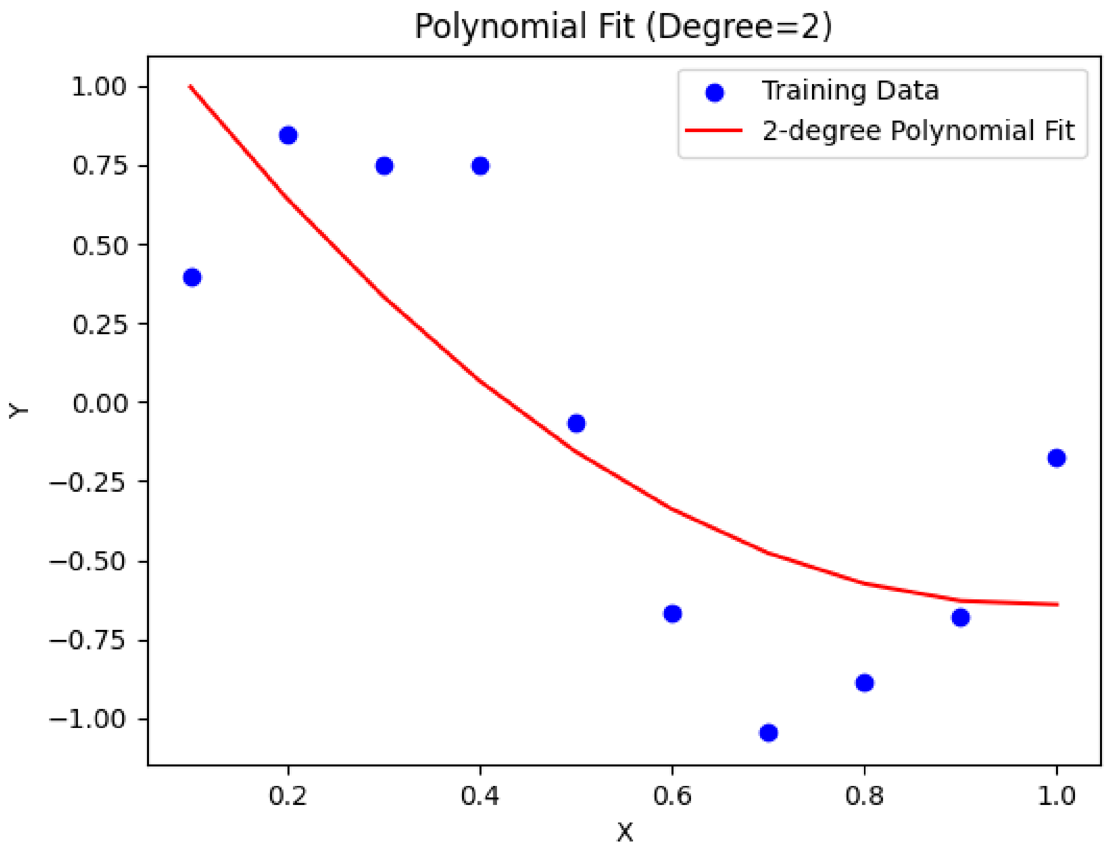
2. Application: the bias-variance trade-of

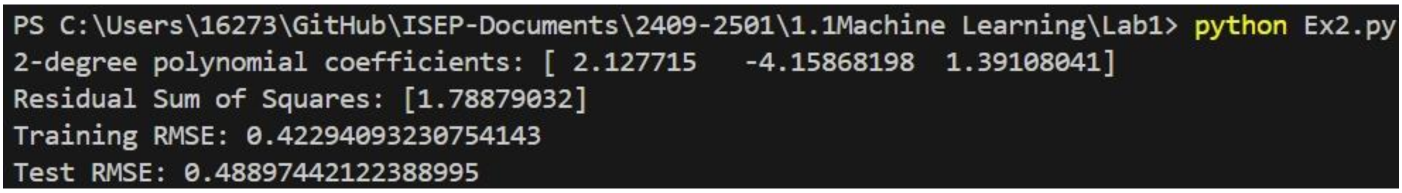
2.2 Choice of model complexity and bias-variance trade-of

1) Read and visualize the data.



