Machine Learning for Data Analysis

MSc in Data Analytics

CCT College Dublin

Underfitting and Overfitting

Week 2

Lecturer: Dr. Muhammad Iqbal*

Email: miqbal@cct.ie

Classification and Regression

- In classification, the goal is to predict a ***class label***, which is a choice from a predefined list of possibilities.
- We used an example of classifying **irises** into one of three possible species.
- Classification is separated into ***binary classification***, which is the special case of distinguishing between exactly two classes, and ***multiclass classification***, which is classification between more than two classes.
- We can think of binary classification as trying to answer a **yes/ no** question.
- Classifying emails as either **spam or not spam** is an example of a binary classification problem.
- In this binary classification task, the **yes/no question** being asked would be “Is this email spam?”

Classification and Regression

- A clear way to distinguish between classification and regression tasks is to ask whether there is some kind of continuity in the output.
- If there is continuity between possible outcomes, then the problem is a **regression problem.**
- Think about predicting annual income. There is a clear continuity in the output.
- Whether a person makes **$40,000** or **$40,001** a year does not make a tangible difference, even though these are different amounts of money; if our algorithm predicts $39,999 or $40,001 when it should have predicted $40,000, we don't mind that much.
- If we consider the language of a website (which is a classification problem), there is no matter of degree. A website is in one language, or it is in another. There is no continuity between languages, and there is no language that is *between* English and French.

Generalization, Overfitting, and Underfitting

- In supervised learning, we want to build a model on the training data and then be able to make accurate predictions on new, unseen data that has the same characteristics as the training set that we used.
- If a model is able to make accurate predictions on unseen data, we say that it is able to ***generalize*** from the training set to the test set.
- In ML models, our objective is to build a model that is able to generalize as accurately as possible.
- If the training and test sets have enough in common, we expect the model to be accurate on the test set.
- However, there are some cases where this can go wrong. For example, if we allow ourselves to build very complex models, we can always be as accurate as we like on the training set.

Generalization, Overfitting, and Underfitting

Which customers are considered to send out promotional emails who are likely to actually make a purchase?

Example data about customers

Age	Number of cars owned	Owns house	Number of children	Marital status	Owns a dog	Bought a boat
66	1	yes	2	widowed	no	yes
52	2	yes	3	married	no	yes
22	0	no	0	married	yes	no
25	1	no	1	single	no	no
44	0	no	2	divorced	yes	no
39	1	yes	2	married	yes	no
26	1	no	2	single	no	no
40	3	yes	1	married	yes	no
53	2	yes	2	divorced	no	yes
64	2	yes	3	divorced	no	no
58	2	yes	2	married	yes	yes
33	1	no	1	single	no	no

- Let's take a novice data scientist that wants to predict whether a customer will buy a boat, given records of previous boat buyers and customers that were interested in buying a boat or not.
- The novice data scientist comes up with the following rule:
- **"If the customer is older than 45, and has less than 3 children or is not divorced, then they want to buy a boat."**
- When asked how well this rule of his does, the data scientist answers, "It's 100 percent accurate!" On the data present in the table, the rule is perfectly accurate.

Generalization, Overfitting, and Underfitting

Which customers are considered to send out promotional emails who are likely to actually make a purchase?

Example data about customers

Age	Number of cars owned	Owns house	Number of children	Marital status	Owns a dog	Bought a boat
66	1	yes	2	widowed	no	yes
52	2	yes	3	married	no	yes
22	0	no	0	married	yes	no
25	1	no	1	single	no	no
44	0	no	2	divorced	yes	no
39	1	yes	2	married	yes	no
26	1	no	2	single	no	no
40	3	yes	1	married	yes	no
53	2	yes	2	divorced	no	yes
64	2	yes	3	divorced	no	no
58	2	yes	2	married	yes	yes
33	1	no	1	single	no	no

- As the age data appears only once, we could say people who are 66, 52, 53, or 58 years old want to buy a boat, while all others don't. We are not interested in making predictions for this dataset as It is clear from this dataset.
- **We want to know if new customers are likely to buy a boat.**
- We need to find a rule that will work well for new customers, and achieving 100 percent accuracy on the training set does not help us there.
- It seems too complex, and it is supported by very little data. For example, the "or is not divorced" part of the rule hinges on a single customer.

