Machine Learning

Introduction

Dariush Hosseini

dariush.hosseini@ucl.ac.uk
Department of Computer Science
University College London

Lecture Overview

1 Lecture Overview

2 Machine Learning: Context & Aims

3 Machine Learning: Paradigms

4 Attributes of Machine Learning Systems

5 Summary

Learning Outcomes for Today’s Lecture

By the end of this lecture you should:

1. Know the **context** and **aims** of machine learning
2. Understand that machine learning algorithms have **paradigmatic motivations**
3. Know the **attributes** of a machine learning system

Lecture Overview

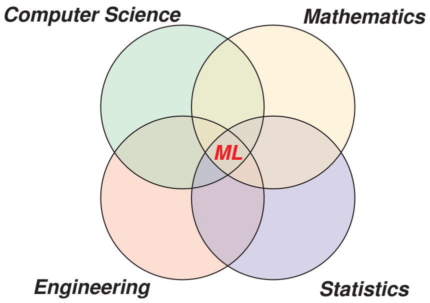
1. Lecture Overview
2. Machine Learning: Context & Aims
3. Machine Learning: Paradigms
4. Attributes of Machine Learning Systems
5. Summary

So what is Machine Learning?

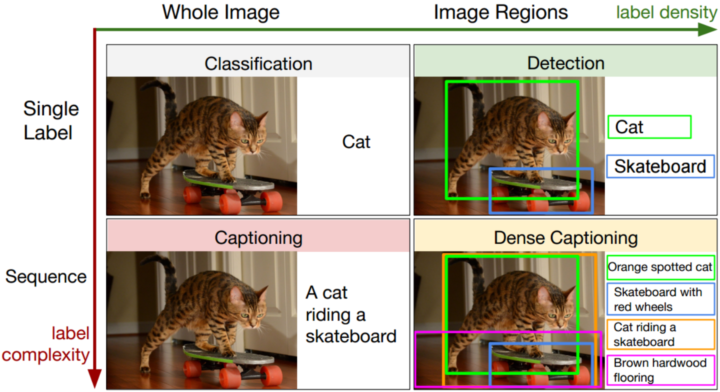
- Allows computers to find hidden patterns...
  - ...without being explicitly programmed to do so.
- Method of data analysis that automates (principled) model building
  - Infers knowledge from data
  - **Generalises** this to unseen data
- Practical yet (often) principled science of **inductive** rather than **deductive** reasoning

Academic Context

- An interdisciplinary field that develops both the **mathematical foundations** and **practical applications** of systems that **learn from data**.



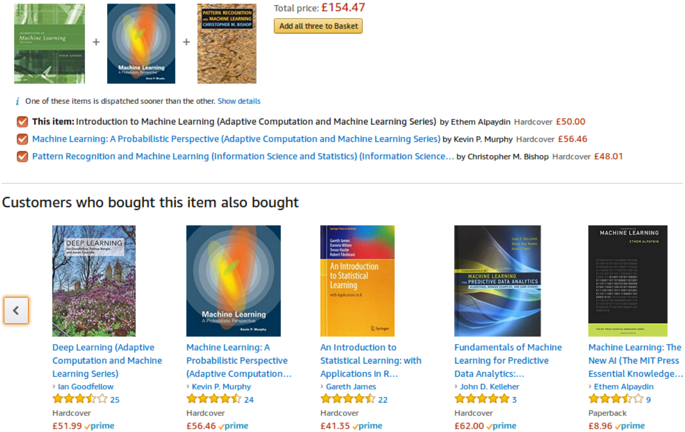
Example: Image Recognition[1]



- Redmon et al, ‘You Only Look Once: Unified, Real-Time Object Detection’ [2015] ▸ Link

[1] Johnson et al, ‘DenseCap: Fully Convolutional Localization Networks for Dense Captioning’ [2015]

