EE214 (25sp) Assignment 1

Euiseok Han 20200690

April 11, 2025

1 Approach and Solution Outline

To analysis two different types of basis functions in depth, I observed MSE while varying degree of polynomial regression model and gaussian basis regression model. I drew MSE-degree graph and found the best model for both regression models and analyzed the graph. In addition to the analysis with respect to the degree, to investigate bias and variance of both models, I splitted the training dataset into 4 sub datasets with $(824/4) = 206$ sample data. By making four models, I observed the bias and variance of both models along with the change of the degree.

For real-world applicability, to obtain insight of the datasets, I reviewed the paper written along with the dataset publication. Though the result of the paper "Modeling of Strength of High-Performance Concrete Using Artificial Neural Networks" by I-Cheng Yeh may not serve as a suitable baseline for my experiment, the paper provides a traditional statistical model as follows:

$$f'_c(t) = aX^b \cdot [cln(t) + (d)]$$

Here, comparing my two models with traditional statistical model given in the paper derived from conventional material modeling process will demonstrate my models' suitability in the real-world. Hence, I compared MSE of our models with the traditional statistical model.

Lastly, while reviewing the paper, I noticed that the concrete compressive strength, which is our target value is function of age. To investigate the sensitivity of regression accuracy to feature selection, I performed polynomial linear regression using three different features: water-to-binder ratio, cement, and age. Since logarithmic function can be expanded into polynomial series, I expected polynomial regression with age as an input feature will show the best result.

2 Results and Analysis

2.1 Identification of Best Model

To find the best model for both polynomial regression model and gaussian basis regression model, I conducted regression while changing degrees. Then I found the best model based on the mean square error. In polynomial regression model, the mean squared error was the lowest when the degree was 3, with a value of approximately 51.1. In the gaussian basis regression model, the lowest mean square error was 179.67 which is model of degree 7. Notice that I designed the mean value of each gaussian basis in the gaussian basis regression model to run from $1/(degree + 1)$ to $(degree + 1)/(degree + 1)$, since I scaled features already. In conclusion, the best degree for polynomial regression model and gaussian basis regression model was 3 and 7.

In addition, the best polynomial regression model showed better performance than the best gaussian basis regression model. As will be discussed in the fourth question, the polynomial regression model is estimated to have shown better performance than the Gaussian basis regression model because it made use of all the input features.

2.2 Underfitting/Overfitting Identification

Observing the overall trend in polynomial regression model and comparing the `train_mse` and `test_mse`, it was found that underfitting occurred at degree 0 and 1, while overfitting became evident

from degree above 4. Since we used polynomial model, it is expected to the function diverge as model complexity increases.

In gaussian basis regression model, as `train_mse` decreases when the degree increases from 1 to 7, I identified underfitting at lower degrees. However, at higher degrees, I was not able to find the evidence of overfitting. Therefore, I drew plots comparing the training data and the test data, but was not able to observe any signs of overfitting. It is considered that the absence of overfitting comes from the characteristic of gaussian function, which doesn't diverge so that the effective of high degree is significantly lower than polynomial regression model. Also, by noticing that the gaussian function with coefficient 0 almost doesn't contribute to the overall model, the absence of overfitting seems feasible.

2.3 Bias–Variance Tradeoff Explanation

To analyze bias and variance of both models, I divided the training dataset into four subsets. Each subsets containing $(824/4 = 206)$ data produce individual model.

As I expected, in polynomial regression model, high bias has observed at lower degrees(degree of 0 and 1). Also, variance has strictly increased as degree increases. Here, I limited y axis to 300 because both bias and variance became too large at higher degrees. The dramatic increase of both bias and variance is reasonable since a polynomial function eventually diverges at the infinity.

In gaussian basis regression model, as in polynomial regression model, bias has decreased until the degree of the best model. Hence, there was high bias at lower degrees. However, variance has indeed increased, but the increasing magnitude was very small. Overall, the $\text{bias}^2 + \text{variance}$ seemed to be stabilized. As I discussed previously, the overall low variance at higher degree compare to polynomial regression model is though to be caused by the characteristic of gaussian function, which converges at infinity and has 0 values everywhere when the coefficient is small.

At lower degrees, bias was high, and at higher degrees, variance was high. Nevertheless, their varying magnitude were different significantly according to the type of basis function. It will be important to choose right basis function depending on the characteristic of a dataset.

