怎样学习机器学习

— Some Tips for Learning Machine Learning

卿来云

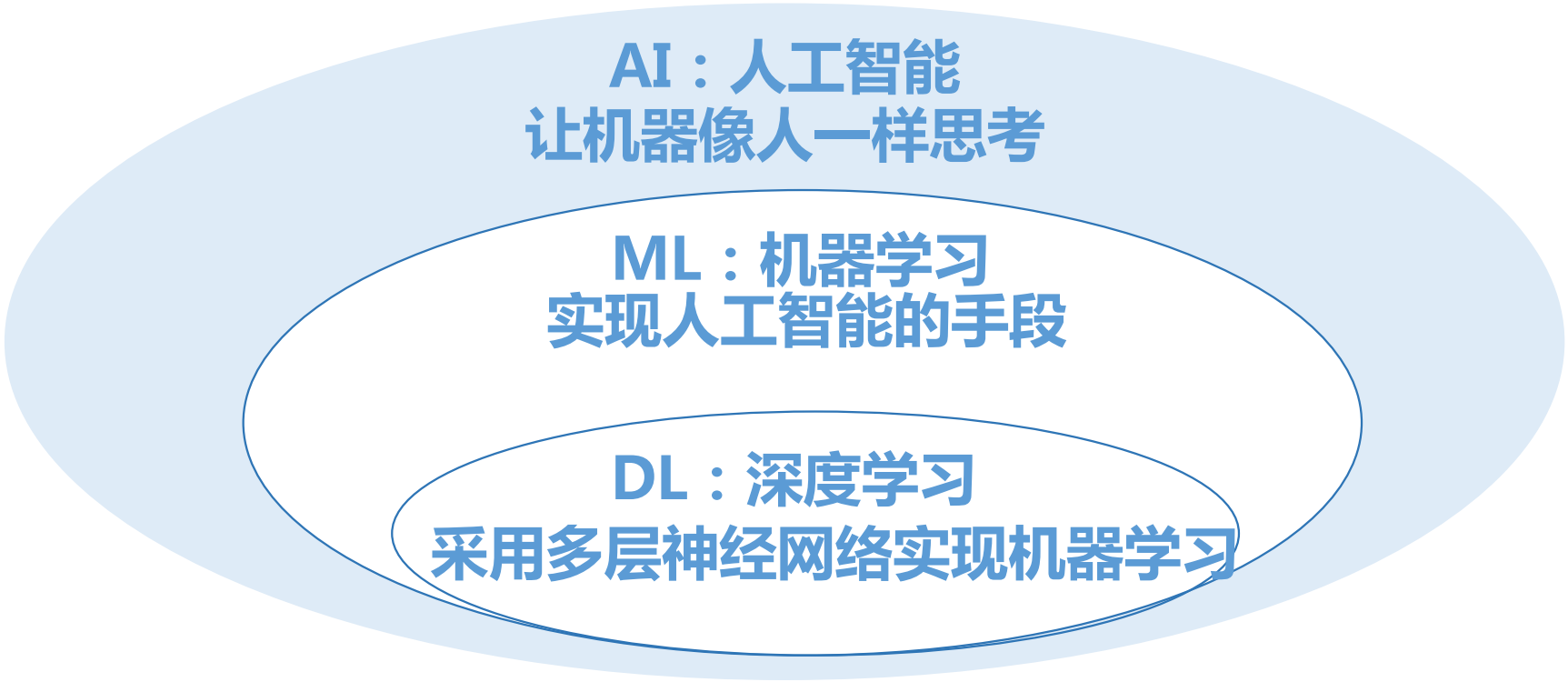
lyqing@ucas.ac.cn

中国科学院大学 计算机科学与技术学院

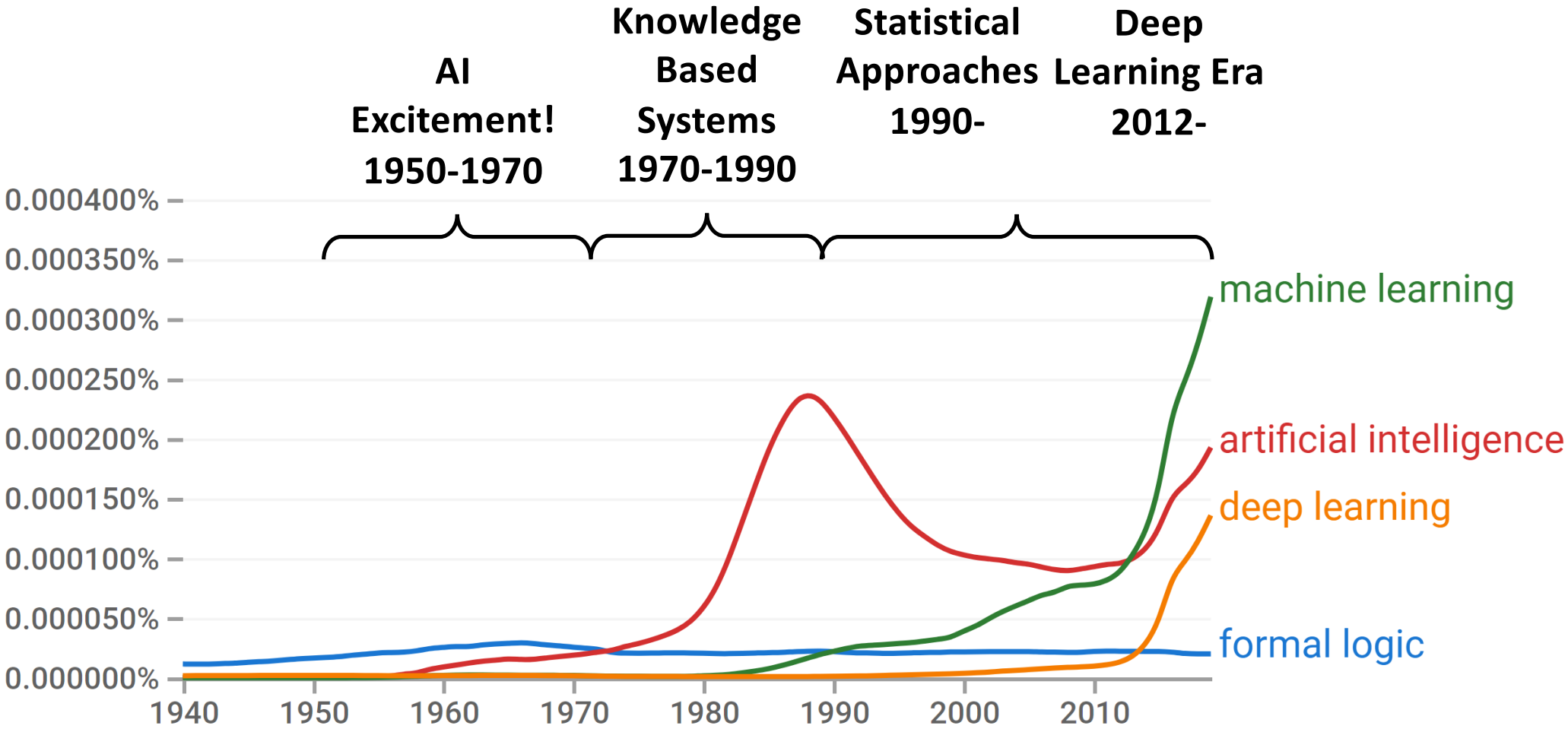
大纲

- 机器学习简介
- 学习目标
- 撸起袖子加油干
- 网络课程推荐

机器学习 vs. 人工智能


AI：人工智能
让机器像人一样思考
ML：机器学习
实现人工智能的手段
DL：深度学习
采用多层神经网络实现机器学习

