

Sketch of Chapter 1

DuLi

2019 年 3 月 15 日

目录

1	Definition of Machine Learning	3
1.1	Arthur Samuel's Definition	3
1.2	Tom Michell's definition	3
2	Why use Machine Learning	4
2.1	Types of Machine Learning Systems	4
3	Supervised/Unsupervised Learning	5
3.1	Supervised Learning	5
3.2	Unsupervised Learning	5
3.3	Semisupervised Learning	6
3.4	Reinforcement Learning	6
4	Batch and Online Learning	6
4.1	Batch learning	7
4.2	Online learning	7
5	Instance-based versus Model-based learning	7
5.1	Instance-based learning	7
5.2	Model-based learning	7
5.3	Training and running a linear model using Scikit-Learn	7
6	Main challenge of Machine Learning	10
6.1	Insufficient Quantity of Training Data	10
6.2	Nonrepresentative Training Data	10
6.3	Poor-Quality Data	10
6.4	Irrelevant Features	10
6.5	Overfitting the training data	10
6.6	Underfitting the training data	11
7	Testing and Validating	11
8	Summarization	11

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 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 applies what it has learned, it is 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 predictions. The book uses 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 capita。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://data.imf.org/?sk=64&sk=64>

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 Capita的数据。

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	Value
Country	Value
Australia	(€) 1.1 20 (€) 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 59 574 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 3.98 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 (€) 16 588 21 409 8.1 62 (€) 2.02 28 434 84 65 443 17.3 16 89 1.5 49 79.1 57 6.7 51.1 4.5 10.06 (€) 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 (€) 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 96 80.8 72 7.5 83 0.7 2.2 15.1
Estonia	6.9 18 1.6 19 860 19 967 4 72 2.13 23 621 90 89 524 15.6 8 82 2.8 94 77.7 51 5.6 67.2 3.1 2.89 14.
Finland	0.6 23 1.8 29 374 27 972 2.7 69 2.33 42 127 95 88 523 19.8 6 94 2.3 89 81.6 70 7.6 82.9 1.4 3.91 15.
France	0.5 21 1.6 31 137 59 476 5 65 4.33 42 992 88 76 496 16.5 13 92 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 48 390 92 86 508 18.3 14 93 2.1 72 80.7 85 7 75.9 0.4 4.6 15.
Greece	0.5 24 1.2 17 002 18 117 17.4 52 16.95 25 124 82 72 458 16.9 18 69 1.9 64 81.1 74 5.2 61.8 1 7.3 (€) 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 (€) 15.0
Iceland	0 24 1.6 (€) 30 453 64 398 2.6 88 0.26 55 984 98 78 481 19.3 3 99 1.6 79 82.5 76 7.5 87 0.9 (€) 15.06 (€) 14.1
Ireland	0.1 21 2.1 25 439 43 493 2.1 65 4.72 51 681 98 80 509 18.7 7 82 0.8 65 81.5 82 7 75.5 0.6 4.68 (€) 15.2

Legend:
€ 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

International Monetary Fund

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

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

World Economic Outlook Database, April 2016

Step 5 of 5

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.此部分代码已经上传至Github的code文件夹下，同学们可以自行取用。这里就不再赘述。

6 Main challenge of Machine Learning

To summarize, "bad algorithm" and "bad data"

6.1 Insufficient Quantity of Training Data

The book demonstrate this point by an example. The result suggest that we may want to reconsider the trade-off between spending time and money on algorithm development versus spending it on corpus development.

6.2 Nonrepresentative Training Data

The US presidential election in 1936. The result of the election predict by Literary Digest was wrong because the list tending to favor wealthier people and ruling out people who don't care much about politics.

6.3 Poor-Quality Data

Errors, outliers, noise.

6.4 Irrelevant Features

Garbage in, garbage out.

6.5 Overfitting the training data

Overfitting means that the model performs well on the training data, but it does not generalize well.

How to resolve overfitting:

- Simplify the model.

- gather more train data.
- fix data errors and remove outliers.

The way to reduce the risk of overfitting is regularization. Find the right balance between fitting the data perfectly and keeping the model simple.

6.6 Underfitting the training data

- Selecting a more powerful model, with more parameters.
- Feeding better features to the learning algorithm.
- Reducing the constraints on the model.

7 Testing and Validating

Cross-validation.

8 Summarization

Simple problems using linear regression, more complex problems using Neural Network.