COMP4388: **MACHINE LEARNING**

Overfitting

Dr. Radi Jarrar
Department of Computer Science
Birzeit University

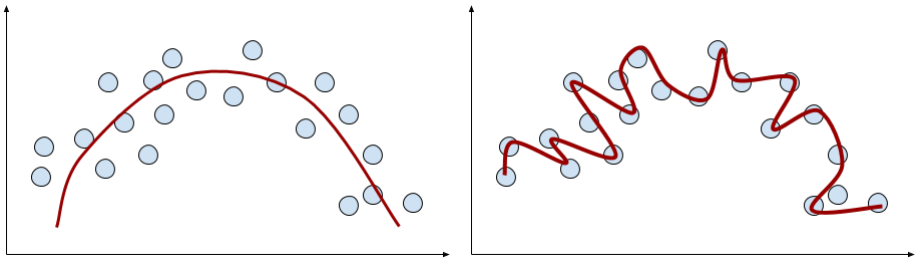
Overfitting

- Overfitting is a common danger that occurs while generating machine learning models
- This could mean that the model has involved in *learning the noise*
- It means the generated model performs well on the training data but it cannot genarlise on any new data
- It occurs when the model fits the data **too well**

STUDENTS-HUB.com Uploaded By: Jibreel Bornat

Overfitting (2)

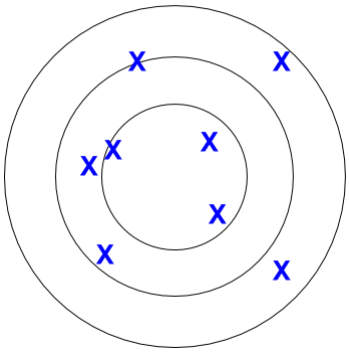
- The model is capturing all details of the data even though some of them (the noise) is considered as outlier and can be removed



Just right | Overfitted model

Overfitting (4)

- This may result when the learning model is more complex than is necessary to fit the data in the target function
- Technically, it means that the model shows low bias and high variance



STUDENTS-HUB.com Uploaded By: Jibreel Bornat

When does it happen?

- It happens when the ML algorithm includes the noise in the data while building the hypothesis (generating the model)
- This means, selecting a hypothesis with much lower error on the training data but it generates a high error on the test data
- One of the reasons of overfitting is applying complex algorithms on small datasets

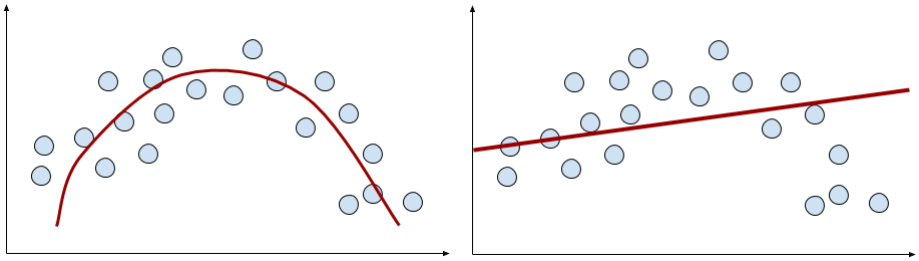
When does it happen? (2)

- Overfitting also results when the number of parameters grows in the dataset (i.e., a very complex hypothesis)
- For instance, if a dataset with 100 observations, each comprises 600 attributes is very prone to overfitting
- On the other hand, a data set with 1000 observations, each contains 6 attributes is much more immune to overfitting

STUDENTS-HUB.com Uploaded By: Jibreel Bornat

Underfitting

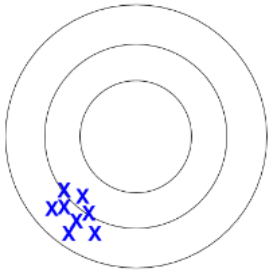
- Underfitting is the opposite of overfitting
- It occurs when the model/algorithm fails to capture underlying trend of the data
- This means that the model/algorithm does not fit the data too well



Just right | Underfitted model

Underfitting (2)

- In the case of underfitting, the model shows low variance and high bias
- It is the result of an excessively simple model
- Both overfitting and underfitting lead to poor prediction on new test data
- Using validation (in specific, cross-validation) while training the model is one method to prevent the over/underfitting



STUDENTS-HUB.com Uploaded By: Jibreel Bornat

Problems of overfitting

- The main problem of overfitting is that the model will appear to perform very well on the training data while it will perform poorly (or even badly) on the test data (or new data from the same problem)
- If a classifier fits the noise along with the signal, then it won't be able to separate signal from noise on new data
- This is a problem in the context on Machine Learning as the ultimate goal is to build a model that generalises to new data

Avoiding Overfitting

- Increasing the number of observations
- Overfitting can be prevented using cross-validation on data to compare their predictive accuracies while training the model
- In corss-validation, the training dataset is split into a number of folds (subsets) that are used to test the performance of the generated model while the training process is taking place

STUDENTS-HUB.com Uploaded By: Jibreel Bornat

Avoiding Overfitting

- Assume a training dataset of 1,000 records
- It can be divided (somehow equally) into 3-subsets each of around 333 records namely set1, set2, and set3

Cross-validation

- Using the three aforementioned sets, the classifier is built during the training process as follows:
  - Two sets (in this case) are used to train and generate the model (assume set1 & set2) and set3 is used to test the generated model and the results are recorded
  - The process is repeated to generate another model using another combination of subsets (assume set2 & set3) and the test is performed on set1
  - The process is repeated again using set1 & set3 and the test is performed on set2
  - The sets are then combined and become the result of the entire dataset