Example: Recommendations



Total price: £154.47

Add all three to Basket

i One of these items is dispatched sooner than the other. Show details

☑ **This item:** Introduction to Machine Learning (Adaptive Computation and Machine Learning Series) by Ethem Alpaydin Hardcover £50.00

☑ Machine Learning: A Probabilistic Perspective (Adaptive Computation and Machine Learning Series) by Kevin P. Murphy Hardcover £56.46

☑ Pattern Recognition and Machine Learning (Information Science and Statistics) (Information Science… by Christopher M. Bishop Hardcover £48.01

Customers who bought this item also bought

Deep Learning (Adaptive Computation and Machine Learning Series)
› Ian Goodfellow
★★★½☆ 25
Hardcover
£51.99 ✓prime

Machine Learning: A Probabilistic Perspective (Adaptive Computation…
› Kevin P. Murphy
★★★★½ 24
Hardcover
£56.46 ✓prime

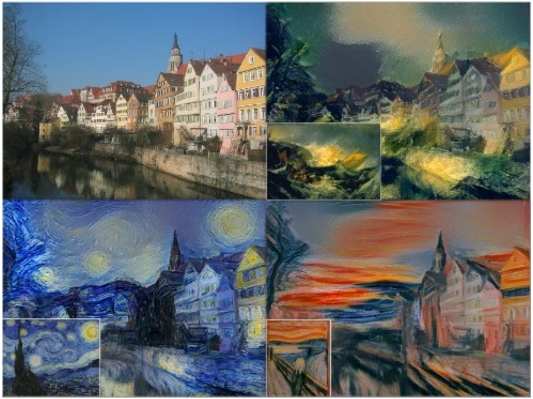
An Introduction to Statistical Learning: with Applications in R…
› Gareth James
★★★★½ 22
Hardcover
£41.35 ✓prime

Fundamentals of Machine Learning for Predictive Data Analytics:…
› John D. Kelleher
★★★★★ 3
Hardcover
£62.00 ✓prime

Machine Learning: The New AI (The MIT Press Essential Knowledge…
› Ethem Alpaydin
★★★½☆ 9
Paperback
£8.96 ✓prime

Machine Learning
Machine Learning A Probabilistic Perspective
PATTERN RECOGNITION AND MACHINE LEARNING CHRISTOPHER M. BISHOP
DEEP LEARNING
An Introduction to Statistical Learning with Applications in R
FUNDAMENTALS OF MACHINE LEARNING FOR PREDICTIVE DATA ANALYTICS
MACHINE LEARNING ETHEM ALPAYDIN

Example: Neural Style[2]



[2] Gatys et al, ‘A Neural Algorithm of Artistic Style’ [2015]

Tasks

- **Supervised Learning**
  - Training data is **labelled** and we seek to predict outputs given inputs
  - **Classification** - where outputs are **discrete**
  - **Regression** - where the outputs are **real-valued**
- **Unsupervised Learning**
  - Training data is **unlabelled** and we seek structure
  - **Dimensionality Reduction**
  - **Clustering**
- **Reinforcement Learning**
  - Exploration / Exploitation
  - **Agent** seeks to learn the **optimal actions** to take based on the outcomes of **past actions**
  - ▸ Link

Lecture Overview

1 Lecture Overview

2 Machine Learning: Context & Aims

3 Machine Learning: Paradigms

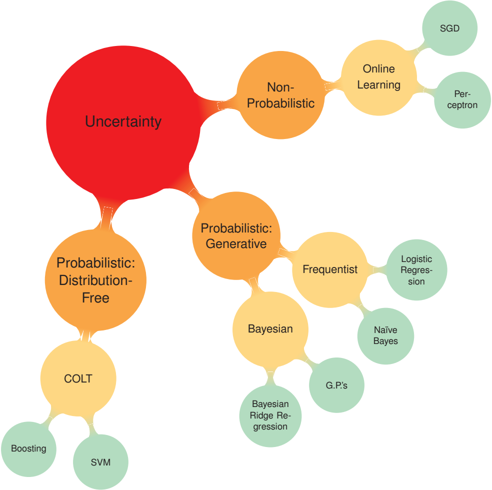
4 Attributes of Machine Learning Systems

