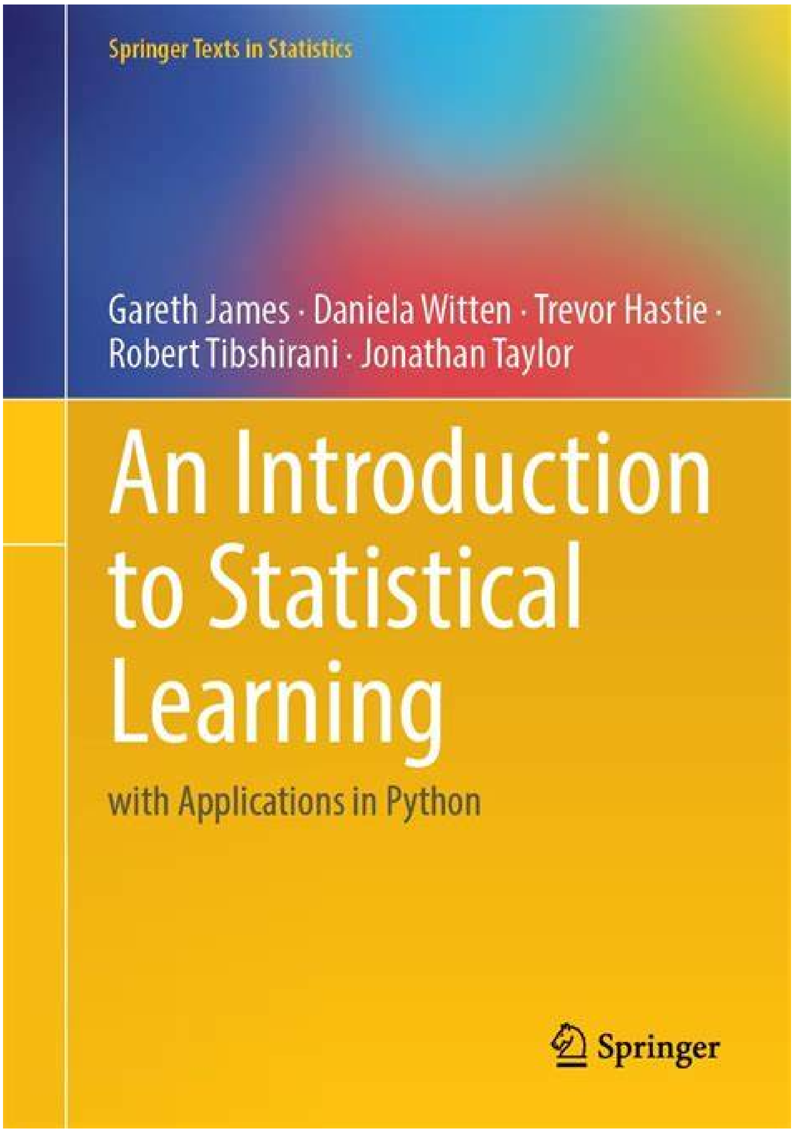
Introduction to Statistical Learning Exercises - Python-Based

Athil George

Updated Last 1/20/24


Springer Texts in Statistics
Gareth James · Daniela Witten · Trevor Hastie · Robert Tibshirani · Jonathan Taylor
An Introduction to Statistical Learning
with Applications in Python
Springer

Contents

1 Linear Regression

1.1 Problem 1.1

For each of parts (a) through (d). indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than the inflexible method. Justify your answer.

1. The same size n is extremely large, and the number of predictors p is small.
2. The number of predictors p is extremely large, and the number of observations n is small.
3. The relationship between the predictors and the response is highly nonlinear.
4. The variance of the error terms is high.

1. For a large n and a small p, flexible statistical learning methods are likely to perform better. With a large sample size, these methods can better capture the underlying patterns in the data, and having fewer predictors helps in avoiding overfitting.
2. In this case, inflexible methods might perform better. When the number of predictors is extremely large compared to the number of observations, flexible methods may struggle with overfitting. Inflexible methods, like linear regression, could be more robust in such situations.
3. If the relationship between the predictors and the response is highly nonlinear, then a more flexible approach would be more appropriate since regression won't be able to capture nonlinear functions.
4. In this case, flexible methods might perform worse. Flexible methods are more prone to overfitting noisy data, so if the variance of the error terms is high, they might capture the noise in addition to the underlying pattern, leading to poor generalization. Inflexible methods can be more robust to high variance in the error terms.

