Machine Learning

Lecture 1

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

[1] https://justinkennethpearson.github.io/

General Information and communication

- Studium will used for announcements and handling assignments.
- Lecture slides can be found at

  `https://intro-ml-1dl034-uu-se.github.io/`

To contact me or the assistants use:

- `it-1dl034@lists.uu.se` to ask questions about the assignments or the projects. The email will go to all the assistants and me.
- `justin.pearson@it.uu.se` to contact me about other things.
- The public versions of the assignments can be found at: `https://github.com/JustinKennethPearson/ml-1dl034-public-assignments`

What is this course about?

- Understanding various machine learning algorithms including:
  - Linear Regression (as a machine learning algorithm), logistic regression,
  - Bayesian classification,
  - support vector machines,
  - decision trees and clustering,
  - Hyper-parameters and how to set them using cross-validation.

What is this course about?

- Understanding various machine learning algorithms including:
  - Linear Regression (as a machine learning algorithm), logistic regression,
  - Bayesian classification,
  - support vector machines,
  - decision trees and clustering,
  - Hyper-parameters and how to set them using cross-validation.
- Common themes behind learning algorithms such optimisation by gradient descent or finding parameters that maximise or minimise some measure of accuracy.

What is this course about?

- Understanding various machine learning algorithms including:
  - Linear Regression (as a machine learning algorithm), logistic regression,
  - Bayesian classification,
  - support vector machines,
  - decision trees and clustering,
  - Hyper-parameters and how to set them using cross-validation.
- Common themes behind learning algorithms such optimisation by gradient descent or finding parameters that maximise or minimise some measure of accuracy.
- Some practical applications in the assignments and the project.

It is important to understand what is going on behind the algorithms so that you know when to apply them.

What is this course about?

- Everything that you should try before you try neural networks or deep learning.

Neural networks are very powerful, but especially with deep learning (lots of data and big neural networks) it is sometimes hard to make them do what you want them to do, and when then do what you want them to do you often do no know why.

Structure of the Course: Overview

- Lectures
- Assignments , done individually.
- Project done is groups of 4-6 (TBA depends on the class size)

Structure of the Course: Assignments

- 4 Practical Python notebooks (solo) where you will learn how to use tools such as scikit-learn, panadas, numpy and Matplotlib.
- 4 more theoretical notebooks (solo) where you will implement or use some machine learning techniques from the lectures.

The Python notebooks are intended to be short. The timebudget for the notebooks and the projects are within the timebudget of 2hp (53.3333 hours). Part of your 3hp (80 hours) spent on the course material is included in learning the material for the more theoretical notebooks.

Structure of the Course: Project

- A project (Groups of 4-6ish). A project that runs throughout the whole course. More information will arrive soon.

Deadlines, hand ins and Help Sessions

For deadline see Studium. You can resubmit as many times during the course up to and including the end of the exam week (Week 11). If you submit an assignment before the deadline, then we will mark it within a week.

When you hand in your notebook either export a PDF[2] or download it as an `html` file and hand that in.

If you look at the timetable in timeedit, you will see a number of sessions marked **Frågestund**. These are help sessions for your assignments. We will not answer questions on the assignments via email. You have to come to these sessions. They are not booked in a lab room but in a classroom. Bring your laptop.
When you start the project you will be assigned an assistant who can help you with the project.

[2] Some students seem to have trouble doing this, but I am not sure why

Textbook

All of the material in the course is standard material, and information can be easily found. The webpage[3]. Has some recommendations. Two books that I recommend are:

- Andriy Burkov's book "The hundred-page machine learning book" Book home page `http://themlbook.com/`. It is available for free, but it is very cheap.
- Chirag Shah "A Hands-On Introduction to Machine Learning". Cambridge University Press. There are some physical copies in the library.

[3] `https://intro-ml-1dl034-uu-se.github.io/resources/books_and_online/index.html`

TODO: Action Points

- Look at `https://intro-ml-1dl034-uu-se.github.io/index.html` and the content.
- Get familiar with Python notebooks. You can either run them locally on your own machine `https://intro-ml-1dl034-uu-se.github.io/resources/python_notebooks_ml/index.html` or you can run them in the cloud `https://www.dataschool.io/cloud-services-for-jupyter-notebook/`.
- When announced. Form a project team. If you want to be randomly assigned then add yourself to the group `I want to be randomly assigned assigned` on Studium. Details on the project will come soon.