5 Summary

A Learning Framework

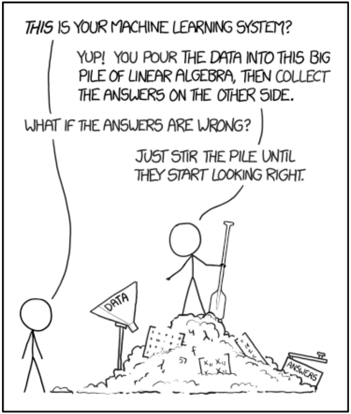
- But Machine Learning is more than a menu of algorithms...
- It also offers a framework for **how** we should create these algorithms...
- It can provide a set of principled approaches to algorithm design, and hence a logic of induction, a way to reason with uncertainty...

Some Approaches to Learning [3]



[3] N.B. This is illustrative - the same algorithm can often be generated by separate approaches.

...And Another[4]


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

[4] https://xkcd.com/

Lecture Overview

1 Lecture Overview

2 Machine Learning: Context & Aims

3 Machine Learning: Paradigms

4 Attributes of Machine Learning Systems

5 Summary

Components of a Machine Learning System

- **Representation**
  - Output of a learning algorithm is a **function** (or **hypothesis** or **model**)
  - We select this function from a set of functions which we provide to the computer
  - This set is called the **function** (or **hypothesis** or **model**) **class** or the **representation** of the learner
- **Evaluation**
  - We provide the computer with an **objective** (or **cost** or **loss**) function with which to distinguish good hypotheses from bad ones
- **Optimisation**
  - We need a procedure for sorting through our hypothesis class, and for selecting the one which produces the **optimal** results upon evaluation

A Motivating Example

A simple motivating example.

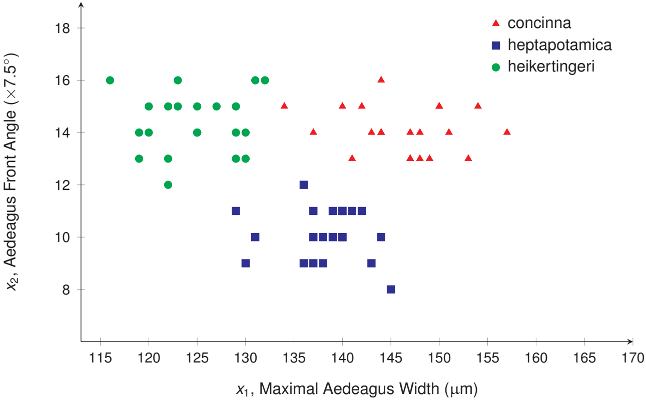
We are interested in automatically identifying the species of a flea based on measurements taken of its body.

This is a **classification** task.

Data

- For each species, we have the measurements of 21 fleas, meaning that we have 63 **data points** in total
- We say that the species of the flea is our **output** or **target** variable
- We take two measurements from each flea and we will refer to these measurements as our **attributes** or **features**, which make up our **input** variable
- Since we have two measurements, we can plot our data in 2-dimensional **feature space**, and colour the data points to indicate their **output class**

Data


▲ concinna
■ heptapotamica
● heikertingeri
x_2, Aedeagus Front Angle (×7.5°)
18
16
14
12
10
8
115
120
125
130
135
140
145
150
155
160
165
170
x_1, Maximal Aedeagus Width (µm)

