Machine Learning for Humans

My journey from ignorance to Oxford
Aim

Why the hype?
Overview of Machine Learning/Data Science
Some code
Give you an idea if it can help you in your day job
Encourage you to try it out
Some buzz words (for you to sound cool & knowledgeable)
About Me

PHD’s = 0
MSc’s = 0
Degrees = 0
A levels = 0
First programme age 11 (zx81)
Coding Professionally > 25years

Therefore:
I am an Old Dog and
Machine Learning is a new trick
Why the Hype?

The volumes of data are massive
Computer languages have machine learning libraries
GPUs are fast and cheap
Machine learning systems are giving insights traditional systems either can’t do at all or aren’t cost effective
They are now beating real people at games like Go
**What is it?**

**Machine learning** is a type of artificial intelligence (AI) that provides computers with the ability to learn **without** being **explicitly programmed**. Machine learning focuses on the development of computer programs that can teach themselves to grow and change when exposed to **new data**.

[whatis.techtarget.com/definition/machine-learning](whatis.techtarget.com/definition/machine-learning)

**Data science** is an interdisciplinary field about processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured, which is a **continuation** of some of the data analysis fields such as **statistics**, data mining, and **predictive analytics**

[en.wikipedia.org/wiki/Data_science](en.wikipedia.org/wiki/Data_science)
Assumptions

I needed to be at MSc level at least

support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.

\[
\frac{1}{n} \sum_{i=1}^{n} \max\left(0, 1 - y_{i}(w \cdot x_{i} + b)\right) + \lambda \|w\|^2.
\]
Some code – Linear Regression

```r
alligator = data.frame(
  lnLength = c(3.87, ... 3.78),
  lnWeight = c(4.87, ... 4.25) )
model <- lm(lnWeight ~ lnLength, data=alligator)

plot(alligator$lnWeight ~ alligator$lnLength)
abline(model)

predict(model, newdata=data.frame(lnLength=4.0))
  5.248326
predict(model, newdata=data.frame(lnLength=4.2))
  5.934545
```
How does it work?

First there are two or three types

• Supervised learning
• Unsupervised learning
• Reinforced learning

Using Mathematics it is attempting to infer a useful result from previously unseen data.
What is it doing?

Classification – two or multiple
Clustering
Anomaly Detection
Regression
It’s all about the data

Python – Iris sample data

• 150 instances, 50 of each class (Iris - Setosa, Versicolour, Virginica)
• 4 numeric predictive attributes (sepal length & width and petal length & width)
• Code
• Great support to help you create Machine Learning models
• Testing your model with training data leaves you with great results and no confidence
Some of the Lingo

Feature – an attribute e.g. petal length

Vector – all the attributes of a single iris

e.g. [sepal length, sepal width, petal length, length width]
What is it good for?

Predictive Maintenance
Marketing
Finance
Operational Efficiency
Energy Forecasting
Internet of Things
Text and Speech Processing
Image Processing and Computer Vision
Should you use it?

It depends 😊

What problem are you trying to solve?
What level of accuracy do you need?
Is the system CPU or memory constrained?
Is there enough good quality training data? (supervised)
Can data be changed to a suitable format?
Real world Machine Learning - Silos

Problem: find out how full without blowing it up
Level of accuracy: Ask sales or Engineering
System constrained: Yes
Good quality training data: Maybe
Data in suitable format: Yes

martinlishman.com/barn-owl-wireless
Can you use it?

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician. @josh_wills
Can you use it?

It depends 😊

Can you learn to programme in Python/R/C/a JVM language?
Can you learn some basic Mathematics? (the more the better)
Can you prepare data?
Can you learn to use libraries?
Easy Start – Toy Data sets

Python – dataset package
• boston house prices - regression
• iris - classification
• diabetes - regression
• digits - classification
• linnerud - multivariate regression
+ other packages

R - Datasets Package
• 80+
Working with data

Gaps

Data features of differing scales
Some more of the Lingo

Interpolate
- fill in the gaps, lots of ways (better Maths will help here)

Mean, Variance and Standard deviation
- By normalising the data you can give equal weight to features
Having knowledge to improve – Metrics

Confusion Matrix

<table>
<thead>
<tr>
<th></th>
<th>setosa</th>
<th></th>
<th>versicolor</th>
<th></th>
<th>virginica</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>14, 0, 0</td>
<td></td>
<td>0, 14, 4</td>
<td></td>
<td>0, 1, 17</td>
<td></td>
</tr>
</tbody>
</table>

Confusion Matrix virginica – [setosa = 0, versicolor = 1, virginica = 17]
# More Metrics

## Classification report

<table>
<thead>
<tr>
<th></th>
<th>precision</th>
<th>recall</th>
<th>f1-score</th>
<th>support</th>
</tr>
</thead>
<tbody>
<tr>
<td>setosa</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>14</td>
</tr>
<tr>
<td>versicolor</td>
<td>0.93</td>
<td>0.78</td>
<td>0.85</td>
<td>18</td>
</tr>
<tr>
<td>virginica</td>
<td>0.81</td>
<td>0.94</td>
<td>0.87</td>
<td>18</td>
</tr>
<tr>
<td>avg / total</td>
<td>0.91</td>
<td>0.90</td>
<td>0.90</td>
<td>50</td>
</tr>
</tbody>
</table>

Precision virginica – correct 17 predicted 21: 17/21 = 0.81  
Recall virginica – correct 17 actual 18: 17/18 = 0.94  
F1-source – mean of precision and recall
Working with text

Text != Numeric

For machine learning Text -> numerical feature vectors.

• Each word is assigned an integer identifier
Real world Machine Learning - Text

Problem: Feature extraction from documents
Level of accuracy: Very high
System constrained: No
Good quality training data: Getting there
Data in suitable format: Yes
Working with text

Code (if there is time)

Text processing

• Vectorisation
• Text feature extraction
• Term Frequencies times Inverse Document Frequency (tf-idf)
• Stop words
What we have covered

What is Machine Learning
Some of the ways Machine Learning can be used
Some code – using and reviewing results
Some buzz words (for you to sound cool & knowledgeable)
Books

www.manning.com/books/introducing-data-science

Questions? & Links

Information:
www.analyticsvidhya.com/blog/2015/08/common-machine-learning-algorithms
www.analyticsvidhya.com/blog/2015/09/full-cheatsheet-machine-learning-algorithms

Start coding:
www.continuum.io/anaconda-overview
www.r-project.org
www.rstudio.com/home

Email: peter@catalystcomputing.co.uk
Web: catalystcomputing.co.uk
Blog: catalystcomputing.co.uk/peter-marriott
Twitter: @peter_marriott
GitHub: github.com/catalystcomputing