1.2 Problem 1.2

Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

1. We collect a set of data on the top 500 forms in the US. For each firm, we record profit, number of employees, industry and the CEO salary. We are interested in understanding what factors affect the CEO salary.
2. We are considering launching a new product and wish to know whether it will be a success or failure. We collect data on 20 similar products that were previously launched. For each product, we have recorded whether it was a success or failure, the price charged for the product, the marketing budget, the competition price, and ten other variables.
3. We are interested in predicting the percent change in the USD/Euro exchange rate to the weekly changes in the world stock markets. Hence, we collect weekly data for all of 2012. For each week, we record the percent change in the UDS/Euro, the percent change in the US market, and the percent change in teh US market.

1. Inference. We are interested in understanding the parameters that affect the CEO salary.
2. Prediction. We are solely interested in predicting whether the product is a success and failure and do not care about the parameters affecting it.
3. Prediction. We are not interested in exploring the parameters that affect the percent change US/Euro exchange rate.

1.3 Problem 1.3

We now revisit the bias-variance decomposition.

1. Provide a sketch of typical (squared) bias, variance, training error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods toward more flexible approaches. The x-axis should present the amount of flexibility in the method, and the y -axis should represent the values for each curve. There should be five curves. Make sure to label each one.
2. Explain why each of the five curves has the shape displayed in part (a).

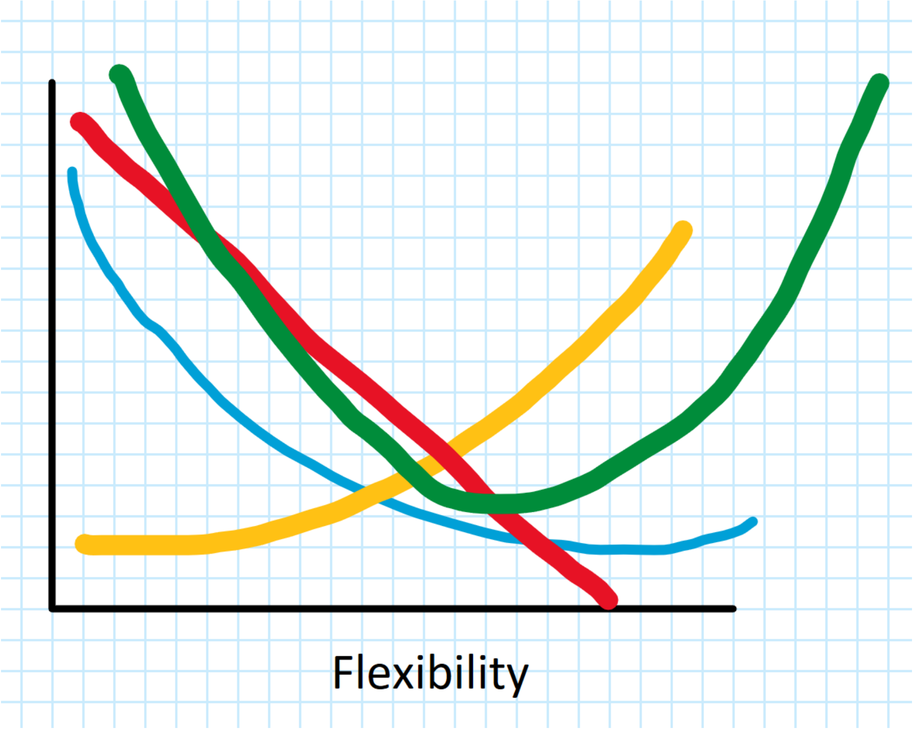


Figure 1: Flexibility in Learning Models. Red: Training Error, Green: Test Error, Yellow: Squared Bias, Blue: Variance

(a) Squared Bias: Initially high as the model is too simple, decreases as flexibility increases but may increase again if the model becomes overly complex.

(b) Variance: Initially low as the model is too simple, increases as flexibility grows, and may decrease if the model becomes too complex.

(c) Training Error: Decreases as flexibility increases since more flexible models can fit the training data better.

(d) Test Error: Initially decreases as flexibility increases, but after a point, the model starts overfitting, causing an increase in test error.

(e) Bayes (Irreducible) Error (Constant line): Represents the inherent noise or uncertainty in the data, which cannot be reduced by the model. It remains constant across different levels of flexibility. Not shown on the graph.

1.4 Problem 1.4

You will now think of some real life applications for statistical learning.

(a) Describe three real life applications in which classification might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

(b) Describe three real-life applications in which regression might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

(c) Describe three real-life applications in which cluster analysis might be useful.