Training Data

- We apply our learning algorithm to a portion of this data (in this case all of it) in order to **train** or **learn** a function (or model) which we can use to classify fleas
- We call this portion the **training data**, $\mathcal{S}$
- More formally, we might represent this training data as:

$$\mathcal{S} = \{(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), \ldots, (\mathbf{x}^{(n)}, y^{(n)})\}$$

where $\mathbf{x}^{(i)} = \left[x_1^{(i)}, x_2^{(i)}\right]^T \in \mathbb{R}^2$, $y^{(i)} \in \{1, 2, 3\}$, and $n = 63$

- Here $\mathbf{x}^{(i)}$ are outcomes of a random variable, $\mathcal{X}$, while $y^{(i)}$, are outcomes of a random variable $\mathcal{Y}$

Model

- And we represent the function (or model) which we wish to learn as f, which maps from our inputs to the set of outputs, such that:

$$f : \mathbb{R}^2 \rightarrow \{1, 2, 3\}$$

- Clearly we wish to learn a particular f such that:

$$y \approx f(\mathbf{x})$$

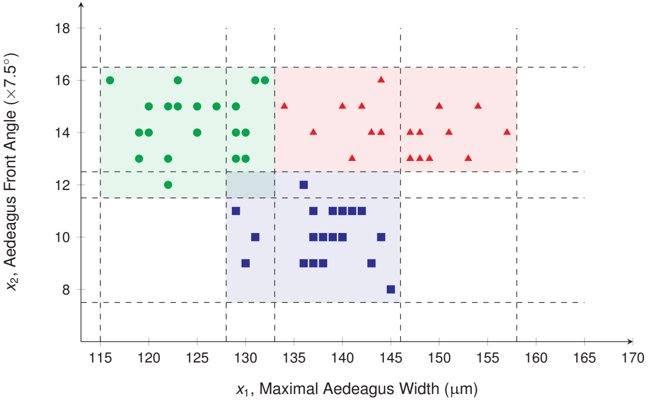
A Simple Model

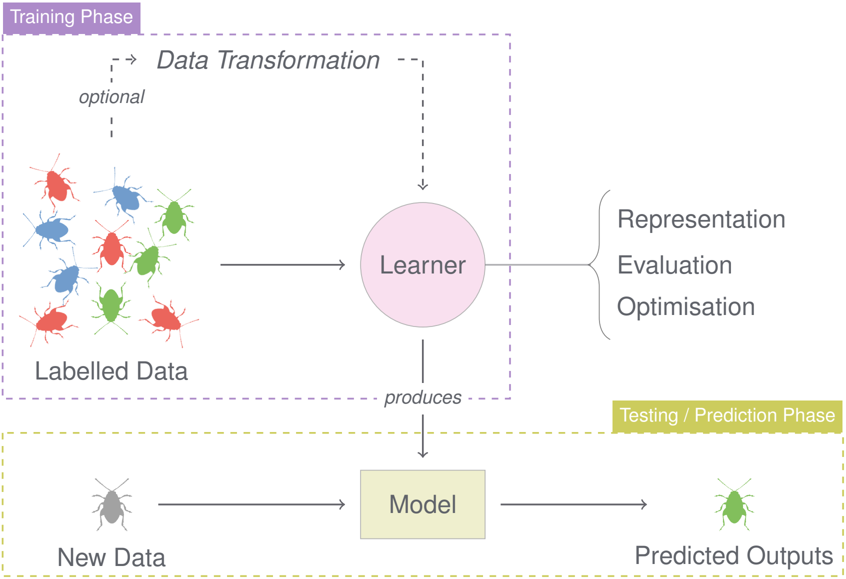
- One of the simplest approaches to perform classification is to produce a set of rules which split our feature space up into distinct regions
- For example, a classification rule for *heptapotamica* might be:

$$\textbf{if}\ \left((x_1 > 128\ \textbf{and}\ x_1 < 146)\ \textbf{and}\ (x_2 > 7.5\ \textbf{and}\ x_2 < 12.5)\right)$$

- So we end up with a set of rules for each species, and we then apply these rules to a new flea to identify its species

