Artificial Intelligence Masterclass

Supervised Learning Additional Topics

H.M. Samadhi Chathuranga Rathnayake

M.Sc in CS (SU), PG.Dip in SML (Othm), PG.Dip in HRM (LRN), B.Sc (Hons) in IS (UOC), B.Eng (Hons) in SE (LMU), P. Dip EP & SBO (ABE), Dip SE, Dip IT, Dip IT & E-Com, Dip B.Mgt, Dip HRM, Dip Eng

Overfitting & Underfitting

Overfitting refers to a model that models the training data too well. This means that the noise or random fluctuations in the training data is picked up and learned as concepts by the model. The problem is that these concepts do not apply to new data and negatively impact the models ability to generalize.

Underfitting refers to a model that can neither model the training data nor generalize to new data. An underfit machine learning model is not a suitable model and will be obvious as it will have poor performance on the training data.

High-Dimensional Data

Most traditional statistical techniques for regression and classification are intended for the low-dimensional setting in which n is much greater than p.

Let's consider predicting the blood pressure based on age, gender and body mass index (BMI). There are three or four (if an intercept is included) predictors in the model and thousands of patients with the predictor values. Hence n >>p and the problem is low-dimensional.

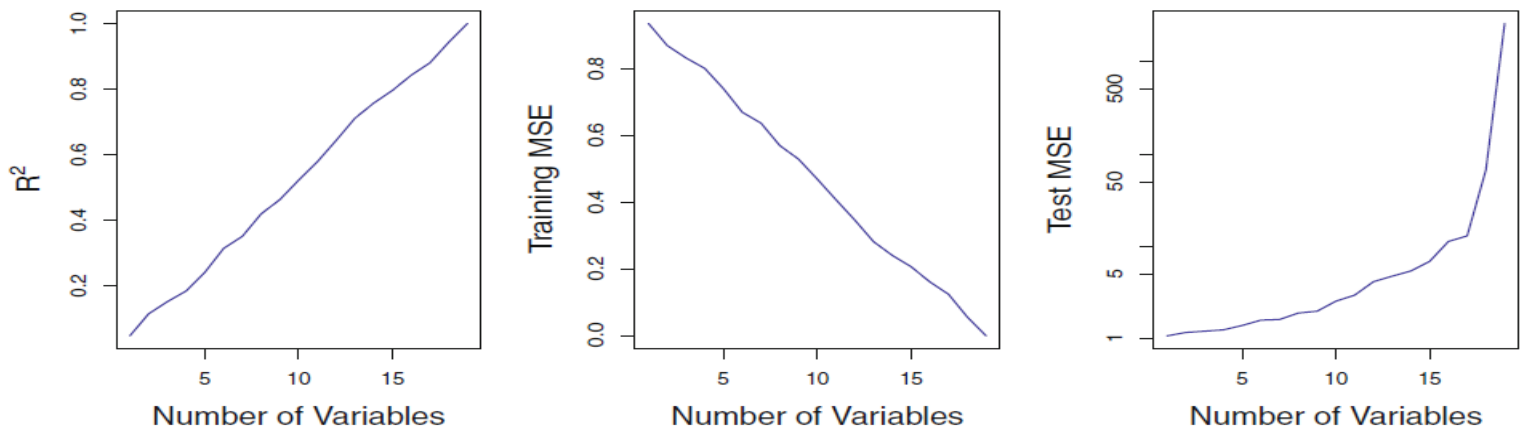
But now it is common to collect a large number of feature measurements (in fields such as Finance, marketing and medicine). While p can be extremely large n is often limited due to cost, sample availability or other considerations.

Datasets containing more features than observations are called as High-dimensional. Classical approaches such as least squares linear regression are not appropriate in this case.

What Goes Wrong in High Dimensions?

It is possible to perfectly fit the training data in the high-dimensional setting, the resulting model will perform extremely poorly on test data.

When p > n or p is close to n, a simple least squares line is too flexible and hence overfits the data. Following plots shows the risk of high dimensions when the number of variables is increased.



Dimensionality reduction

Simplify complex high-dimensional data. Summarize data with a lower dimensional data.