Teaching Style and Expectations

Why do I have lectures?

- All of the material in the course is standard material, there are lots of youtube videos, web pages and books that teach the same material.
- The advantage of coming to lectures, is unlike watching a random video on youtube is that you can ask me questions. One way of using the lectures would be to look at the material before the lecture and ask me questions about things that you understand.
- Even if you follow the lecture, you will probably have to look at the text book or other supplementary material in order to understand the material. The lectures will give you a framework for you to digest the material.
- There is a lot of material in the course, following the lectures will give you a schedule of when to learn things and give you the knowledge to be able to do the assignments on time.

Any questions about course mechanics?

- Questions?

What is learning?

From the Oxford English Dictionary

> *Learning: The acquisition of knowledge or skills through study, experience, or being taught.*

What is Machine Learning?

> *Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed. Arthur Samuel (1959)*

Machine learning grew out of Artificial Intelligence (AI), and is now considered a separate field. Machine learning has been around for a very long time.

What is Machine Learning?

> *A Well-posed Learning Problem: A computer program is said to learn from experience E with respect some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.*

This is a very broad definition and could cover everything from spam filters to self driving cars to Skynet.

Machine Learning and Data

Our focus will be statically given data.

- Our machine learning algorithm will be trained on a subset of the data (a training set)
- The performance will be then be measured by how well the algorithm predicts the correct answer on the data.

What we are not going to study

- Reactive agents. We are not going to consider an algorithm in an environment that continually learn.
- No reinforcement learning.
- No biologically inspired algorithms such as genetic algorithms, ant colony optimisation or sheep herd inspired optimisation.
- We will only briefly touch on neural networks, although one of our algorithms such as logistic regression are closely related to perceptrons.

Early History of AI

AI is as old as computer science.

- McCulloch and Pitts : "Logical calculus of the ideas immanent in nervous activity" (1943). First abstract mathematical model of the neuron.
- Alan Turing considered the possibility of thinking machines. Invented the Turing test: "Computing Machinery and Intelligence" (1950) If it walks like a duck and quacks like a duck then it is a duck.
- Checkers (draughts in English, dam på svenska) playing program 1959 by Arthur Samuel.

Motivations for AI

- Understand how people think or how people and animals do things by making models with computers.
- To make intelligent machines.
- To make machines that think.
- To make computers do things without telling them explicitly how to do things.

What does it even mean to make a machine conscious? Are you conscious?

Types of AI

- Symbolic AI: Everything is about logic and symbols build systems that work on these principles. (GOFAI)
- Sub-symbolic AI: Build machines that look like the brain, neural networks or algorithms inspired by statistics.

Some people confuse AI with anything that appears non-trivial or clever. An easy mistake to make, but is google maps using AI when it finds you the quickest route to somewhere?.

Roots of AI

- Statistics : Modelling data production processes as probabilistic systems to make predictions.
- Probability
- Logic : Symbolic AI (still useful).
- Optimisation : Finding the best set of parameter values that achieve a certain goal.
- Control theory : A feedback system that tries to control the environment.

Machine Learning

> *Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.*

Definition taken from IBM.

Data

- Machine learning is now successful because it is easy to get hold of data.
- Large data sets can be very effective see[4]. Also AlphaGo involved training times that took weeks.
- Good data is sometimes hard to find.
- More data is not always the answer. You have to decide what features to use and how to prepossess the data.

[4] 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`)

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.

