Classification and Representation Learning Autonomous Work 2

Aujasvi, *Master student in Data Science,* University of Nantes (Polytech Nantes)

Abstract—The Autonomous work-2 is about polynmial regression on a given dataset ie. 'polynome.data'. Here, we are using some of the built in functions of numpy library in python like the polyfit and polyval. Polyfit is a numpy function which fits a polynomial p(x) = p[0] * x**deg + ... + p[deg] of degree to points (x, y) and returns a vector of coefficients p that minimises the squared error, whereas polyval evaluates a polynomial at specific values.

Index Terms—Polynomial regression, quadratic error, generalization error, empirical error, K-fold cross-validation.

1 Questions

1.1 Question 1

Run the script crossval.py in order to visualize the data (16 Examples). According to you which order should the polynome have at least, in order to perform the good fit?

Answer From the visual observation, the order of polynome should be atleast 2 when we are plotting the dataset points on the x-y plane they follow a parabola shaped curve.

1.2 Question 2

Complete **crossval.py** in order to apply the polynomial regression to the data contained in polynome.data, with degrees ranging from 1 (linear regression) to 20. Visualize all the fits.Starting from which degree do they look OK?

Answer By applying the polynomial regression with degree ranging from 1 to 20 with the help of **for** loop, we observe starting from **order 3** onwards the curve starts satisfying the dataset points and looks OK.

1.3 Question 3

For each of these hypothesis, print the quadratic error on the training set. How does the training error evolve when the degree of the polynome is increased? What is the risk by taking the hypothesis with the smallest training error?

Answer Quadratic error on the training set is given by $MSE = ((Y - y) * *2).mean(axis = None)$. As the order of the polynomial increases the quadratic error decreases. The main drawback and risk in taking this hypothesis is that the higher order curve tries to **overfit** the data and due to which robustness decreases.

1.4 Question 4

Apply simple cross-validation to find the degree for the polynomial regression. Separate the data set into S-train (70 percent) and S-test (30 percent). Train each polynome with and compute the generalization error on S-test. Which degree of the polynome gives the minimal empirical error?Why?

Answer Generalization error on each degree of the polynome is computed and degree of **order 4** gives the minimal empirical error because it is generalizing well. But when we are increasing the degree of the polynome then curve tends to overfit. In case of overfitting the train error is high and test error is low. Cross-validation provides us with the capability to more accurately estimate the test error, which we will never know in practice.

1.5 Question 5

70 percent of the 16 training examples is really not a lot for a good fit. Apply the K-fold cross validation (with k = 1 or 2). Does it change something to the optimal degree of the polynome?

Answer Applying K-Fold cross validation, minimum average mse is for **order 5**. So, optimal degree of the polynome is 5.

- *Aujasvi, Department of Data Science, University Of Nantes (Polytech Nantes), Nantes, 44000.*
E-mail: aujasvimoudgil@gmail.com