AI、ML和DL的历史变迁



https://books.google.com/ngrams

机器学习应用举例：个性化推荐

Everything is personalized

Over 75% of wha
people watch
comes from a
recommendation

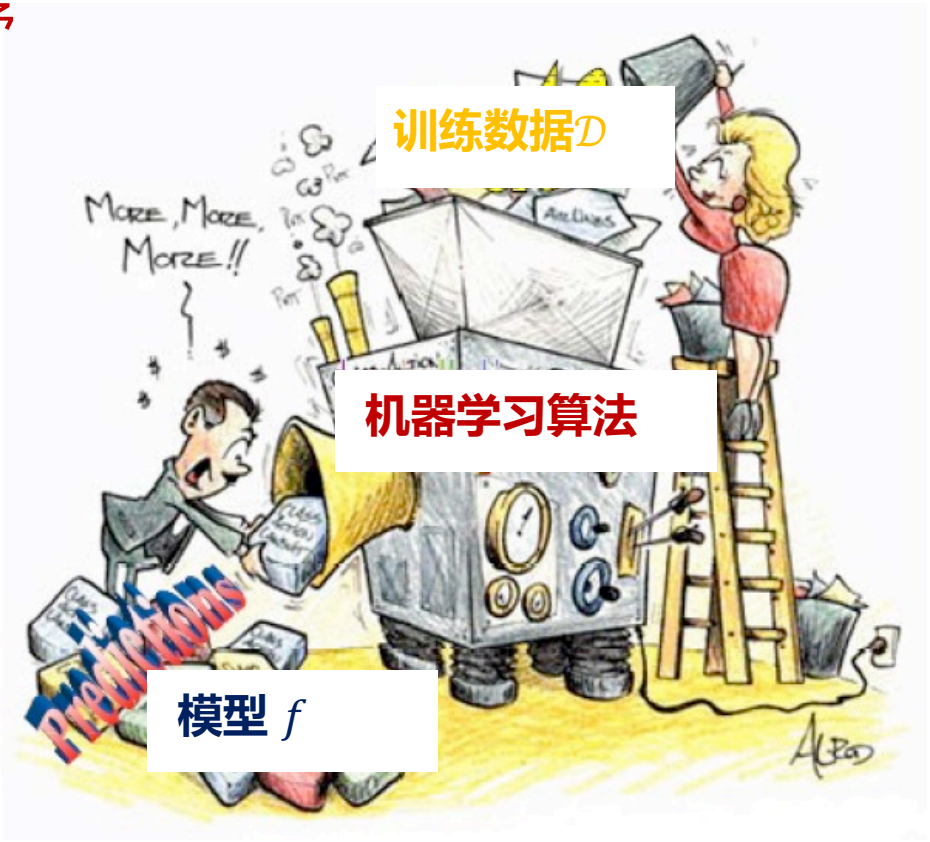
NETFLIX

图片： https://digital.hbs.edu/platform-rctom/submission/promise-and-peril-for-machine-learning-at-netflix/