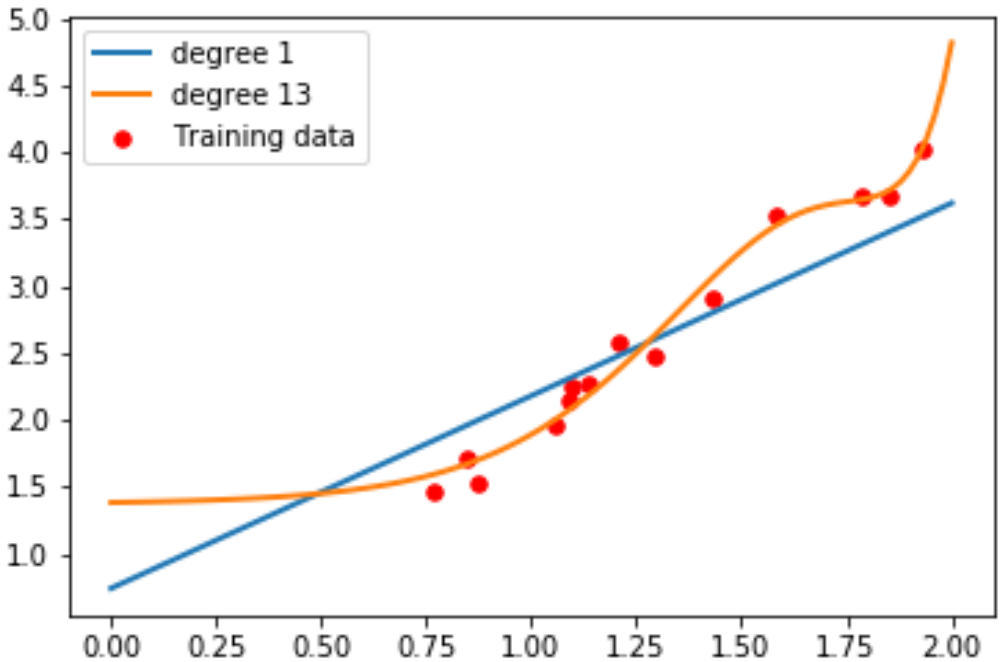
Of course there is a lot of overlap, and the two fields inform each other. A statistician will often make assumptions on the assumed distribution of data more clear than people do in machine learning.

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:

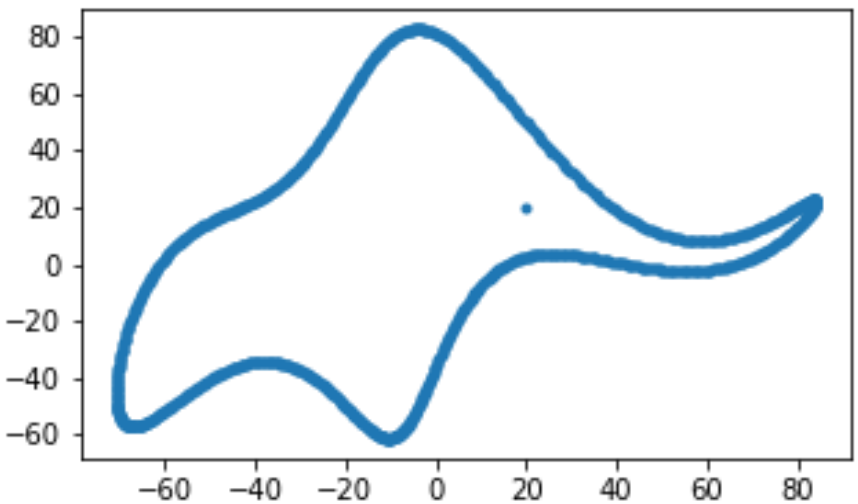
- Not enough data
- Not enough prepossessing of the data
- The wrong machine learning algorithm

Over fitting



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?

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. Statisticians spend a lot of time thinking about this, but for you it is important to think about how you pick your training data.

Supervised and Unsupervised Learning

Two main types of algorithms:

Supervised You are given labelled data. For each data-point you know what the correct prediction should be.

Unsupervised You just have data which is not labelled. This is given to algorithm. The most common algorithms do some sort of clustering, data-points that are similar are grouped together.

Getting good labelled data can sometimes be a problem, especially for deep neural networks where sometimes 100,000s of data-points need to gathered to train the network. Do you want to label 100,000 pictures of cats?

