

# Sketch of Chapter 1

DuLi

2019 年 3 月 13 日

## 目录

<b>1</b>	<b>Definition of Machine Learning</b>	<b>3</b>
1.1	Arthur Samuel's Definition . . . . .	3
1.2	Tom Michell's definition . . . . .	3
<b>2</b>	<b>Why use Machine Learning</b>	<b>4</b>
2.1	Types of Machine Learning Systems . . . . .	4
<b>3</b>	<b>Supervised/Unsupervised Learning</b>	<b>5</b>
3.1	Supervised Learning . . . . .	5
3.2	Unsupervised Learning . . . . .	5
3.3	Semisupervised Learning . . . . .	6
3.4	Reinforcement Learning . . . . .	6
<b>4</b>	<b>Batch and Online Learning</b>	<b>6</b>
4.1	Batch learning . . . . .	7
4.2	Online learning . . . . .	7
<b>5</b>	<b>Instance-based versus Model-based learning</b>	<b>7</b>
5.1	Instance-based learning . . . . .	7
5.2	Model-based learning . . . . .	7
5.3	Training and running a linear model using Scikit-Learn . . . . .	7

## 1 Definition of Machine Learning

### 1.1 Arthur Samuel's Definition

*Machine Learning is a field of study that gives computers the ability to learn without being explicitly programmed.*

这个定义看来是非常general了，基本上是从字面上解释了这个词组。稍微提一下Arthur Samuel这个人，直接把Wikipedia的简介搬过来。大体就是上古大神的意思了。



图 1: Arthur Samuel

Arthur Lee Samuel (December 5, 1901 – July 29, 1990) was an American pioneer in the field of computer gaming and artificial intelligence. He coined the term "machine learning" in 1959. The Samuel Checkers-playing Program was among the world's first successful self-learning programs, and as such a very early demonstration of the fundamental concept of artificial intelligence (AI). He was also a senior member in the TeX community who devoted much time giving personal attention to the needs of users and wrote an early TeX manual in 1983.

### 1.2 Tom Michell's definition

A computer program is said to learn from experience  $E$  with respect to some task  $T$  and some performance measure  $P$ , if its performance on  $T$ , as measured by  $P$ , improves with experience  $E$ .

乍一看这个写的比较拗口，但实际上也就是说通过对experience E的学习，使原有task T的performance P有了提高。对于Tom Michell的了解大概是因为那本机器学习的教材，薄薄一本，当时应该是看过同学的，没有看太多，也不好评价。这个定义就显得更像在描述一件事情而非定义。本书接下来用一个简单的例子来说明了一下，那就是spam filter，并将Tom Michell的定义套用做了讲解。



图 2: Tom Michell

## 2 Why use Machine Learning

依旧以写一个spam filter为例，当使用传统的编程技术时，需要一个长长的规则清单，并且这个规则是难以把握且复杂的。相反，当使用Machine Learning Techniques时，可以让机器自主的学习知识，学习垃圾邮件中的词频，这显然更加简洁高效。To summarize, Machine Learning is great for:

1. Problems require a lot of hand-tuning or long list of rules.
2. Complex problems is no good solution using traditional approach.
3. Fluctuating enviornment
4. Getting insights about complex problems and large amount of data

### 2.1 Types of Machine Learning Systems

- Whether or not they are trained with human supervision(Supervised Unsupervised Semisupervised Reinforcement Learning 监督非监督半监督强化学习)

- Whether or not they can learn incrementally(Online versus Batch learning 在线学习 VS 批量学习)
- Whether work by comparing new data to known data or instead detect patterns in the training data and build a predictive model(instance-based versus model-based learning 基于模型 vs 基于实例)

## 3 Supervised/Unsupervised Learning

### 3.1 Supervised Learning

The training data you feed to the algorithm includes the desired solutions, called *labels*. Here are some of the most important supervised learning algorithms(在后文中都有体现):

- k-Nearest Neighbors
- Linear Regression
- Logistic Regression
- Support Vector Machines
- Decision Trees and Random Forests
- Neural networks

### 3.2 Unsupervised Learning

The training data is unlabeled. Some important unsupervised learning algorithms:

- Clustering
  - k-Means
  - Hierarchical Cluster Analysis(HCA)

- Expectation Maximization
- Visualization and dimensionality reduction
- Principle Component Analysis(PCA)
- Kernel PCA
- Locally-Linear Embedding(LLE)
- t-distributed Stochastic Neighbor Embedding(t-SNE)
- Association rule learning
- Apriori
- Eclat

### 3.3 Semisupervised Learning

Algorithms can deal with partially labeled training data(usually a lot of unlabeled data and a little bit of labeled data) is called semisupervised learning

### 3.4 Reinforcement Learning

Reinforcement learning (RL) is an area of machine learning concerned with how software agents ought to take actions in an environment so as to maximize some notion of cumulative reward.

## 4 Batch and Online Learning

Whether the system can learn incrementally from a stream of income data or not is the criterion to classify batch learning and online learning.

### 4.1 Batch learning

First the system is trained, and then it is launched into production and runs without learning anymore, it just apply what it has learned, it also called offline learning.

### 4.2 Online learning

In online learning, you train the system incrementally by feeding it data instances sequentially, either individually or by small groups called mini-batches.