什么是机器学习？

对于某类任务T和性能度量P，如果计算机程序在T上以P衡量的性能随着经验E而自我完善，就称这个计算机程序从经验E学习。

—— Tom Mitchell

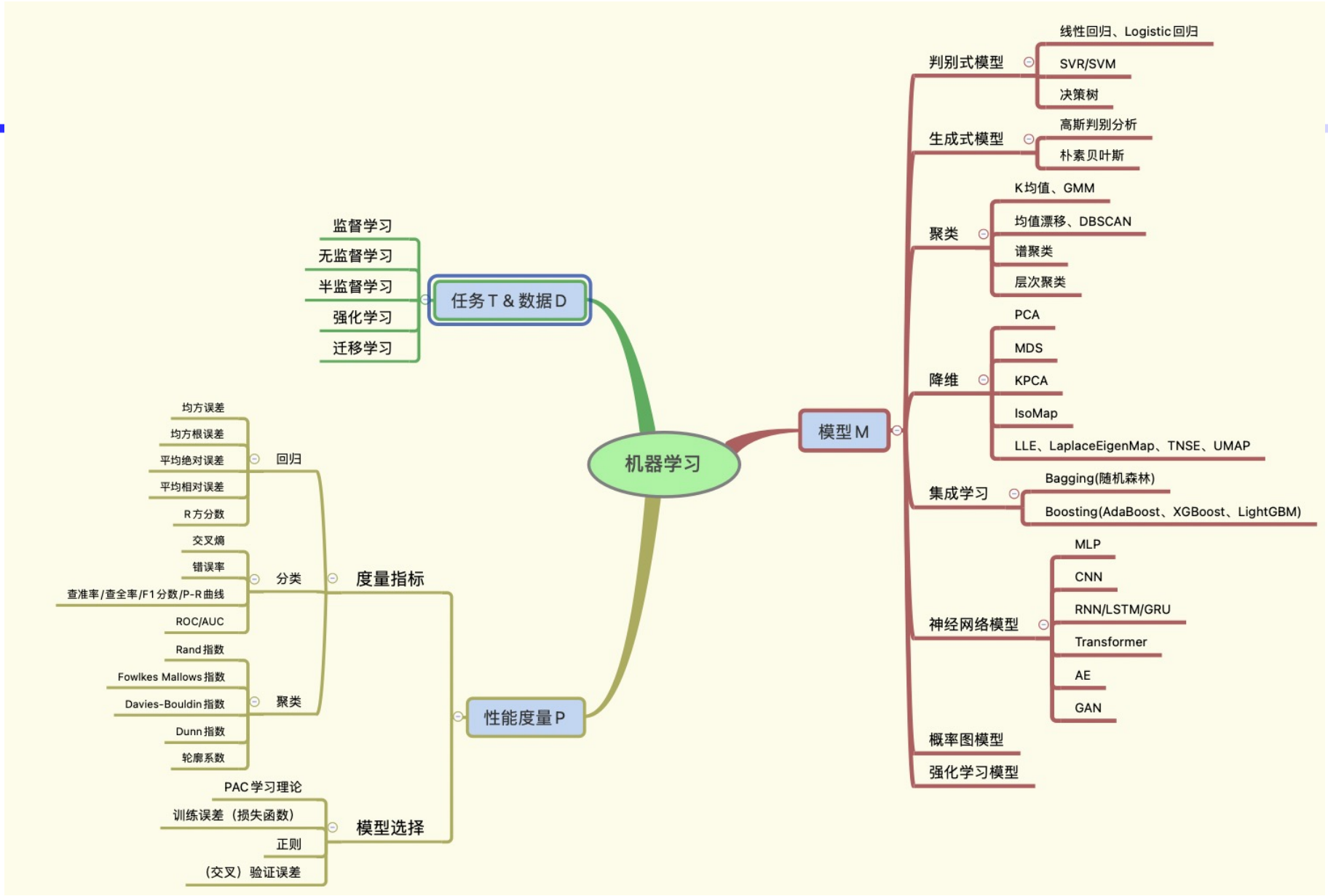


经验E: 训练数据$\mathcal{D}$

模型：预测函数f

机器学习算法：怎样训练数据$\mathcal{D}$中得到模型f

性能评价：模型有多好


机器学习
任务 T & 数据 D
监督学习
无监督学习
半监督学习
强化学习
迁移学习
性能度量 P
度量指标
回归
均方误差
均方根误差
平均绝对误差
平均相对误差
R 方分数
分类
交叉熵
错误率
查准率/查全率/F1 分数/P-R 曲线
ROC/AUC
聚类
Rand 指数
Fowlkes Mallows 指数
Davies-Bouldin 指数
Dunn 指数
轮廓系数
模型选择
PAC 学习理论
训练误差（损失函数）
正则
（交叉）验证误差
模型 M
判别式模型
线性回归、Logistic 回归
SVR/SVM
决策树
生成式模型
高斯判别分析
朴素贝叶斯
聚类
K均值、GMM
均值漂移、DBSCAN
谱聚类
层次聚类
降维
PCA
MDS
KPCA
IsoMap
LLE、LaplaceEigenMap、TNSE、UMAP
集成学习
Bagging(随机森林)
Boosting(AdaBoost、XGBoost、LightGBM)
神经网络模型
MLP
CNN
RNN/LSTM/GRU
Transformer
AE
GAN
概率图模型
强化学习模型

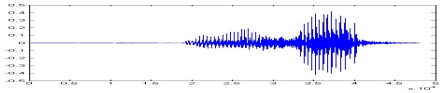
机器学习之数据$\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^{N}$

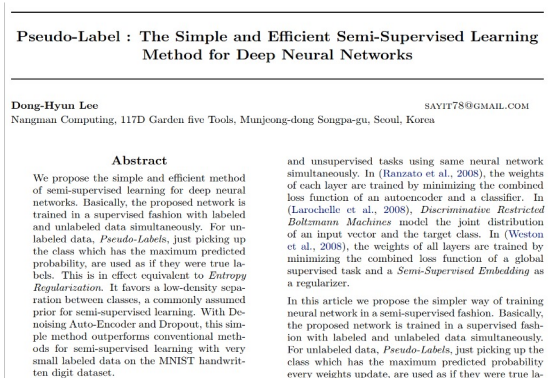
- 结构型数据

可能需要对原始数据进行适当的预处理

电视广告费用	广播广告费用	报纸广告费用	产品销量
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	9.3
151.5	41.3	58.5	18.5
180.8	10.8	58.4	12.9

- 非结构数据：特征提取 → 表示学习（深度模型）





Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks

Dong-Hyun Lee SAYIT78@GMAIL.COM
Nangman Computing, 117D Garden five Tools, Munjeong-dong Songpa-gu, Seoul, Korea

Abstract

We propose the simple and efficient method of semi-supervised learning for deep neural networks. Basically, the proposed network is trained in a supervised fashion with labeled and unlabeled data simultaneously. For unlabeled data, *Pseudo-Labels*, just picking up the class which has the maximum predicted probability, are used as if they were true labels. This is in effect equivalent to *Entropy Regularization*. It favors a low-density separation between classes, a commonly assumed prior for semi-supervised learning. With Denoising Auto-Encoder and Dropout, this simple method outperforms conventional methods for semi-supervised learning with very small labeled data on the MNIST handwritten digit dataset.