(a) i. Engine Health Monitoring: The classification response variable is the health status (normal, degraded, faulty). The predictors could be various sensor data such as temperature, pressure and mechanical vibration during engine operation. The goal of this application is to predict the health status of the engine based on the collected data to proactively schedule maintenance and avoid failures.

ii. Flight Safety Assessment: The classification response variable is the safety status of the flight (safe, unsafe). The predictors could be weather conditions and aircraft performance parameters. The goal of this application is inference and to infer the safety status based on the given predictors, helping the airline make real-time decisions.

iii. Aircraft System Fault Detection: The classification response is the identification of specific system faults, such as landing gear, control surface, and structural malfunction. The predictors could be the system sensor data, control inputs, or historical fault data. The goal for this application is to predict faults in the aircraft systems before they occur, facilitating proactive maintenance.

(b) i. Fuel Efficiency Prediction: The response is the fuel efficiency of an aircraft measured by the parameter SFC, or specific fuel consumption. The predictors could be the altitude, speed, weight, engine type, and atmospheric conditions. The goal of this application is to predict fuel efficiency based on the given predictors to optimize flight planning and reduce operational costs.

ii. Aircraft Performance Optimization: The response could be performance metrics such as speed, and endurance. The predictors could be Aircraft design parameters, weather conditions, and operational parameters. The goal of this application is to optimize the aircraft's performance by predicting the outcomes of different design or operational choices.

iii. Material Fatigue Life Prediction: The response is the fatigue life of aircraft materials. The predictors could be material properties such as stress, temperature and usage history. The latter is usually measured through the number of takeoff/landing cycles that the aircraft has gone through. The goal is to predict the remaining fatigue life of materials, enabling timely replacements and avoiding structural failures.

(c) i. Air Traffic Pattern Analysis: Clustering can help identify common air traffic patterns, improving air traffic control strategies and optimizing routes for fuel efficiency.

1.5 Problem 1.5

What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible one? When might a less flexible approach be preferred?

The advantages of a flexible approach are:

- Better Fit to Complex Patterns.
- Adaptability to Diverse Datasets.
- Higher Predictive Accuracy.

The disadvantages of a flexible approach are:

- Overfitting.
- Computational Intensity.
- Difficulty in Interpretability.

A more flexible approach might be preferred when:

- Modeling complex, nonlinear relationships.
- Working with Heterogeneous Datasets.
- accuracy is the primary goal.

A less flexible approach might be preferred when:

- Interprebility is preferred
- Computational resources are limited.
- There is a risk of overfitting due to limited data or noisy features.

1.6 Problem 1.6

Describe the differences between a parametric and non-parametric statistical learning approach. What are the advantages of a parametric approach to regression or classification (as opposed to a non-parametric approach)? What are its disadvantages?

1.7 Problem 1.7

The table below provides a training data set containing six observations, three predictors, and one qualitative response variable. Suppose we wish to use this data set to make a prediction for Y when $X_1 = X_2 = X_3$ using K-nearest neighbors.

(a) Compute the Euclidean distance between each observation and test point, $X_1 = X_2 = X_3 = 0$.

(b) What is our prediction with K = 1? Why?

(c) What is our prediction with K = 3? Why?

(d) If the Bayes decision boundary in this problem is highly nonlinear, then would we expect the best value for K to be large or small? Why?

(a) The Euclidean distance is defined as the following.

$$E(x, x_i) = \sqrt{\sum_{i=1}^{N} (x - x_i)^2}$$

N is the number of predictors.

Obs.	X_1	X_2	X_3	E	Y
1	0	3	0	3	RED
2	2	0	0	2	RED
3	0	1	3	$\sqrt{10}$	RED
4	0	1	2	$\sqrt{5}$	GREEN
5	-1	0	1	$\sqrt{2}$	GREEN
6	1	1	1	$\sqrt{3}$	RED

Table 1: Problem 1.7 Table with Euclidean Errors

(b) For $K = 1$, the prediction is green since the 1st nearest neighbor is observation 5, whose classification is green.

(c) For $K = 1$, the prediction is red since red is the most frequent classification of the first three neighbors.

(d) If the Bayes decision boundary is highly nonlinear, it means that the relationship between the input features and the classes is not well-captured by a simple linear model. Therefore, we use a larger K for improved performance.

1.8 Problem 1.8 - Applied

Exploration of this data set seems to show that private colleges are more selective than public colleges with the metric for selectivity being acceptance rate.

1.9 Problem 1.9 - Applied

(a) The quantitative predictors in this dataset are given as follows.

