Clarifications on Tutorial 4

Rishabh Anand, TG-06

Contents

1 Relationships **1**
1.1 Definitions . . . 1
1.1.1 Model and Algorithm . . . 1
1.1.2 Bias . . . 2
1.1.3 Variance . . . 2
1.1.4 Noise . . . 2
1.1.5 Complexity of Target . . . 3
1.1.6 Bias Variance Decomposition . . . 3

2 Tutorial Clarifications **4**
2.1 Question 1 . . . 4
2.2 Question 2 . . . 4
2.3 Question 3 . . . 5

Abstract

This document contains answers to questions raised in class on Tutorial 4. It covers topics such as Bias, Variance, Model fitting, and Noise, and the causal relationships between them. I also cover and explain the tutorial questions and the students' doubts on them.

1 Relationships

Bias, Variance, Model fitting (over and under), and Noise (deterministic and stochastic) are all related in some way or the other. Knowing what will happen when is important in diagnosing your model's behaviour. But first, let's define these terms one more time for brevity:

1.1 Definitions

1.1.1 Model and Algorithm

The Model is the final representation or relationship between x and y. We assume this model has the best fit on the given data. The algorithm is the procedure by which we aim to obtain this model (Linear Regression, Logistic Regression, k-NN, etc.).

In reality, algorithm and model are one and the same. When said, we assume it's already best fit. No one calls it "ML Algorithm" specifically.

1.1.2 Bias

It's is the algorithm's tendency to pick a simpler model over a relatively complex model. High bias occurs when your algorithm misses out on catching certain aspects or patterns in the data. This is typically called **Underfitting**.

Note: Do not confuse Bias with **Inductive Bias**, which is the model's implicit assumptions on training data (ie. the observed learnings) which allows it to predict well on unseen data.

1.1.3 Variance

It's the algorithm's sensitivity to tiny fluctuations or minor changes (i.e., noise) in your dataset. High variance occurs when your algorithm learns to model the noise in the data aside from the features that it's actually supposed to focus on. This causes **Overfitting**. As such, your model works very well on training data but performs poorly on testing data.

1.1.4 Noise

There are two types of Noise – deterministic and stochastic. Stochastic noise (SN) comes from the dataset and cannot be minimised or removed. Suppose we have the best possible (ideal) model function $f(x)$:

$$\hat{y} = f(x) + \text{SN} \tag{1}$$

SN lies completely out of the model $f(x)$'s control. Deterministic noise (DN) comes from the model making explicit assumptions about the data. DN is a part of y that cannot be modeled well. Suppose we have a hypothesis function $h(x)$ that aims to be very similar to $f(x)$:

$$\begin{aligned} \hat{y} &= f(x) \\ &= h(x) + \text{DN} \end{aligned}$$

DN is part of the hypothesis function (acts like the Bias) and can be minimised. It's a result of the model being underfit or overfit. If the model is not fit well, DN is high, while well-fit models have minimal DN. A diagrammatic representation can be seen in Figure 1.

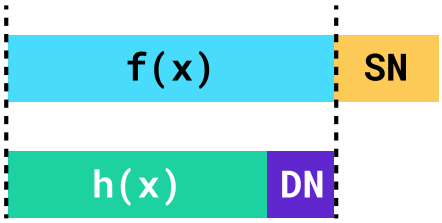

Figure 1: This figure represents the relationship between model fitting and the two types of noise. DN can be minimised by getting a good generalisation/fit on the dataset. SN cannot be removed or minimised as it's part of the data, not the model or the algorithm's ability to learn.

1.1.5 Complexity of Target

When measuring the "amount" of overfitting, we use the degree of the polynomial as the proxy. This "degree of polynomial" is called the complexity of the target. Suppose we have data that's modeled by a polynomial 5. When our algorithm comes up with a model of degree 3, we say the model has underfit the data since there's a difference of +2 degrees. The concept of overfitting also applies when the model's degree is much higher than the data's required degree.

> **Note:** Of course, in CS3244, we have oversimplified the notion of overfitting as the difference between the data's degree and model's degree. In reality, overfitting cannot be quantitatively described – only proxies and approximations can be used/made.

When I say model's *degree* or data's *degree*, I mean the degree of the polynomial of the model or data.

1.1.6 Bias Variance Decomposition

You've come across this equation in lecture before:

$$\text{Error}(Y) = \text{Bias}^2 + \text{Variance} + \text{Stochastic Noise} \tag{2}$$

In the lecture notes, you see *Irreducible Error* mentioned a lot. This is just another term for Stochastic Noise which I use in the equation above. Also, a lot of students had issues with decomposing this equation and its components. Let's break it down further:

$$\begin{aligned}\textbf{Error}(Y) &= \textbf{E}[(h(x) - f(x))^2] \\ &= \mathbb{E}[(h(x) - \bar{h}(x))^2] \;+\; \mathbb{E}[(\bar{h}(x) - f(x))^2] \;+\; \sigma^2\end{aligned}$$

where $\bar{h}(x) \approx \frac{1}{K}\sum_{i=1}^{K} h_i(x)$ is an average learner over many datasets (average the performance to get this) and $f(x)$ is the target function we're trying to approximate. $\textbf{E}[(h(x) - f(x))^2]$ is the total MSE error we wish to decompose, $\mathbb{E}[(h(x) - \bar{h}(x))^2]$ is the Variance, $\mathbb{E}[(\bar{h}(x) - f(x))^2]$ is the Bias /

Deterministic Noise, and σ^2 is the Stochastic Noise.