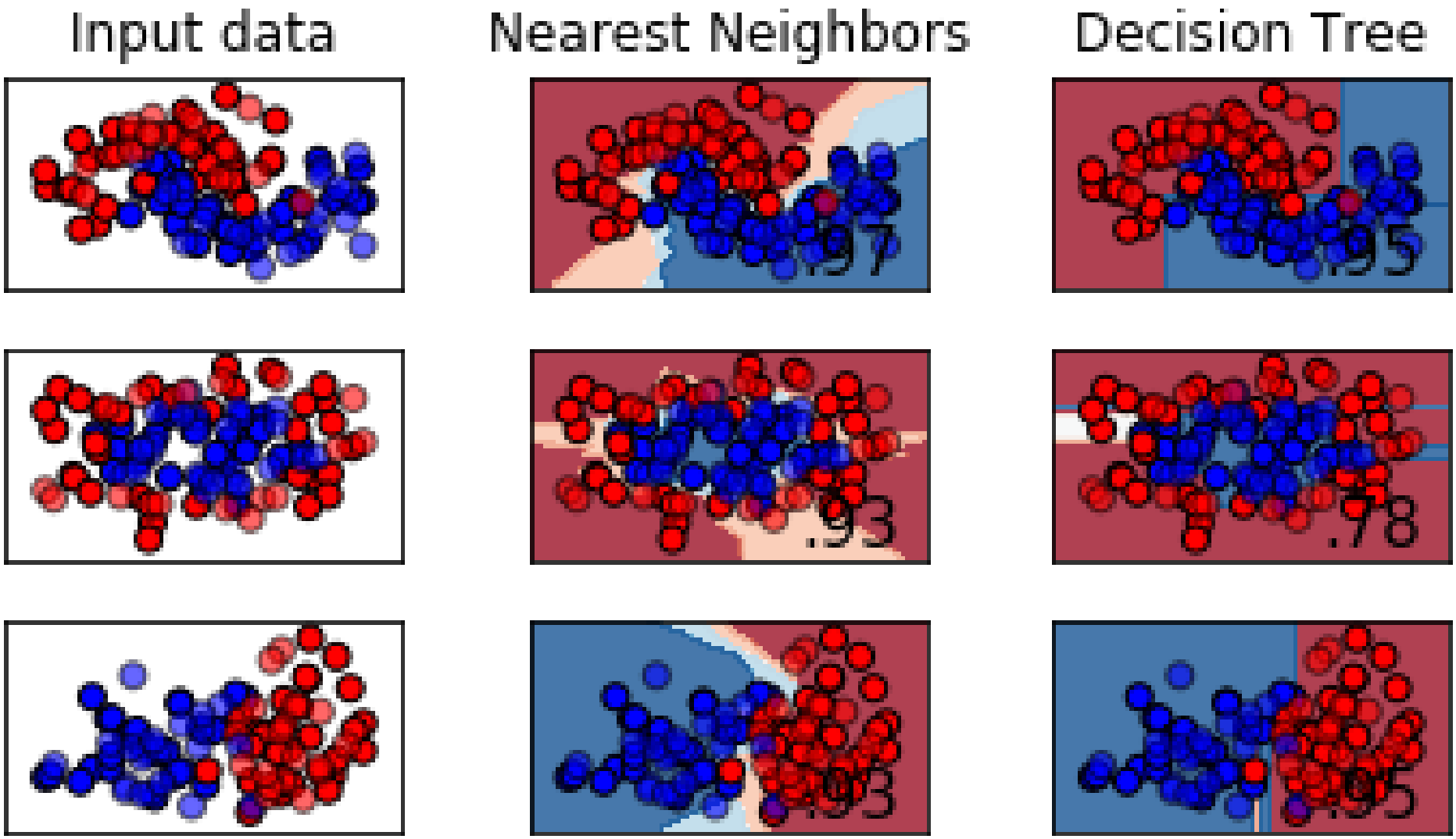
Classification and Regression

Classification Each data point should be put into one of a finite number of classes. For example email should be classified as Spam or Ham. Pictures should be classified into pictures of cats, dogs, or sleeping students.