- MPG
- Cylinders
- Displacement
- Horsepower
- Weight
- Acceleration
- Year

The quantitative predictors in this dataset are given as follows.

- Origin
- Name

1.10 Problem 1

You are given 3 data points.

$$P_1 = (3,3)$$
$$P_2 = (2,1)$$
$$P_3 = (1,1)$$

Find a line in the form

$$y = ax + b$$

such that the square error is minimized. This problem is called method of least squares. The square of trinomials formula might be useful.

$$(x+y+z)^2 = x^2 + y^2 + z^2 + 2xy + 2yz + 2xz \tag{1}$$

Start by defining an error cost.

$$\phi(a,b) = \sum_{i=1}^{N} (ax_i + b - y_i)^2$$

For this case, we have three points, so $N = 3$.

$$\phi(a,b) = (a+b-1)^2 + (2a+b-1)^2 + (3a+b-3)^2$$

Expand and simplify using the square of trinomials formula:

$$\phi(a,b) = 14a^2 + 12ab - 24a + 3b^2 - 10b + 11$$

In order to optimize this function, we have to take its derivative and set it equal to 0. Note there are two parameters, so we will have two equations.

$$\frac{\partial \phi}{\partial a} = 28a + 12b - 24 = 0$$
$$\frac{\partial \phi}{\partial b} = 12a + 6b - 10 = 0$$

You should find that $a = 1$ and $b = -1/3$.

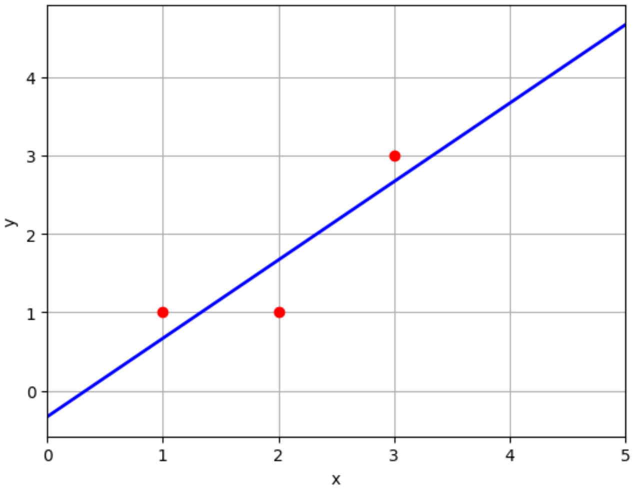


Figure 2: Least Squares Optimization

1.11 Problem 2

In this exercise, we will create a Python program that numerically performs linear regression given a set of data. The variables will be x and y, which are both scalar values. This means that the regressions polynomial will be on the XY plane. The polynomial we will use will be 1st order (Linear) and 2nd order (Quadratic). Follow the following steps to construct the code.

(a) Define a matrix, P, that packages the x and y data. Structure the matrix in the following form:

$$P = \begin{bmatrix} X_1 & Y_1 \\ . & . \\ . & . \\ . & . \\ X_N & Y_N \end{bmatrix}$$

For example, if we are given 3 data points $(1, 1)$, $(2, 3)$, and $(3, 6)$, the matrix P will be:

$$P = \begin{bmatrix} 1 & 1 \\ 2 & 3 \\ 3 & 6 \end{bmatrix}$$

(b) The objective of the numerical approach is to find the best-fit parameters a, b, and c for the Linear and Quadratic polynomials below.

$$y = ax + b$$

$$y = ax^2 + bx + c$$

Note that the process of finding the parameters and predicting y given x for both the Linear and Quadratic polynomials is called Linear Regression. To find these best-fit parameters, we need to minimize the following cost function which minimizes the **least squares** error between the predicted and observed y values. According to the above description, the cost function, denoted as ϕ can be written as follows.

$$\phi(a, b) = \sum_{i=1}^{N} (y - y_i)^2 \tag{2}$$

Define a function in python that computes and returns this cost given the P matrix defined previously. The inputs should contain the fitting parameters only.

(c) Use the fmin function in Python to numerically minimize the cost function.

The The Python implementation can be found here.