A Simple Model





Training Phase

Data Transformation

optional

Labelled Data

Learner

Representation

Evaluation

Optimisation

produces

Testing / Prediction Phase

New Data

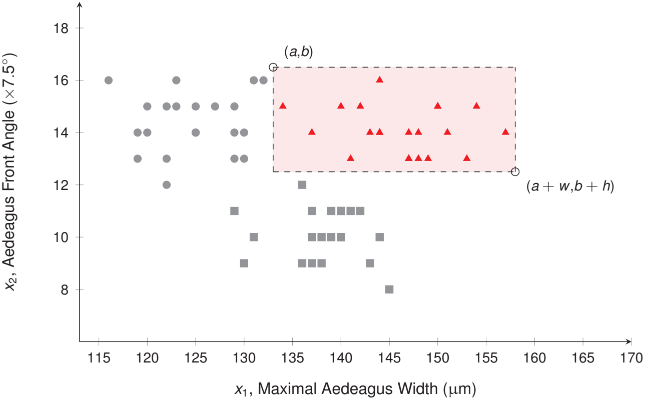
Model

Predicted Outputs

Representation

- Now let's modify the flea classification setting, such that we now seek to classify a flea as *concinna* or not
- Correspondingly our output set is modified such that $y \in \{0, 1\}$
- And a particular rule-based f_θ is characterised by the parameters: $\theta = [a, b, w, h]$ as follows:

Representation



Representation

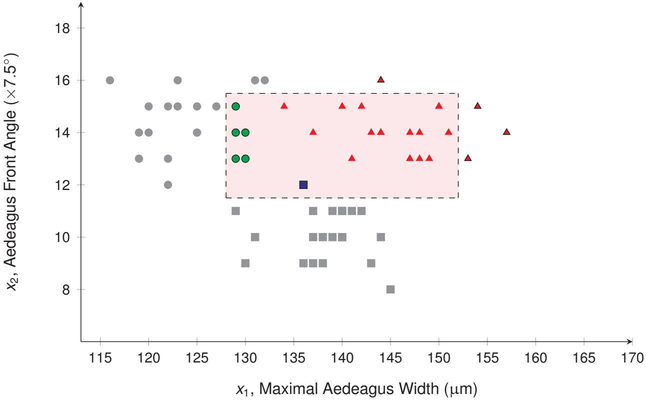
- Since our hypothesis uses four parameters to represent our positive class, we can define the space of all possible combinations of these parameters as:

$$\mathcal{F} = \{f_\theta | \theta = [a, b, w, h], \ \ \forall \, [a, b, w, h] \in \mathbb{R}^4\}$$

- This is our **hypothesis space** and is the space of all such possible rectangles

- The learning process of our classifier is the task of **searching** $\mathcal{F}$ for the **best possible** hypothesis: $f_\theta \in \mathcal{F}$

Error Evaluation



Error Evaluation

- In order to assess how good a particular model is we need a way of evaluating the error.
- Clearly this error should take into account the training data and the function which we are evaluating
- For this we need to first define a **loss measure**, $\mathcal{E}$, a similarity mapping between two inputs, $a \in \mathbb{R}$, $b \in \mathbb{R}$:

$$\mathcal{E} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$$

Error Evaluation

- There are numerous functions that we can use as loss measures. Some common ones include:

  Squared Error: $\mathcal{E}(f_\theta(\mathbf{x}), y) = (f_\theta(\mathbf{x}) - y)^2$

  Absolute Error: $\mathcal{E}(f_\theta(\mathbf{x}), y) = |f_\theta(\mathbf{x}) - y|$

  Misclassification Error: $\mathcal{E}(f_\theta(\mathbf{x}), y) = \mathbb{I}[y \neq f_\theta(\mathbf{x})]$

- The choice of loss measure will depend on the task and on the representation being used.