- Given data points in d dimensions
- Convert them to data points in r<d dimensions
- With minimal loss of information

Dimensionality reduction

There are several ways to do dimensionality reduction.

1. Feature Selection
2. Feature Engineering
3. Feature Extraction

Feature Selection

Recursive Feature Elimination

- In Python, above discussed techniques are not available.
- Instead of these techniques, Python provides a better way which is called the Recursive Feature Elimination (RFE).
- The goal of recursive feature elimination (RFE) is to select features by recursively considering smaller and smaller sets of features.
- First, the estimator is trained on the initial set of features and the importance of each feature is obtained either through any specific attribute or callable.
- Then, the least important features are pruned from current set of features.
- That procedure is recursively repeated on the pruned set until the desired number of features to select is eventually reached.

Class Imbalance Problem in Classification

Consider the following example where we have to predict Yes/ No using a model. Consider the response variable data in the training dataset.

Category	Number of Observations
Yes	1500
No	300

Here we can clearly see that the observations in both classes are not balanced. This problem is called the Class Imbalance Problem.

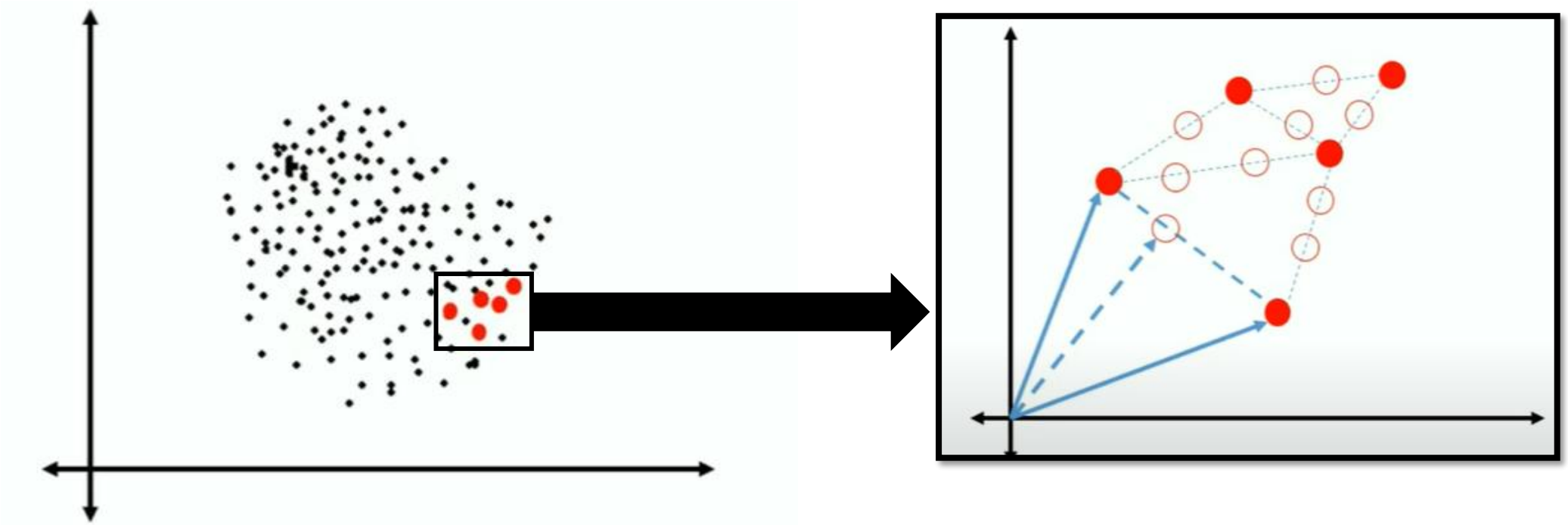
Class Imbalance Problem in Classification

For dealing with this problem, we have several options.

- SMOTE: Synthetic Minority Oversampling Technique
- ADASYN: Adaptive Synthetic Sampling Approach
- Hybridization: SMOTE + Tomek Links
- Hybridization: SMOTE + ENN

SMOTE: Synthetic Minority Oversampling Technique

Consider the following example.



Two Class Classification

No-Fraud → 99.5%
Fraud → 0.5%

ADASYN: Adaptive Synthetic Sampling Approach

Consider the following example.