Given the current dataset, we can see that $\mathbb{E}[(h(x) - \bar{h}(x))^2]$ tells us the difference between the current best-fit model and the average learner. Now, imagine we had infinite data. $\mathbb{E}(\bar{h}(x) - f(x))^2$ tells us the maximum extent or limit we can reach in terms of learning even with infinite data. σ^2 is the noise in the data we can never remove or minimise.

2 Tutorial Clarifications

2.1 Question 1

Here, we test your understanding of the causal relationships for Overfitting. From the lecture we know that increasing the number of samples *decreases* Overfitting, while the two types of noise *increase* the chance of Overfitting. H is the hypothesis function and f is the target function we're trying to approximate. I urge you to look at the diagram in Figure 1 when answering this question. Think of it from a "capacity perspective" – when one goes down, the other has to go up, and so on.

1a. From the lecture notes, we know increasing samples decreases overfitting. Increasing noise makes increases the chance of overfitting because it forces the model to learn the noise as well (the model can't tell apart what kind of noise it's learning). Increasing target complexity causes underfitting, not overfitting, since the data's degree is higher than the model's degree.

1b. The target function has a much higher degree now compared to the hypothesis function. The deterministic noise now increases because there is a lot more noise in the more complex dataset that cannot be modeled or understood by the model. Since the model's degree is lower than the data's degree, the model will likely underfit.

1c. The model has a relatively much higher degree than what it was (though, still $\leq$ than the target function's degree). The deterministic noise now decreases because the target function's complexity. Since the model's degree is relatively higher and closer to the data's degree, the model will likely overfit.

2.2 Question 2

Many students asked if the band (blue and green) represents the mean accuracy, minimum accuracy, and maximum accuracy. This is **NOT** true. From our 10-fold Cross Validation, we get 10 values for accuracy, each for training (blue) and validation (green).

For each set of 10 values, we calculate the mean μ and standard deviation σ. The top bound of the band represents the accuracy +1 s.d. and the bottom bound represents the accuracy -1 s.d. away.

Since s.d. is a measure of how further data is from the mean, we can use this method to show the graph. Also, we since s.d. and Variance are related (Var = σ^2), we can use the Variance as well.

2.3 Question 3

Most students had some issue breaking down these terms and decomposing Total Error as a function of Bias, Variance, and Irreducible Error.

> **Fun fact:** k-NNs can be used for Regression! If you remember from the lecture, you can consider the real-valued predictions of the k nearest points around the new test point and get their average. This average is then assigned as the prediction for the new test point.

3a. Let's look at the colour-coded equation above. Let's also break down the whole statement into some smaller expressions. The aim of the question is to derive the equation provided, not start from there. There is no point looking at the different components. We have to start from somewhere else given our current knowledge and land up at the provided equation.

The question tells us $\mathbb{E}(\epsilon) = 0$ and $\text{Var}(\epsilon) = \sigma^2$ which is the case for stochastic noise; we've seen this before in the lecture. We can conclude σ^2 is the Stochastic Noise / Irreducible Error from the data that cannot be removed.

We know Variance is defined as the spread of predictions across the true value (if plotted on some target board; look at my tutorial slides for more on that). This shows that the model has learned the underlying noise as well which should not happen theoretically. We aim to capture the variance of such predictions from the model $\hat{f}(x_0)$ in this context. This is why we look at $Var(\hat{f}(x_0))$. After some manipulation and expansion, we can get the term for the Variance. You can find my annotated working on the slides.

Next, we know Bias as the "closeness of predictions to the truth". We can measure closeness in terms of distance between true value and prediction. This is also given by our squared error metric $(f(x_0) - \mathbb{E}[\hat{f}(x_0)])^2$. We get slightly different predictions for the same input around the true value, which is why we look at the expected value of the regression model's prediction. The detailed expansion can be found on the slides.

Finally, we have 3 terms to put together to get the final Total Error. We add them to get the provided equation:

$$\text{Error}(x_0) = \sigma^2 + (f(x_0) - \frac{1}{k}\sum_{i=1}^{k} f(x_i))^2 + \frac{\sigma^2}{k} \tag{3}$$

where σ^2 is the stochastic noise, $(f(x_0) - \frac{1}{k}\sum_{i=1}^{k} f(x_i))^2$ is the Bias, and $\frac{\sigma^2}{k}$ is the Variance.