STUDENTS-HUB.com Uploaded By: Jibreel Bornat

Cross-validation (2)

- In corss-validation, the training dataset is split into a number of folds (subsets) that are used to test the performance of the generated model while the training process is taking place
- Assume a training dataset of 1,000 records
- It can be divided (somehow equally) into 3-subsets each of around 333 records namely set1, set2, and set3

Bias and Variance

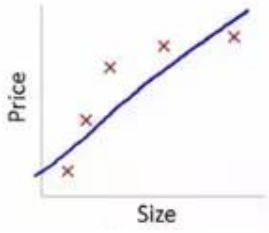
- These are the two main source of errors in ML
- The balance between overfitting and underfitting the training data is a problem known as the bias-variance tradeoff
- The bias/variance measures of what will happen if the model is retrained many times over different subsets of the training data
- Bias: when the learner is consistently learning the same wrong thing
- Variance: when the learner learns random things irrespective to the real signal

STUDENTS-HUB.com Uploaded By: Jibreel Bornat

Bias Error

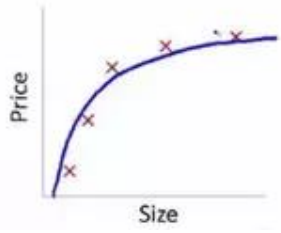
- High bias results in underfitting the data
- High bias means that the algorithm is missing important trends in the features
- High bias algorithms are easy to learn but have lower predictive performance (e.g., linear and non-parametric models)
- Low bias assumptions result in lower assumptions (e.g., non linear and parametric models)

Bias/Variance



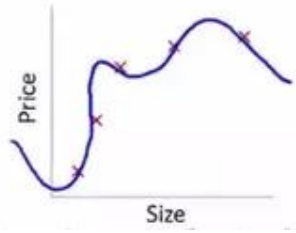
$h(x) = \alpha_0 + \alpha_1 x$

High bias
(underfit)



$h(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2$

Good fit



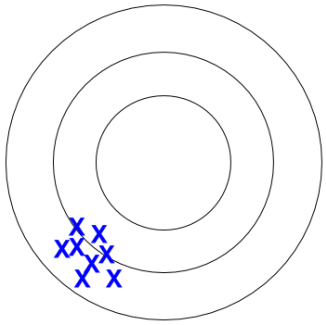
$h(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 + \alpha_4 x^4$

High variance
(overfit)

STUDENTS-HUB.com Uploaded By: Jibreel Bornat

Bias/Variance (2)

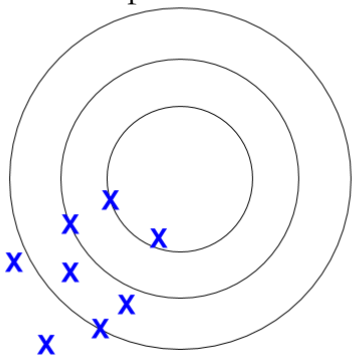
- High bias: when the borderline between two classes isn't clear and the learner is unable to induce it
- The model performs poorly on the training data too (i.e., the linear line does not fit the data very well—underfitting)



Low-Variance, High-Bias

Bias/Variance (3)

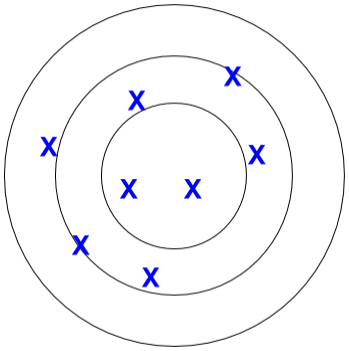
- High variance and high bias
- Decision trees have high variance: decision trees generated from different subsets generated by the same phenomenon



High-Variance, High-Bias

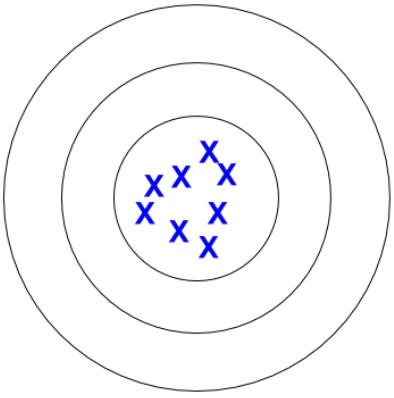
STUDENTS-HUB.com Uploaded By: Jibreel Bornat

Bias/Variance (4)



High-Variance, Low-Bias

Bias/Variance (5)



Low-Variance, Low-Bias

STUDENTS-HUB.com Uploaded By: Jibreel Bornat

Regularization

- Adding a Regularization term to the hypothesis function is another method to avoid overfitting (reducing variance)
- $h(x) = \frac{1}{2m}\sum_{i=1}^{m}(h(x_i) - y_i)^2 + \frac{\lambda}{2m}\sum_{j=1}^{n}\alpha_j^2$
- Minimise the cost function and restrict the parameters not to be too large
- It penalises large parameters. So minimising the cost function includes two terms: the MSE term and the regularisation term

Regularization (2)

- Everytime a parameter is updated to become very large, it increases the value of the cost function
- As a result, it will penalised and updated to a smaller function
- This will favour simple classifiers that have less room to overfit in comparison to classifiers of complex structures

STUDENTS-HUB.com Uploaded By: Jibreel Bornat

Final note

- High variance leads to overfitting
- High bias leads to underfitting
- Avoiding both needs a perfect classifier
- It is hard to know beforehand what is the technique that will do the best
- Statistical tests can be performed (such as Chi-Square) to test if adding a new feature will change the significance of the data or not

STUDENTS-HUB.com Uploaded By: Jibreel Bornat