2) Fitting a polynomial model





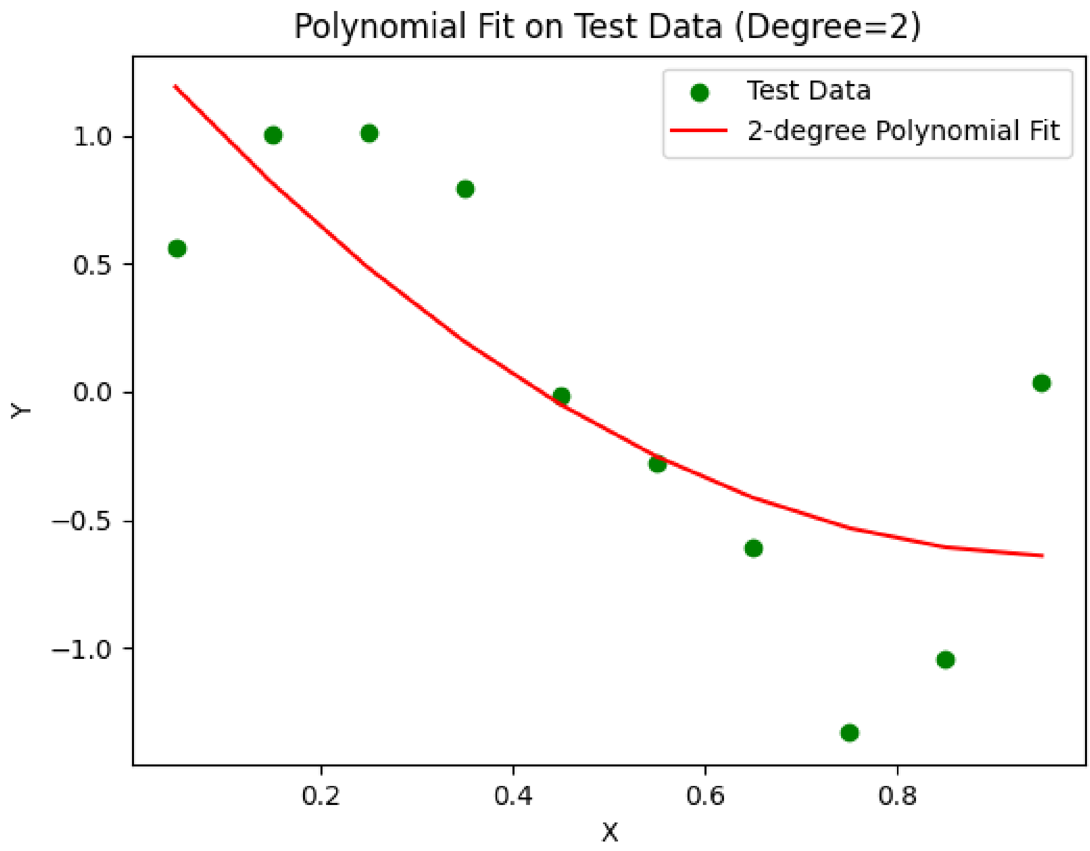
```
PS C:\Users\16273\GitHub\ISEP-Documents\2409-2501\1.1Machine Learning\Lab1> python Ex2.py
2-degree polynomial coefficients: [ 2.127715   -4.15868198  1.39108041]
Residual Sum of Squares: [1.78879032]
Training RMSE: 0.42294093230754143
Test RMSE: 0.48897442122388995
```

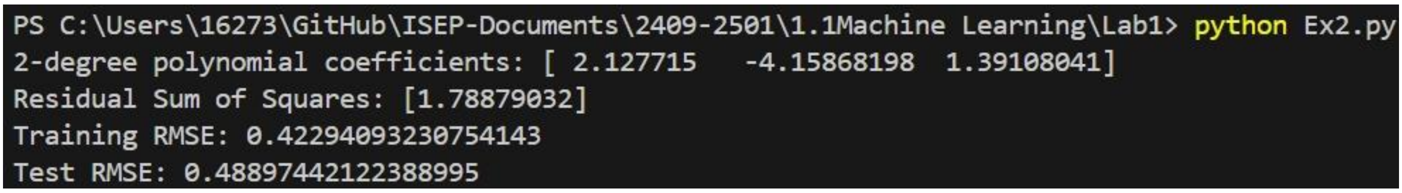
a) RSS = 1.78879, which indicates that the model has a small fitting error on the training data.

b) Training RMSE = 0.423, showing that the model fits the training data very well.

c) Under the 2nd-degree polynomial model, the model successfully captures the nonlinear trend in the data, and the fitting effect is good. This is evident from the low RMSE, indicating that the model is capable of accurately modelling the underlying data structure.