Error Evaluation

- After we have defined a loss measure, we must then use it to define a **loss function**, L.
- The loss function is a mapping from the loss measure, and some aspect of the data (either in or out of sample), which aggregates the loss measure evaluated at the data points as an expectation.
- The loss function allows us to evaluate how well a particular model performs with respect to a particular loss measure.

Error Evaluation

- Evaluated in sample, for a realised dataset, $\mathcal{S}$, we define the **empirical loss**, $\mathsf{L} : (\mathcal{E}, \mathcal{S}, f_\theta) \mapsto \mathsf{L}(\mathcal{E}, \mathcal{S}, f_\theta) \in \mathbb{R}$, as follows:

$$\mathsf{L}(\mathcal{E}, \mathcal{S}, f_\theta) = \mathbb{E}_{\mathcal{S}}[\mathcal{E}(f_\theta(\mathcal{X}), \mathcal{Y})] = \frac{1}{n}\sum_{i=1}^{n} \mathcal{E}(f_\theta(\mathbf{x}^{(i)}), y^{(i)})$$

- Evaluated out of sample, for a data generating distribution, $\mathcal{D}$, characterised by a **probability distribution function**, $p_{\mathcal{X},\mathcal{Y}}$, from which $\mathcal{S}$ is drawn (i.e. $\mathcal{S} \sim \mathcal{D}^n$), we define the **generalisation loss**, $\mathsf{L} : (\mathcal{E}, \mathcal{D}, f_\theta) \mapsto \mathsf{L}(\mathcal{E}, \mathcal{D}, f_\theta) \in \mathbb{R}$, as follows:

$$\mathsf{L}(\mathcal{E}, \mathcal{D}, f_\theta) = \mathbb{E}_{\mathcal{D}}[\mathcal{E}(f_\theta(\mathcal{X}), \mathcal{Y})] = \iint \mathcal{E}(f_\theta(\mathbf{x}), y) p_{\mathcal{X},\mathcal{Y}}(\mathbf{x}, y)\, d\mathbf{x}\, dy$$

Error Evaluation

- So, for example, for the squared error loss measure:

$$\mathsf{L}(\mathcal{E}, \mathcal{S}, f_\theta) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} (f_\theta(\mathbf{x}^{(i)}) - y^{(i)})^2$$

$$\mathsf{L}(\mathcal{E}, \mathcal{D}, f_\theta) = \mathbb{E}_{\mathcal{D}} \left[\frac{1}{2} (f_\theta(\mathcal{X}) - \mathcal{Y})^2 \right]$$

Optimisation

- Now that we have a way of measuring the loss of a particular model we can proceed to select the optimal one, characterised by θ^*, which exhibits minimal loss.

- Given our **representation** (hypothesis class), we will typically use an **optimisation** approach coupled with a suitable **evaluation** method (loss function) to help us search for the optimal values of θ. We can write this formally as[5]:

$$\theta^* = \underbrace{\underset{\theta}{\operatorname{argmin}}}_{\text{Optimisation}} \; \underbrace{\mathsf{L}\left(\quad \mathcal{E}, \quad \mathcal{S}, \quad \overbrace{f_\theta}^{\text{Representation}} \quad \right)}_{\text{Evaluation}}$$

[5] N.B. Here we use the empirical loss for illustration. We shall look more carefully at the generalisation loss and its approximations in subsequent lectures.

Lecture Overview

1 Lecture Overview

2 Machine Learning: Context & Aims

3 Machine Learning: Paradigms

4 Attributes of Machine Learning Systems

5 Summary

Lecture Summary

1. We can characterise Machine Learning in a number of different ways: academic discipline; task; algorithm; etc.
2. It is important to remember the value of understanding the **theoretical motivation** of a learning algorithm
3. A useful guide in our examination of learning algorithms will be the **Representation + Evaluation + Optimisation** view

In the next lecture we will begin to disucss the mathematical tools that will be useful in understanding the origin of many of the algorithms which we will encounter over this term.