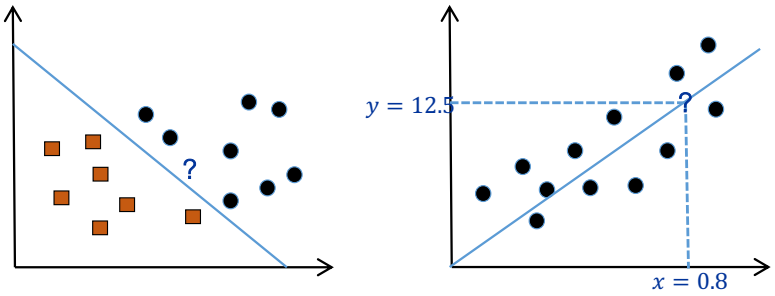
and unsupervised tasks using same neural network simultaneously. In (Ranzato et al., 2008), the weights of each layer are trained by minimizing the combined loss function of an autoencoder and a classifier. In (Larochelle et al., 2008), *Discriminative Restricted Boltzmann Machines* model the joint distribution of an input vector and the target class. In (Weston et al., 2008), the weights of all layers are trained by minimizing the combined loss function of a global supervised task and a *Semi-Supervised Embedding* as a regularizer.

In this article we propose the simpler way of training neural network in a semi-supervised fashion. Basically, the proposed network is trained in a supervised fashion with labeled and unlabeled data simultaneously. For unlabeled data, *Pseudo-Labels*, just picking up the class which has the maximum predicted probability every weights update, are used as if they were true la-

机器学习之任务类型 & 数据$\mathcal{D} = \{x_i, y_i\}_{i=1}^N$

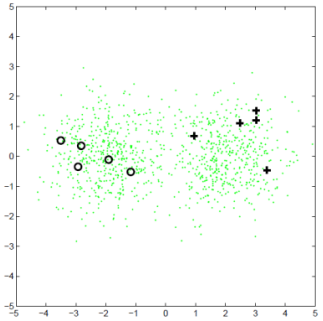
- 监督学习$\mathcal{D} = \{x_i, y_i\}_{i=1}^N$

  $y_i \in \{1,2,\dots,C\}$　　$y_i \in \mathbb{R}$

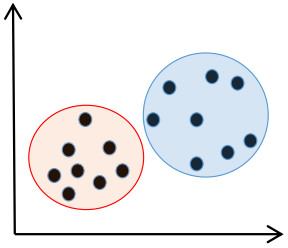


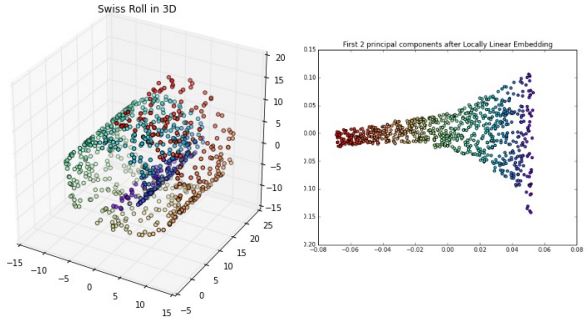
- 半监督学习$\mathcal{D} = \{x_i, y_i\}_{i=1}^L$，