3) Test set prediction and error calculation





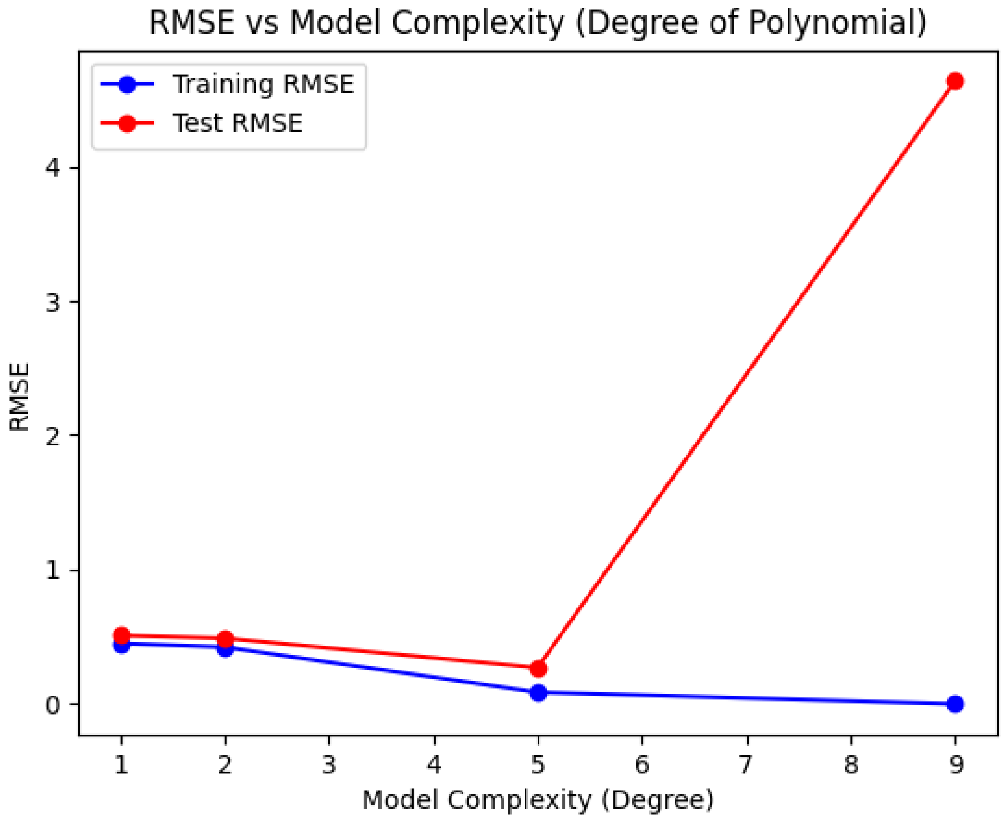
```
PS C:\Users\16273\GitHub\ISEP-Documents\2409-2501\1.1Machine Learning\Lab1> python Ex2.py
2-degree polynomial coefficients: [ 2.127715   -4.15868198  1.39108041]
Residual Sum of Squares: [1.78879032]
Training RMSE: 0.42294093230754143
Test RMSE: 0.48897442122388995
```

a) Test RMSE = 0.489, slightly higher than the RMSE of the training set.

b) The prediction error of the model on the test data is slightly larger, but still within a reasonable range.

c) On the test data, the performance of the quadratic polynomial model is similar to that of the training data, and the fitting curve can better capture the trend of the test data.

4) Model complexity and bias-variance trade-off



a) **Training set RMSE:** As the polynomial order increases, the training set RMSE continues to decrease. This is because the model complexity increases, allowing it to better fit the training data.

b) **Test set RMSE:** For the 2nd and 3rd-degree polynomials, the test set RMSE remains small and performs well. However, starting from the 5th-degree polynomial, the test set RMSE rises sharply, indicating that the model is overfitting the training data, which reduces its ability to generalize to new, unseen data.

c) **Bias-variance trade-off (1st to 3rd degree polynomials, Low-order models):** have a good bias-variance balance, with small RMSE for both training and test sets, and stable performance.

d) **Bias-variance trade-off (5th and above polynomials, High-order models):** As the model complexity increases, the RMSE of the training set decreases, but the RMSE of the test set increases significantly, indicating that the model is overfitting.