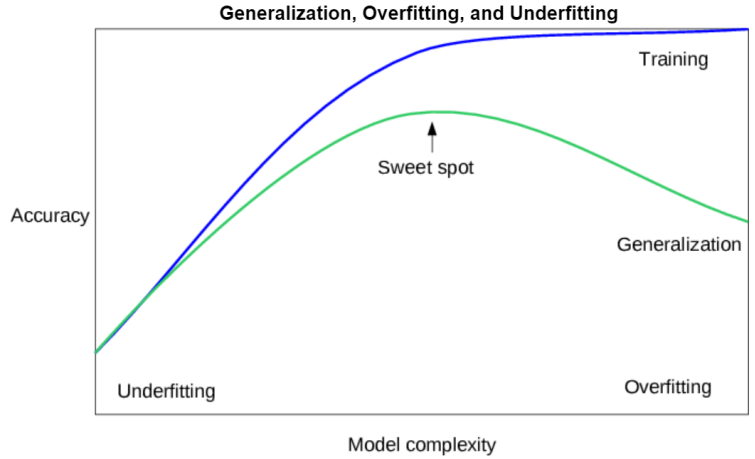
Generalization, Overfitting, and Underfitting

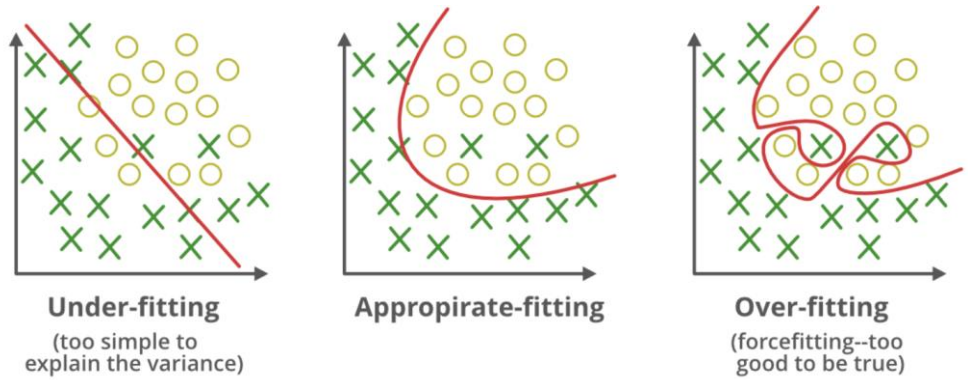
- The only measure of whether an algorithm will perform well on new data is the evaluation on the test set.
- However, we expect simple models to generalize better to new data.

- If the rule was “People older than 50 want to buy a boat,” and this would explain the behaviour of all the customers, we would trust it more than the rule involving children and marital status in addition to age.

- We always want to find the simplest model. Building a model that is too complex based on the amount of information in the table is called ***overfitting*.**

- **Overfitting** occurs when you fit a model too closely to the particularities of the training set and obtain a model that works well on the training set but is not able to generalize to new data.