  $\mathcal{D}_U = \{x_{L+1}, x_{L+2}, \dots x_{L+U}\}$



- 无监督学习$\mathcal{D} = \{x_i\}_{i=1}^N$

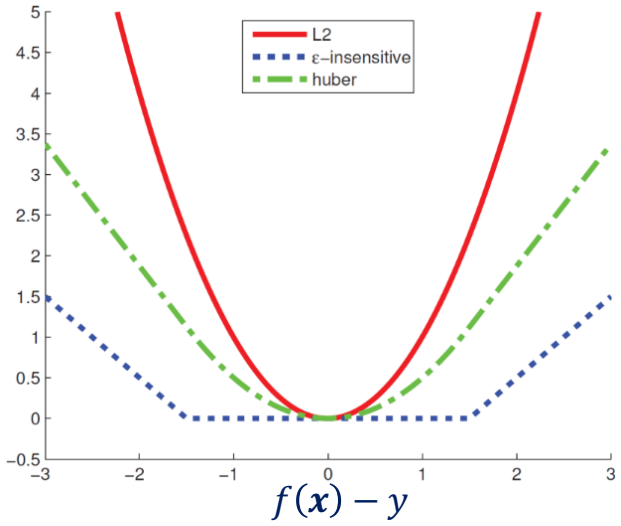




- 强化学习$\mathcal{D} = \{x_i, y_i\}_{i=1}^L$

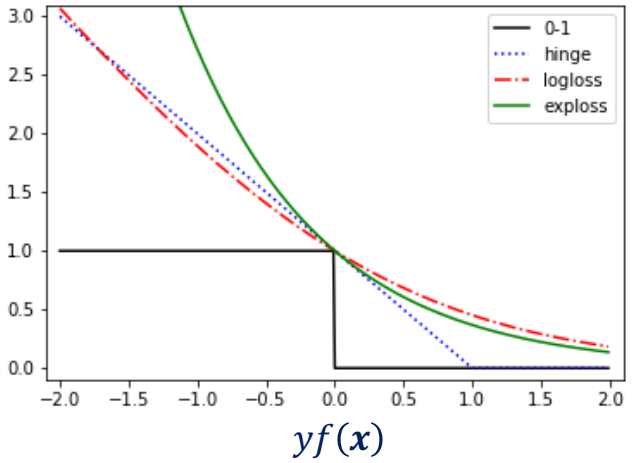
机器学习之模型性能度量

- 回归任务常用误差函数$L(f(\boldsymbol{x}), y)$



常用度量：平均平方误差、均方根误差、平均绝对误差、R2分数、…

- 分类任务常用误差函数$L(f(\boldsymbol{x}), y)$



常用度量：正确率/错误率、查全率/P、查准率/R、 F1分数、 P-R曲线、ROC/AUC、…

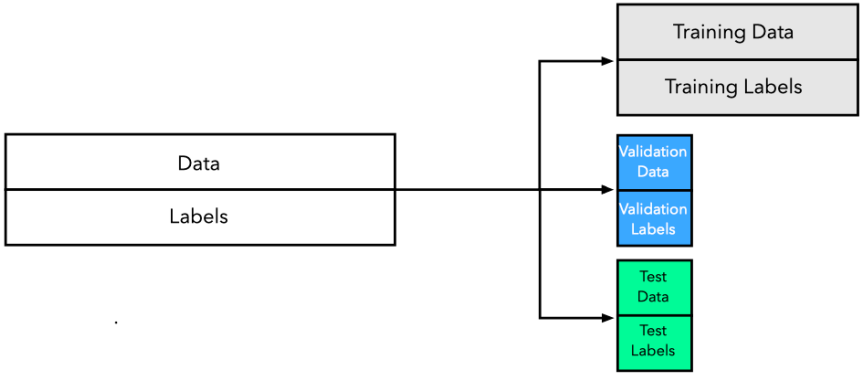
机器学习之模型选择

- 机器学习的目标：在未来的数据上性能好，但我们只有可见的训练数据
- PAC学习理论：在一定条件下，训练误差可近似测试误差
  - 给定小的$\varepsilon > 0$， $P(|e_{train} - e_{test}| > \varepsilon) \leq \delta$，其中 $\delta = 4\frac{(2N)^{d_{VC}}+1}{e^{\frac{1}{8}\varepsilon^2 N}}$
    - 只要训练样本数N足够大和假设空间的VC维d_{VC}足够小，δ就足够小，训练误差可近似测试误差。
  - 等价表示：$\varepsilon = \frac{1}{8}\ln\left(4\frac{(2N)^{d_{VC}}+1}{\delta}\right)$

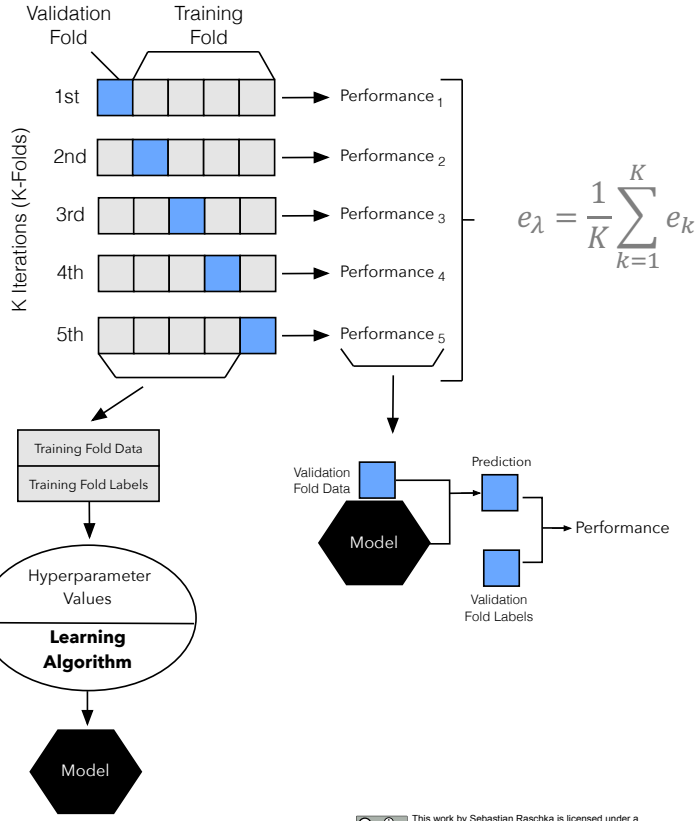
    $e_{train} - \sqrt{\varepsilon} \leq e_{test} \leq e_{train} + \sqrt{\varepsilon}$
    - 要想测试误差小，既要训练误差小，训练样本数N足够大和假设空间的VC维d_{VC}足够小。

