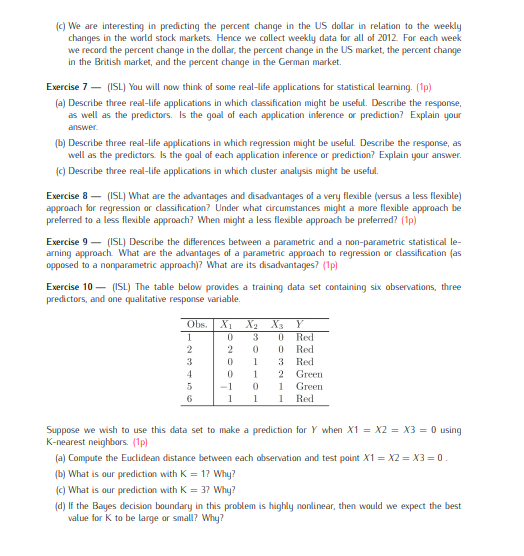
Obraz zawierający tekst

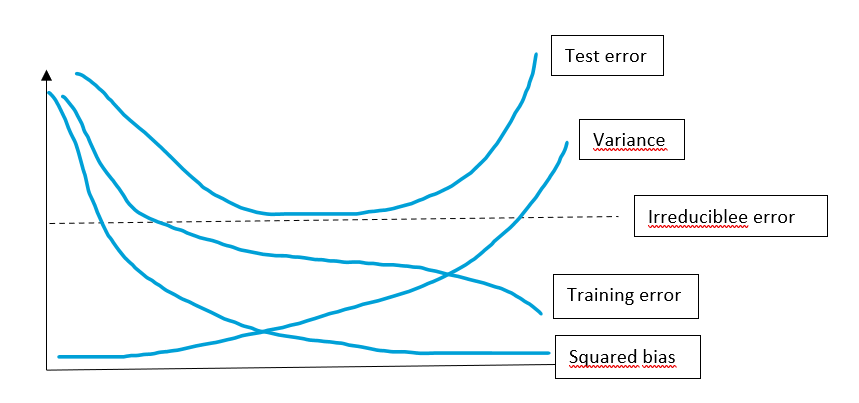
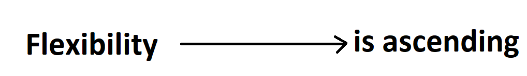
Opis wygenerowany automatycznie



Exercise 4 (During the exercises)

1. In this case a flexbile method will be better, we can use it to get more informations from the large n pool
2. Oposite, in this case we need to use inflexible method, becouse there is great risk of overfitting
3. Non-linear relation, in this case then inflexible method will prove badly, we choose flexible then
4. If we have extemly high variance oft he error terms we should expect a lot of noise in the relations between samples. We should choose inflexible method then

Exercise 5



Let’s explain what we have here. Starting with test error, it declines as flexbility is increasing but start going up at some point later.

The next is variance and it will increase monotonically as the squared bias will be decreasing in the same way (the more flexible the method, the more the variance will increase and the squared bias will decrease).

The irreducible error is standard constant line used for determination, test error will allways be above the irreducible error curve becouse it allways will have greater value than variance.

The training error will be declining monotonically as flexbility will be increasing.

Exercise 6

1. We know that the CEO salary is continuous variable, we have then a regression problem. In this case we will be interested in inference, becouse we want to know how CEO salary will be impacted by given factors. The sample size n = 500 (500 frims providing data) and predictors p = 3 (amount of employers, profit gain and industry).
2. Now we will have classification problem, becouse we will be trying to determine if a launch of a new product will be succes/failure. But in this case we also trying to predict succes or failure with less concern for understanding the underlying relatnionships. The sample size n = 20 (Data collected on 20 similiar products allready launched) and predictors p = 13 (all variables mentioned)
3. In the last one we will have regression problem and prediction problem, we have situation in which % change in the US dollar value is quantitative dependent variable. We also trying to predict becouse we are interested in predicting the % change value. The sample size n = 52 (becouse it’s weekly data over 2012) and predictors p = 3 (becouse % change in USA, UK and in Germany).

