Machine Learning

Lecture 2: Overfitting

Justin Pearson[1]
mailto:it-1dl034@lists.uu.se

2021

[1] http://user.it.uu.se/~justin/Hugo/courses/machinelearning/

Data

- Machine learning is now successful because it is easy to get hold of data.
- Large data sets can be very effective see[2].
- But large data sets can be computationally expensive. AlphaGo involved training times that took weeks. Can you justify the greenhouse emissions of your machine learning algorithm?

[2] See *The Unreasonable effectiveness of data* (Halevy, Norvig, and Pereira. `https://ieeexplore.ieee.org/abstract/document/4804817`) and *Scaling to very very large corpora for natural language* (Banko and Brill, `https://dl.acm.org/doi/10.3115/1073012.1073017`)

Data

- Good data is sometimes hard to find.
- A lot of the work of a data scientist or machine learning engineer in the real world, is collecting and cleaning data.
- More data is not always the answer. You have to decide what features to use and how to prepossess the data.

Statistics and Machine Learning

The relationship between statistics and machine learning is a bit complicated.

- A statistician is interested in modelling in order to understand the relationship between variables. As in machine learning the models are data driven.
- In machine learning we use data to train an algorithm in order to make predictions. Ultimately how well the algorithm does depends on how accurate predictions it makes.

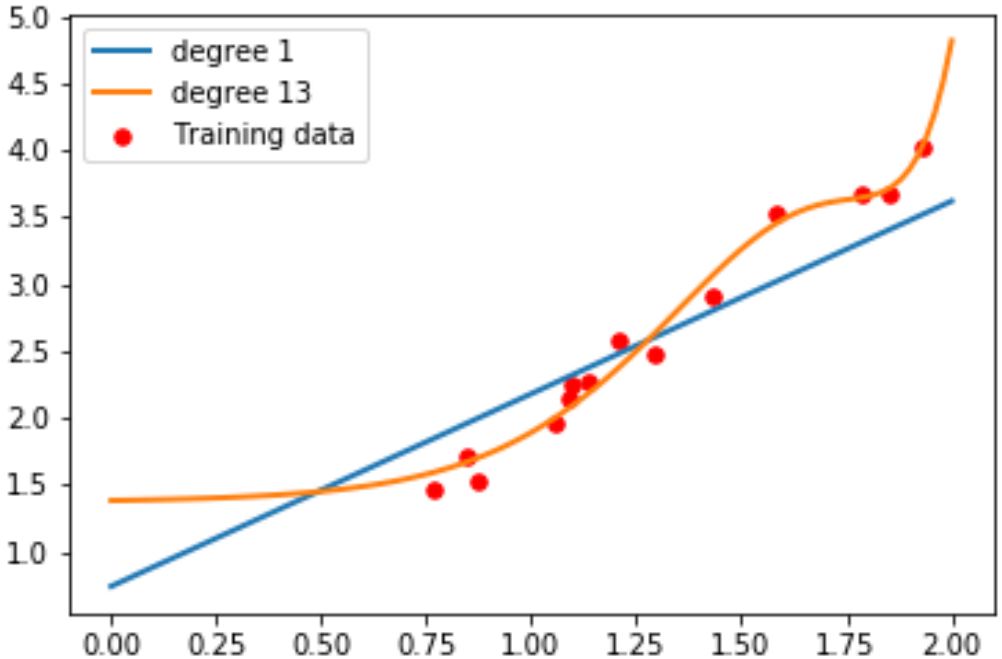
Of course there is a lot of overlap, and the two fields inform each other. A statistician will often make assumptions on the assumed distribution of data more clear than people do in machine learning.

Why might a Machine Learning perform badly?

There are lots of reasons and throughout the course we will try to understand them, but they include:

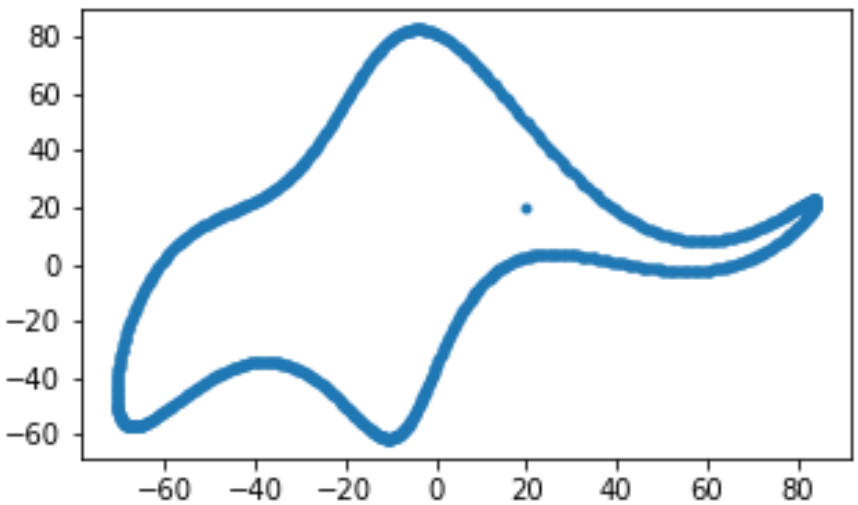
- Not enough data
- Not enough prepossessing of the data
- The wrong machine learning algorithm

Overfitting



Which model is better? Degree 1 or degree 13? If you don't have much data and you can learn complicated models you are in danger of over fitting.

Over fitting



With four parameters I can fit an elephant, and with five I can make him wiggle his trunk.

Attributed to John von Neumann.

Bias

Imagine that alien anthropologists, Zog and Zag, came to Sweden in the 1980s. After some analysis they worked out that type of car that people drove is important for their social standing. They want to analyse the different types of cars that are mod. Zog and Zag sit disguised on Strandvägen. In one hour they observe:

- 100 Volvos
- 75 Saabs

They report back to their science academy that on earth people only drive Volvos annd Saabs.
What is wrong with this picture? Sample bias, making an unwarranted generalisation from the data. They need more data to make predictions about the types of cars available on earth.

How do machine learning algorithms work?

To make things easier we will concentrate on supervised learning.

- The ultimate goal of a machine learning algorithm is to make predictions.
- The algorithm learns a number of parameters given the input data. This is called the *hypothesis*.
- The goal is find a hypothesis that minimises some measure of error, sometimes called the *cost function* or the *loss function*.

Training and Test Sets

One of the most important things to remember in machine learning. You must divide that data into two parts:

Training Set This is the data you use to find the best parameters of the model or hypothesis. Machine learning can be seen as an optimisation problem find the parameters that best explain the data under some error/cost or loss function.

Test Sets The test set is what you use to validate your model. You are interested in the error/cost on this set.

In scikit-learn there is a function `train_test_split` that does randomised splitting of the data. You can decide what percentage of the data is used for training.

Training and Test Set Split

- If your training set is too small then you results might not generalise.
- If your training set is too large then you might get sample bias. You over fit to your data.

Unless you are doing some deep statistical reasoning then you have to use a rule of thumb. A lot of people use the 80/20 or 90/10 split for the training and test split.

Other techniques

- Regularisation.
- Cross validation.
- Hyper-parameter selection methods.
- Picking different learning algorithms.