## 5 Instance-based versus Model-based learning

Two generalization methods: Instance-based and model-based learning.

### 5.1 Instance-based learning


The system learns the examples by heart, then generalizes to new cases using a similarity measure.

### 5.2 Model-based learning

Building a model of the examples, then use the model to make predications. The book using a linear regression problem as an example.

### 5.3 Training and running a linear model using Scikit-Learn

在对数据进行处理前，需要先获得数据，本书都使用真实数据来进行处理。第一个程序所用的程序来自OECD(Organisation for Economic Co-Operation and Development)。本书给出的是一个短链接<https://stats.oecd.org/index.aspx?DataSetCode=BLI>可以在网站直接下载每一年的数据。具体的情况大家可以自行下载或者看下图表头。

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT  **OECD.Stat**

Click here to Login | Contact us | User Guide | Home  
English | Français

Search  [Getting Started](#)

**Better Life Index - Edition 2017**

Customise Export Draw chart My Queries

Measure	Value	
Inequality	Total	
Indicator	Housing	Income
Unit	Dwellings without basic facilities	Housing expenditure Rooms per person Householder net adjusted disposable income Household net financial wealth Labour market insecurity Employment rate Long-term unemployment rate Personal earnings Quality of support network Educational attainment Student skills Years in education Air pollution Water quality Stakeholder engagement for developing regulations Life expectancy Self-reported health Life satisfaction Feeling safe walking alone at night Homeless rate Employees working very long hours Time devoted to leisure and personal care
Unit	Percentage	Percentage Ratio US Dollar US Dollar Percentage Percentage Percentage US Dollar Percentage Percentage Average score Years Micrograms per cubic metre Percentage Average score Percentage Years Percentage Average score Percentage Ratio Percentage Hours
Country		
Australia	(E) 1.1	20 (E) 2.3 33 417 57 462 4.3 72 1.36 52 063 94 80 502 21.2 5 92 2.7 91 82.5 85 7.3 63.6 1 13.2 14
Austria	1	21 1.6 32 544 59 674 2.7 72 1.94 48 295 92 85 492 17.1 16 93 1.3 75 81.3 70 7 80.7 0.4 6.78 14
Belgium	2.3	21 2.2 29 968 104 084 4.8 62 3.88 49 587 92 75 503 18.2 15 84 2.2 89 81.1 75 6.9 70.7 1 4.31 15
Canada	0.2	22 2.5 29 850 85 758 3.9 73 0.81 48 403 93 91 523 16.7 7 91 3 68 81.5 88 7.3 80.9 1.4 3.73 14
Chile	9.4	18 1.9 (E) 16 588 21 409 8.1 62 (E) 2.02 28 434 84 65 443 17.3 16 69 1.5 49 79.1 57 6.7 51.1 4.5 10.06 (E) 14
Czech Republic	0.6	24 1.4 21 103 24 258 1.8 72 1.71 23 722 89 93 491 17.3 20 87 2.6 59 78.7 61 6.6 68.3 0.8 5.77 (E) 15.0
Denmark	0.6	24 1.9 28 950 73 543 2.3 75 1.35 52 580 95 81 504 19.7 9 94 2.1 86 80.8 72 7.5 83 0.7 2.2 15
Estonia	6.9	18 1.6 18 665 16 967 4 72 2.13 23 621 90 89 524 15.8 8 82 2.8 64 77.7 51 5.6 67.2 3.1 2.69 14
Finland	0.5	23 1.9 29 374 27 972 2.7 69 2.33 42 127 95 88 523 19.8 6 94 2.3 69 81.6 70 7.5 82.9 1.4 3.91 15
France	0.5	21 1.8 31 137 59 479 5 65 4.33 42 992 88 78 496 16.5 13 82 2.1 75 82.4 68 6.4 69.6 0.6 7.76 16
Germany	0.1	20 1.8 33 652 57 356 2 75 1.68 46 389 92 86 508 18.3 14 93 2.1 72 80.7 65 7 75.9 0.4 4.6 15
Greece	0.5	24 1.2 17 002 18 117 17.4 52 18.95 25 124 82 72 458 16.9 18 69 1.9 64 81.1 74 5.2 61.8 1 7.3 (E) 14.6
Hungary	4.3	18 1.2 16 821 23 289 4.8 67 2.42 21 711 84 83 474 16.6 19 76 1.2 62 75.7 56 5.3 50.7 1.2 3.05 (E) 15.0
Iceland	0	24 1.6 (E) 30 453 64 398 2.8 86 0.26 55 984 98 78 481 19.3 3 99 1.6 79 82.5 76 7.5 87 0.9 (E) 15.06 (E) 14.1
Ireland	0.1	21 2.1 25 439 43 493 2.1 65 4.72 51 681 96 80 509 18.7 7 82 0.8 65 81.5 82 7 75.5 0.6 4.66 (E) 15.2

Legend:  
E: Estimated value

Data extracted on 13 Mar 2019 13:41 UTC (GMT) from OECD.Stat

Powered by Stat technology | © OECD. All rights reserved. Terms & Conditions | Privacy Policy

Click here to Login | Contact us | Site Map | User Guide | Home

图 3: OEDC 网站截图

```

1 import matplotlib
import matplotlib as plt
3 import numpy as np
import pandas as pd
5 import sklearn

```