1.12 Problem 3

In the previous problem, Python was leveraged to find a numerical solution to the linear regression problem to find the polynomial coefficients by minimizing a cost function defining a least squares error. It turns out that there is a unique analytical solution to the least squares problem. Redefine the cost function as follows. Define a new variable θ and X (upper case). We will now deviate from the single input variable case and consider multiple input variables. Let θ be an array storing the fitted parameters.

$$\theta = \begin{bmatrix} a \\ b \\ . \\ . \\ . \end{bmatrix} \tag{3}$$

Let X be defined the following way.

$$\begin{pmatrix} X_{1,1} & X_{1,2} & \cdots & X_{1,n} \\ X_{2,1} & X_{2,2} & \cdots & X_{2,n} \\ \vdots & \vdots & \vdots & \vdots \\ X_{m,1} & X_{m,2} & \cdots & X_{m,n} \end{pmatrix} \tag{4}$$

$X_{m,n}$ corresponds to the nth input variable of the mth sample. For example, consider a model with 2 input variables and one output variable (x_1,x_2,y). The data points are (1,2,0), (2,3,2), (3,4,4), (4,3,7). The associated X matrix is:

$$X = \begin{bmatrix} 1 & 2 \\ 2 & 3 \\ 3 & 4 \end{bmatrix}$$

Similarly, let the matrix Y be:

$$Y = \begin{bmatrix} Y_1 \\ . \\ . \\ . \\ Y_m \end{bmatrix}$$

The new cost function can be rewritten as:

$$\phi(\theta) = (X\theta - Y)^T(X\theta - Y)$$

(a) Find an analytical solution for the linear regression problems. This solution extends to polynomials beyond 2nd order!

(b) What are the drawbacks to the analytical solution from a mathematical standpoint? Hint: What condition needs to be satisfied to guarantee a unique solution?

(c) Create a python function to execute the regression using the analytical solution.

(a) Use the following matrix calculus identities for the following derivation.

$$\frac{\partial(AX)}{\partial X} = A^T$$

$$\frac{\partial X^T A}{\partial X} = A$$

$$\frac{\partial X^T A X}{\partial X} = AX + A^T X$$

Expand cost function.

$$J(\theta) = ((X\theta)^T - Y^T)(X\theta - Y)$$

$$J(\theta) = (\theta^T X^T - Y^T)(X\theta - Y)$$

$$J(\theta) = \theta^T X^T X\theta - \theta^T X^T Y - Y^T X\theta + Y^T Y$$

The approach will be to take the first derivative of the cost function and set it equal to 0. Using the matrix identities and differentiating term by term:

$$\frac{\partial J(\theta)}{\partial \theta} = 2X^T X\theta - X^T Y - X^T Y = 0$$

$$\frac{\partial J(\theta)}{\partial \theta} = X^T X\theta - X^T Y = 0$$

Solving this algebraic matrix equation:

$$\boxed{\theta = (X^T X)^{-1} X^T Y}$$

(b) The condition that needs to be satisfied for a unique solution is that $X^T X$ must be invertible. If this quantity is not invertible, there are infinite solutions that could represent the best fit parameters. For the matrix, X, to be invertible, all the features must be linearly independent. For example, lets say there are three feature, namely, X_1, X_2, and X_3. Suppose that $X_3 = X_1 + X_2$. The matrix X would not be invertible do to multicollinearity.

(c) The The Python implementation can be found here.

2 Chapter 3 Exercises

2.1 Problem 1

Describe the null hypotheses to which the p-values given in Table 3.4 correspond. Explain what conclusions you can draw based on these p-values. Your explanation should be phrased in terms of sales, TV, radio, and newspaper, rather than in terms of the coefficients of the linear model. This implies that TV and radio have a larger effect on sales while newspaper advertising statistically plays a small role in sales.

The null hypothesis for each value would correspond to the coefficients of each predictor being equal to 0. Therefore, from the p-values in Table 3.4, we can conclude that the intercept, TV, and radio coefficient are non-zero, while the newspaper coefficient is likely to be zero.

2.2 Problem 2

Carefully explain the differences between the KNN classifier and KNN regression methods.

The KNN classifier returns the class that has the most occurrences in the K nearest point of the test point x_0.

$$\boxed{\hat{y} = \arg\max(\sum_{i=1}^{K} h(y_i = c))}$$

The function h is equal to 1 if $y_i = c$ and 0 if $y_i \neq c$.

$$\boxed{\hat{y} = \frac{1}{K}\sum_{i=1}^{K} y_i}$$

The KNN regression method chooses the K nearest points closest to x_0 and the estimate is the mean of those points.

2.3 Problem 2

Suppose we have a data set with fve predictors, X_1 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = Interaction between GPA and Level. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to ft the model, and get β_0 = 50, β_1 = 20, β_2 = 0.07, β_3 = 35, β_4 = 0.01, β_5 = 10

The regression model is:

$$Y = \tag{5}$$