Sketch of Chapter 1

DuLi

2019 年 3 月 14 日

目录

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个事情而非定义。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 desire solutions,called *labels*.Here are some 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)

- Expection Maximization

• Visualization and dimensionality reduction

- Principle Component Analysis(PCA)

- Kernel PCA

- Locally-Linear Embedding(LLE)

- t-distributed Stochastic Neighbor Embedding(t-SNE)

• Associaton 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 anymo·re,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 similiarity 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。第一个程序所用的数据来自OECD(Organisation for Economic Co-Operation and Development)和IMF(International Monetary Foundation)，分别是life satisfaction 和 GDP per captia。本书给出的是一个短链接，这里直接贴出原始链接https://stats.oecd.org/index.aspx?DataSetCode=BLI和 https://

www.imf.org/external/pubs/ft/weo/2016/01/weodata/weorept.aspx?pr.x=32&pr.y=8&sy=2015&ey=2015&scsm=1&ssd=1&sort=country&ds=.&br=1&c=512,668,914,672,612,946,614,137,311,962,213,674,911,676,193,548,122,556,912,678,313,181,419,867,513,682,316,684,913,273,124,868,339,921,638,948,514,943,218,686,963,688,616,518,223,728,516,558,918,138,748,196,618,278,624,692,522,694,622,142,156,449,626,564,628,565,228,283,924,853,233,288,632,293,636,566,634,964,238,182,662,453,960,968,423,922,935,714,128,862,611,135,321,716,243,456,248,722,469,942,253,718,642,724,643,576,939,936,644,961,819,813,172,199,132,733,646,184,648,524,915,361,134,362,652,364,174,732,328,366,258,734,656,144,654,146,336,463,263,528,268,923,532,738,944,578,176,537,534,742,536,866,429,369,433,744,178,186,436,925,136,869,343,746,158,926,439,466,916,112,664,111,826,298,542,927,967,846,443,299,917,582,544,474,941,754,446,698,666&s=NGDPDPC&grp=0&a#download 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。唯一要注意的是，书上所使用的数据集是2015年的数据。只取life satisfaction和GDP per Captia的数据。