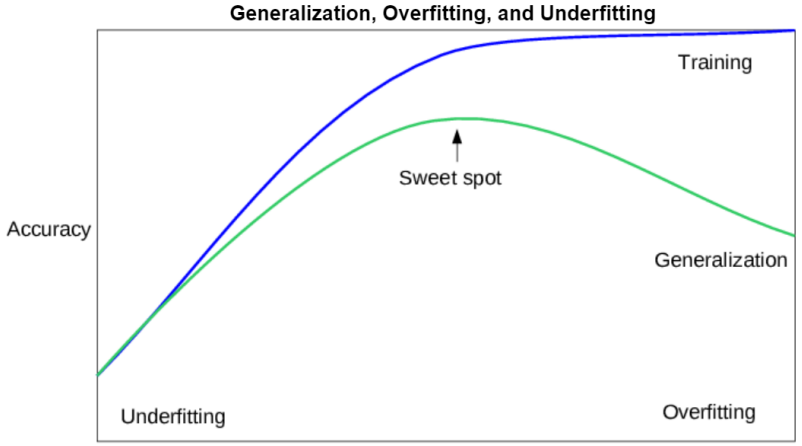
Generalization, Overfitting, and Underfitting

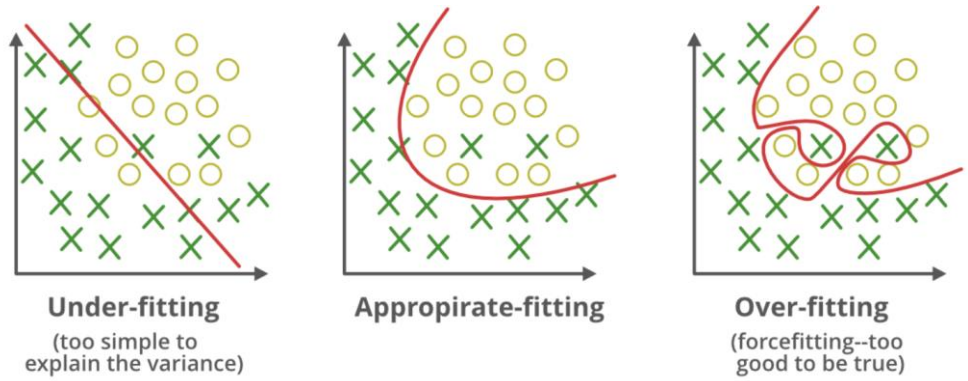
- If our model is too simple, “Everybody who owns a house buys a boat” then we might not be able to capture all the aspects of and variability in the data, and the model will do badly even on the training set. Choosing too simple a model is called ***underfitting***.

- The more realistic we allow our model to be, the better we will be able to predict on the training data. However, if our model becomes too complex, we start focusing too much on each individual data point in our training set, and the model will not generalize well to new data.

- There is a **sweet spot** in between that will yield the best generalization performance.

- This is the model we want to find. The trade-off between **overfitting** and **underfitting** is illustrated in Figure.





Relation of Model Complexity to Dataset Size

- It's important to note that model complexity is linked with the variation of **inputs contained in your training dataset**: If the dataset contains a larger variety of data points, we can construct a more realistic and complex ML model without overfitting.

- The more complex models can be formed appropriately in the case of supervised learning tasks having more data.

- In the real world, we have the ability to decide how much data to collect, which might be more beneficial than tweaking and tuning our model.

- Based on the conceptual understanding of Underfitting and Overfitting, Machine Learning modellers should check the model complexity to the available side of the data.

Resources/ References

- Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition, Aurélien Géron, O'Reilly Media, September 2019, ISBN: 9781492032649.
- Introduction to Machine Learning with Python, Andreas C. Müller and Sarah Guido, O'Reilly Media, Inc. October 2016.
- Data Mining And Machine Learning, Fundamental Concepts And Algorithms, MOHAMMED J. Zaki, Wagner Meira, Jr., Cambridge CB2 8BS, United Kingdom, 2020.
- Discovering Knowledge In Data: An Introduction To Data Exploration, Second Edition, By Daniel Larose And Chantal Larose, John Wiley And Sons, Inc., 2014.
- UCI Repository: http://www.ics.uci.edu/~mlearn/MLRepository.html
- Statlib: http://lib.stat.cmu.edu/
- Some images are used from Google search repository (https://www.google.ie/search) to enhance the level of learning.