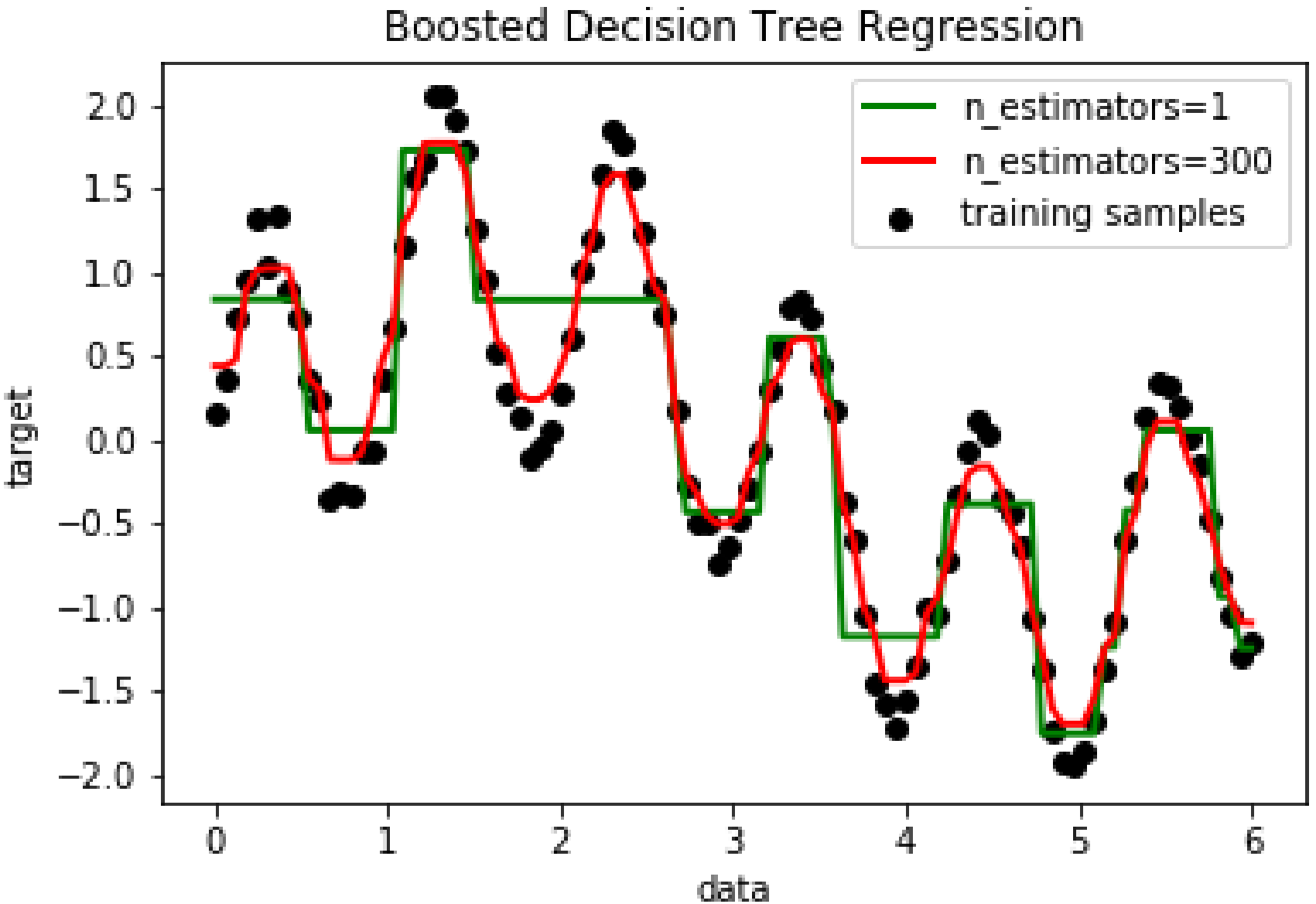
Regresssion Given the data the required prediction is some value. For example predicting house prices from the location and the size of the house.

Even though everything is finite in a computer it is easier mathematically to consider everything as a continuous variable.

Classification



Regression


Boosted Decision Tree Regression
n_estimators=1
n_estimators=300
training samples
target
data
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
0
1
2
3
4
5
6

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*.