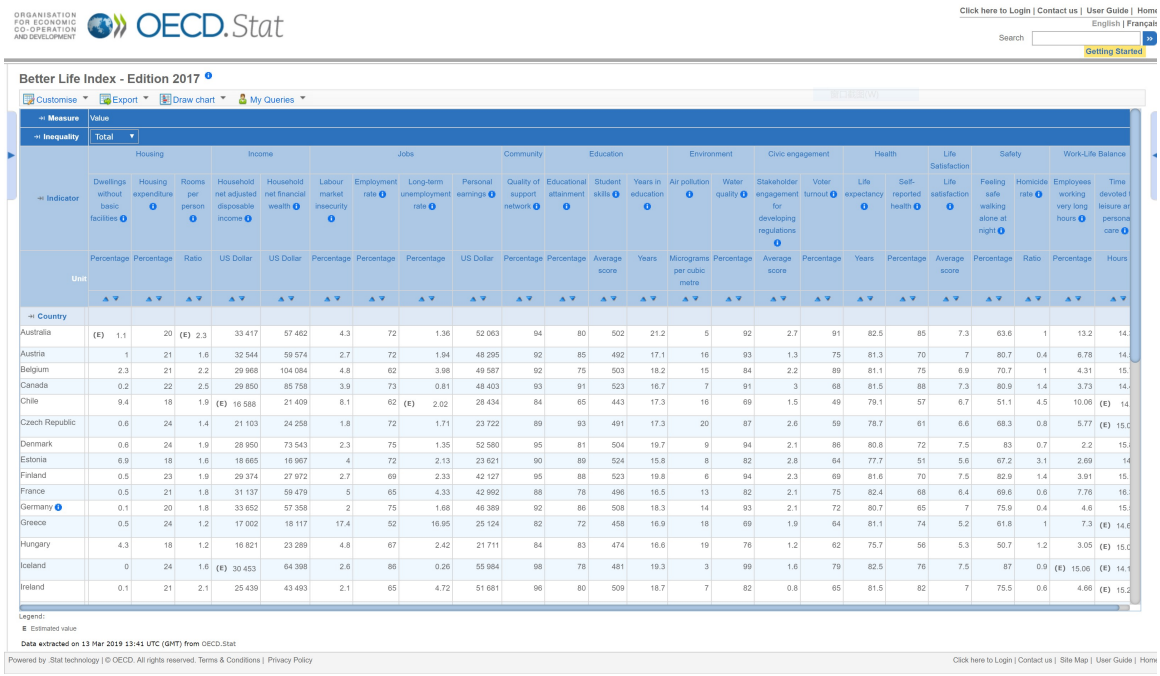
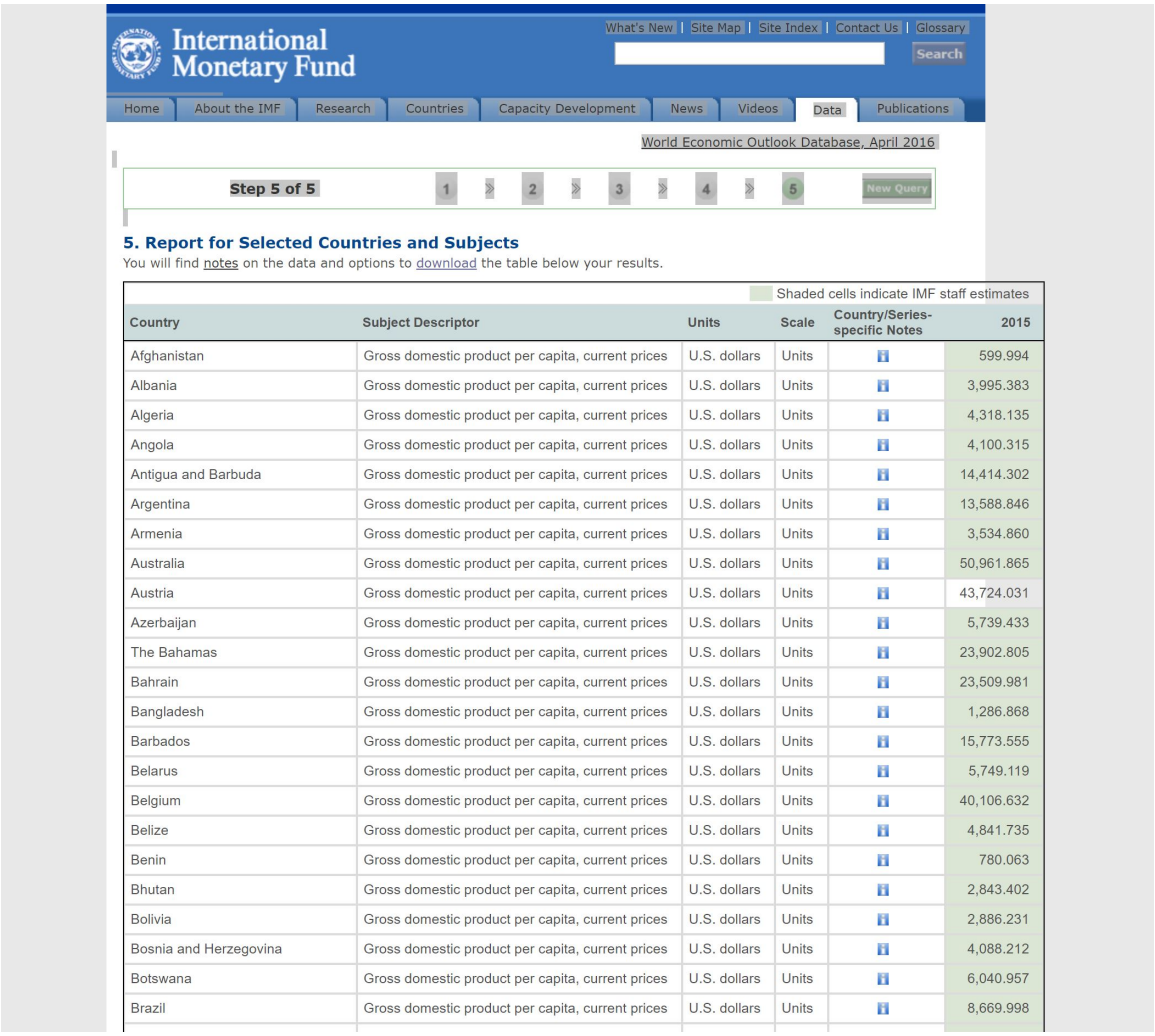