机器学习之模型选择

- 模型选择：选择测试误差最小的模型
  - 训练误差$e_{train} = \frac{1}{N}\sum_{i=1}^{N} L\left(y_i, f(\boldsymbol{x}_i)\right)$小
  - 模型VC维小，复杂度低：正则值$R(f)$小
- 目标函数： $J(f) = \frac{1}{N}\sum_{i=1}^{N} L\left(y_i, f(\boldsymbol{x}_i)\right) + \lambda R(f)$
- 实际操作：数据 = 训练集 + 验证集 + 测试集



交叉验证



机器学习模型

- 机器学习 ≈ 找到一个函数f

$f($  $)=$ “猫”

- 机器学习模型的三要素
  - 1. 函数集合/假设空间：$\{f_1, f_2, \dots\}$
  - 2. 目标函数$J(f)$：函数的好坏 $J(f)=\frac{1}{N}\sum_{i=1}^{N} L\big(y_i, f(\boldsymbol{x}_i)\big)+\lambda R(f)$
  - 3. 优化算法：找到最佳函数 $f^*=\underset{f}{\operatorname{argmin}} J(f)$

机器学习模型

每个机器学习模型特有：

- 函数集合：$f(x)$的形式
- 目标函数
  - 损失函数：极大似然、几何
  - 正则项
  - 超参数
- 数据预处理

各机器学习模型共有：

- 优化算法
- 模型选择：模型评价指标、超参数调优

大纲

- 机器学习简介
- 学习目标
- 撸起袖子加油干
- 网络课程推荐

个人目标

■ 机器学习的三重境界

第一境：昨夜西风凋碧树。独上高楼，望尽天涯路。

第二境：衣带渐宽终不悔，为伊消得人憔悴。

第三境：众里寻他千百度，暮然回首，那人正在灯火阑珊处。

调包侠：会用工具包和框架解决实际问题（特征工程、参数调优）

武林高手：理解算法背后的数学原理，知其然，知其所以然

一代宗师：创造新的算法，为机器学习领域添砖加瓦

调包侠：仗剑走天涯

- 机器学习包：Sklearn

六个部分：各种机器学习算法接口统一

分类（Classification）
回归（Regression）
聚类（Clustering）