Hypothesises

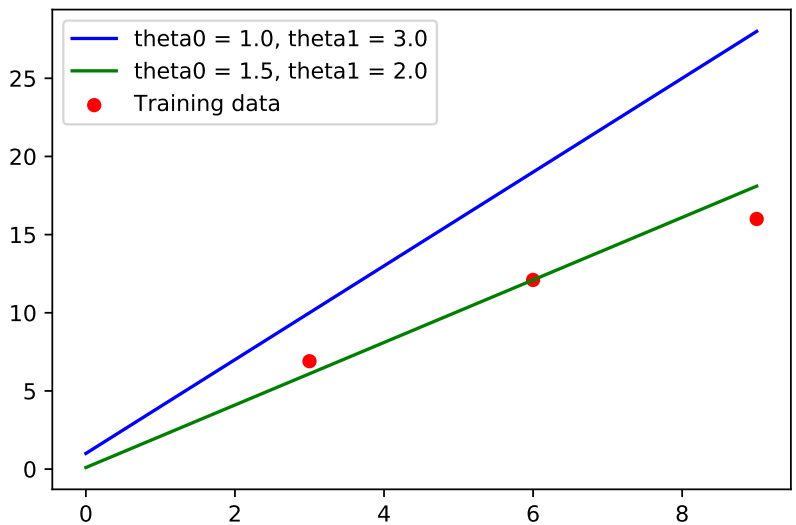
Consider a very simple data set

$$x = (3, 6, 9)$$
$$y = (6.9, 12.1, 16)$$

We want to fit a straight line to the data. Our hypothesises is a function parameterised by θ_0, θ_1

$$h_{\theta_0,\theta_1}(x) = \theta_0 + \theta_1 x$$

Hypothesises



Just looking at the training data we would say that the green line is better. The question is how to we quantify this?

Measuring Error - RMS

Root Mean Squared is a common cost function for regression. In our case given the parameters θ_0, θ_1 the RMS is defined as follows

$$J(\theta_0, \theta_1, x, y) = \frac{1}{2m} \sum_{i=1}^{m} (h_{\theta_0,\theta_1}(x^i) - y^i)^2$$

We assume that we have m data points where x^i represents the ith data point and y^i is the ith value we want to predict. Then $h_{\theta_0,\theta_1}(x^i)$ is the model's prediction given θ_0 and θ_1.
For our data set we get

$$J(1.0, 3.0) = 33.54$$
$$J(1.5, 2.0) = 2.43$$

Obviously the second is a better fit to the data.
Question why $(h_\theta(x) - y)^2$ and not $(h_\theta(x) - y)$ or even $|h_\theta(x) - y|$.

Learning

The general form of regression learning algorithm is as follows:

- Given training data $x = (x^1, \ldots, x^i, \ldots, x^m)$ and $y = (y^1, \ldots, y^i, \ldots, y^m)$
- A set of parameters Θ where each $\theta \in \Theta$ gives rise to a hypothesis function $h_\theta(x)$;
- A loss function $J(\theta, x, y)$ the computes the error or the cost for some hypothesis θ for the given data x,y;
- Find a (the) value θ that minimises J.

How we do this will be the topic of a later lecture.

Classification

Remember two classes of machine learning algorithms: Regression where we want to predict a value (or multiple values) and Classification where we want to predict which class the data belongs to.
Examples of classification problems include:

- Email spam detection, is my current email spam or ham?
- Given some medical information such as the size of a tumour is the tumour cancerous or not.
- Given an image is it a cat, dog or a horse?

Classification — Representation

Typically classes are represented by integer variables.

- $Y = 0$ or $Y = 1$ where $Y = 1$ means that it is spam.
- $Y \in \{0, 1, 2\}$ where 0 means a cat, 1 means a dog and 2 means a horse[5].

Using RMS to measure classification error would not make much sense (Why?)

[5] We'll see later that this is not such a good representation. One-hot encoding is better.

Classification — Measuring Error

One approach is to use probability. We develop an algorithm that gives a probability that the input data is in a particular class.
We then want to maximise the probability that algorithm makes true predictions.

Terminology

Because machine learning has its roots in many different areas of mathematics there are a lot of words (and symbols) that mean the almost same thing:

- Minimising regret, Minimising root mean square error, minimising classification error.
- Parameters of algorithm are the things that learned from the data. In neural networks you would call this the weight space.

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 Sets

- Why is a bad idea to train on the whole data set?

Training and Test Sets

- Why is a bad idea to train on the whole data set?
  - One of the main reasons is that you want to avoid over fitting.
- What is the main factor that might affect what percentage of the data is used for training?

Training and Test Sets

- Why is a bad idea to train on the whole data set?
  - One of the main reasons is that you want to avoid over fitting.
- What is the main factor that might affect what percentage of the data is used for training?
  - How much data you have. If you have too little data then you algorithm will biased towards your data set and your conclusions will not generalise.

Later on we will look at cross validation which splits up the data into multiple training and test sets.