2.4 Real-World Applicability Analysis

As I discussed earlier, concrete compressive strength is a function of multiple features. It is hard to model the output with a single feature. To compare the result with our models, I performed regression with respect to age, since the traditional statistical model is function of age, and got approximately 80.46 `test_mse` value. This result indicates incredible performance of the best polynomial regression model. However, the driving force of the outstanding performance of the best polynomial regression model comes from its non-linearity. Computing higher-order terms introduces non-linearity.

To verify the effect of non-linearity, I performed regression in terms of single feature, cement. Interestingly, the `test_mse` showed similar result with the gaussian basis model, which also used cement as a single feature input. In addition to the regression result in terms of cement, I performed extra regression with respect to `water_to_biner_ratio` and `age`. As one can expect, polynomial regression with respect to age showed the best performance among three regression results, each were conducted with respect to single feature: `water_to_binder_ratio` and `cement` and `age`. In the traditional statistical model, the concrete compressive strength was logarithmic function of age, which can be expended into polynomial series. In this respect, it is considered that the polynomial regression model can be performed better when the output function can be expanded into polynomial series, which is plausible.

Polynomial regression model was more robust to the non-linearity feature of a dataset, but more vulnerable to the possibility of overfitting. In contrast, gaussian regression model was more stable than the polynomial model, but it is harder to take account of multiple features. In conclusion, one should select adequate basis function by considering characteristic of a dataset.

2.5 Alternative Basis Functions Analysis

Along with polynomial regression model and gaussian basis regression model, I tried regression with different basis model: polynomial function with regularization and fourier function. I didn't tried gaussian function with regularization as a basis function, because there wasn't no significant sign of overfitting in the gaussian basis regression model.

As one can expect, polynomial function with regularization showed the best performance among all experiment that I conducted. The best model was degree of 4 with lambda 1.0. Also, the `train_mse` and `test_mse` shows that the overfitting at higher degrees reduced significantly. Despite the improvement in the performance, the overfitting at the higher degrees was still significant.

I also performed regression with fourier function, since every analytic function can be converted into fourier series. Though it showed quite decent output values, its performance was poorer than both polynomial and gaussian model. Also, fourier model showed overfitting at very high degrees.

Polynomial with regularization function and fourier function both showed decent output values. However, both models also weren't able able to overcome their intrinsic defect - overfitting. Again, one should select appropriate model and hyperparameters by considering the characteristic of a dataset.

3 AI usage

I only used chatGPT for this assignment and most of questions were related to concept, debugging, and verification.

I asked questions related to concept as the following. Why the degree should degree + 1 in `X = np.ones((n_samples, degree + 1))`?. How to calculate covariance matrix? How `np.linalg.eig` outputs its result: in column vector or row vector, is it normalized or not? Does `PolynomialFeatures` also calculate combined terms? Is StandardScalar normalize just mean value or variance is also normalized? What is binder in conventional material modeling?

I also asked questions for debugging such as the following. Why `X = data.drop('Concrete compressive strength(MPa, megapascals)', axis=1).values` doesn't working? (There supposed to be space after the last ')'). Is `matmul` and `@` operator different?

Overall, I mainly used AI tool to enhance my conceptional knowledge and find specific debugging errors. AI tool was extremely helpful for fixing grammar error and finding adequate library function. Without AI tool, it would take much longer while finding and learning how to use a specific library function. Rather than, for designing experiments and writing code was based on myself and the knowledge by studying codes given in the regular coding session.

4 reflection

I performed polynomial regression and gaussian basis regression and observed the results. I mainly focused on relating the performance of regression model with the characteristic of a dataset. While designing the experiment, selecting adequate single feature or combination features was challenging. I tried as much as possible to compare both polynomial regression model and gaussian basis model with the traditional statistical model, which indicates the applicability of both models in real world.

Since polynomial regression model with degree of two showed better performance than conventional statistical model, There is a non-negligible possibility that it could be applied in real-world scenarios. Nevertheless, low stability of multi-variable regression will make polynomial model hard to actual use in real-world. Although this experiment has limitations in gaining insights into multi-variable regression due to the lack of diverse feature selection, conducting more comprehensive and varied experiments could lead to meaningful results.

References

[1] I.-C. Yeh, "Modeling of strength of high-performance concrete using artificial neural networks," *Cement and Concrete Research*, vol. 28, no. 12, pp. 1797–1808, 1998.

A notice

* Please understand that I had changed the title of the given dataset for parsing.