Exercise 7

1. Classification example 1 – Will this student pass Data Mining Exercises or not? (Response: Succes/Failure, Predictors: Time invested, Possible interests and hobbies, Commitment and determination, Predispositions, etc., Goal: Prediction)

Classification example 2 – Will the SARS-CoV-2 virus cases be decreased or not (Response: Will the cases SARS-CoV-2 decreased or not?, Predictors: Resarching programs, Medical investment, Involvement of the people, Dedication of the medical service, Human sacrafice, Advertising campaigns, etc., Goal: Prediction)

Classification example 3 – Will the new Amazon game be successfull or not (Response: Successfull/Not Successfull, Predictors: Invested funds, Advertising campaigns, Interest in the product, Game developing experience, etc., Goal: Prediction)

1. Regression example – Increase in fuel prices in the world (Response: What will be the standarized price per barrel for 2040, Predictors: Oil reserves deplitation, Influences of the OPEC members, Fuel consumption ratio, Alternative power sources development, etc., Goal: Inference)

Regression example 2 – Increase in housing prices in the Dolnośląskie region (Response: What will be the actual price of the houses in Dolnośląskie region in the next year, Predictors: Number of flats, Demand and interest in the region, Possible earnings, Population growth rate, etc., Goal: Inference)

Regression example 3 – Possibilities of a new water supply system (Response: According to system specification parameters, the value X will determine effectivness of the system, Predictors: Type of installation, Water supply management guidelines, System operation form, Efficiency of individual system components, Degree of defectiveness of installation components, etc., Goal: Inference)

1. Cluster example 1 – Classify people into high/middle/lower classes in the XVIII century (Response: This human should be classified as high/middle/lower class, Predictors: Place of birth and family name, Amount of property and wealth, Education and skills, Social estimate, etc., Goal: Prediction)

Cluster example 2 – Tax classification in Germany (Response: That person will get that tax class, Predictors: Marital status, Amount of income, Number of jobs, Type of partner's tax class, Civil Status, etc., Goal: Prediction)

Cluster example 3 – PEGI classification for computer games (Response: This game will be rated in category 3/7/12/16/18 according to PEGI, Predictors: Scenes of Violence, Type and form of violence, Scary scenes, Rapidly changing images, Nudity, Scenes with Drugs, Profanity, etc., Goal: Prediction)

Exercise 8

Very flexible approach will be generally better when the relationship is non-linear, and there is a lot of data sample. Flexbile approach should also be preffered when irreducible error is low.   
  
Inflexible approach will be better when the relationship is highly linear, we don’t have lot of data samples and the overall irreducible error is high.

Exercise 9

Parametric approach reduces the estimation down to the given set of parameters, and non-parametric don’t. In this case non-parametric will perform better with non-linear patterns given, and a large pool of samples. On the other hand parametric approach will allow to simplyfy the overall structure to few parameters without need for large sample pool. Also when there is high risk of overfitting a non-parametric will perform worse. Futhermore we should also consider that in case of non-parametric approach the data will be harder to interpret, becouse we don’t use interpretation structure of parameters in this case.

Exercise 10

1. Observation 1

Observation 2

Observation 3

Observation 4

Observation 5

Observation 6

1. We know that the nearest point to the test point (0,0,0) is given by the Observation 5 (-1,0,1) with euclidean distance . If we know that the Observation 5 was Green, that K=1 test point will also be Green.
2. In this case we know that the test point (0,0,0) has three nearest points given by Observation 5 (-1,0,1) with euclidean distance , Observation 6 (1,1,1) with euclidean distance and Observation 2 (2,0,0) with euclidean distance 2. If we know that the Observation 5 was Green, Observation 6 was Red and Observation 2 also was Red, we can predict that K=3 test point will also be Red.
3. We should expect best value for K tob e small, becouse a highly non-linear Bayes boundary suggest that there is less advantage with futher generalization due to high variance