图 3: OEDC 网站截图



International Monetary Fund

What's New | Site Map | Site Index | Contact Us | Glossary

Search

Home | About the IMF | Research | Countries | Capacity Development | News | Videos | Data | Publications

World Economic Outlook Database, April 2016

Step 5 of 5 1 » 2 » 3 » 4 » 5 New Query

5. Report for Selected Countries and Subjects

You will find notes on the data and options to download the table below your results.

Shaded cells indicate IMF staff estimates

Country	Subject Descriptor	Units	Scale	Country/Series-specific Notes	2015
Afghanistan	Gross domestic product per capita, current prices	U.S. dollars	Units		599.994
Albania	Gross domestic product per capita, current prices	U.S. dollars	Units		3,995.383
Algeria	Gross domestic product per capita, current prices	U.S. dollars	Units		4,318.135
Angola	Gross domestic product per capita, current prices	U.S. dollars	Units		4,100.315
Antigua and Barbuda	Gross domestic product per capita, current prices	U.S. dollars	Units		14,414.302
Argentina	Gross domestic product per capita, current prices	U.S. dollars	Units		13,588.846
Armenia	Gross domestic product per capita, current prices	U.S. dollars	Units		3,534.860
Australia	Gross domestic product per capita, current prices	U.S. dollars	Units		50,961.865
Austria	Gross domestic product per capita, current prices	U.S. dollars	Units		43,724.031
Azerbaijan	Gross domestic product per capita, current prices	U.S. dollars	Units		5,739.433
The Bahamas	Gross domestic product per capita, current prices	U.S. dollars	Units		23,902.805
Bahrain	Gross domestic product per capita, current prices	U.S. dollars	Units		23,509.981
Bangladesh	Gross domestic product per capita, current prices	U.S. dollars	Units		1,286.868
Barbados	Gross domestic product per capita, current prices	U.S. dollars	Units		15,773.555
Belarus	Gross domestic product per capita, current prices	U.S. dollars	Units		5,749.119
Belgium	Gross domestic product per capita, current prices	U.S. dollars	Units		40,106.632
Belize	Gross domestic product per capita, current prices	U.S. dollars	Units		4,841.735
Benin	Gross domestic product per capita, current prices	U.S. dollars	Units		780.063
Bhutan	Gross domestic product per capita, current prices	U.S. dollars	Units		2,843.402
Bolivia	Gross domestic product per capita, current prices	U.S. dollars	Units		2,886.231
Bosnia and Herzegovina	Gross domestic product per capita, current prices	U.S. dollars	Units		4,088.212
Botswana	Gross domestic product per capita, current prices	U.S. dollars	Units		6,040.957
Brazil	Gross domestic product per capita, current prices	U.S. dollars	Units		8,669.998

图 4: IMF 网站截图

书上并没有给出merge这两部分数据的代码(The code assume that prepare_country_stats() is already defined:it merges the GDP and life satisfaction data into a single pandas dataframe),接下来我们就首先define一下prepare_country_ function.

```
1
import pandas as pd
```