数据降维（ Dimensionality Reduction

模型选择（Model Selection）
数据预处理（Preprocessing）

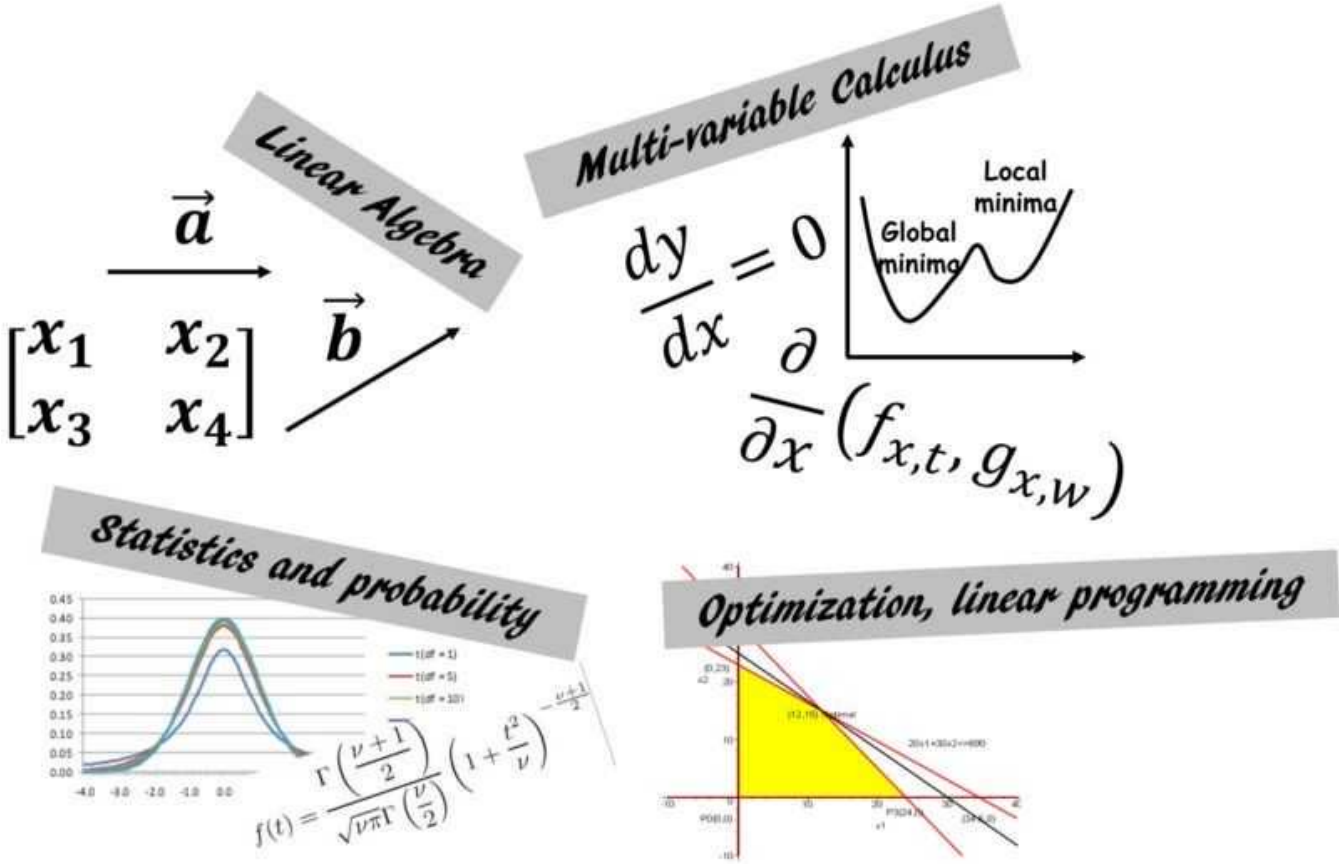
官网：http://scikit-learn.org/stable/

Intel® Extension for Scikit-learn* variant:
Intel® AI Analytics Toolkit (AI Kit)

调包侠：仗剑走天涯

- 深度学习框架

武林高手：数学基础



- 用到数学的时候，再去学习
- 缺失的部分知识在网上搜寻各种视频资料学习
- 如果只是应用，数学部分可适当简化
- 用人的思维方式去理解机器学习算法的基本思想
  - K近邻：近朱者赤，近墨者黑
  - 集成学习：三个臭皮匠顶个诸葛亮
  - 聚类：物以类聚，人以群分

武林高手：手撕代码

惊为天人，普林斯顿博士：手写 30 个主流机器学习算法，代码超3万行，全都开源了！

2019-11-16 雪凝星月_ 阅 44 分享 全屏 转藏

机器之心报道

参与：思源、一鸣、张倩

用 NumPy 手写所有主流 ML 模型，普林斯顿博士后 David Bourgin 最近开源了一个非常剽悍的项目。超过 3 万行代码、30 多个模型，这也许能打造「最强」的机器学习基石?

NumPy 作为 Python 生态中最受欢迎的科学计算包，很多读者已经非常熟悉它了。它为 Python 提供高效率的多维数组计算，并提供了一系列高等数学函数，我们可以快速搭建模型的整个计算流程。毫不负责任地说，NumPy 就是现代深度学习框架的「爸爸」。

尽管目前使用 NumPy 写模型已经不是主流，但这种方式依然不失为是理解底层架构和深度学习原理的好方法。最近，来自普林斯顿的一位博士后将 NumPy 实现的所有机器学习模型全部开源，并提供了相应的论文和一些实现的测试效果。

- 项目地址：https://github.com/ddbourgin/numpy-ml

根据机器之心的粗略估计，该项目大约有 30 个主要机器学习模型，此外还有 15 个用于预处理和计算的小工具，全部.py 文件数量有 62 个之多。平均每个模型的代码行数在 500 行以上，在神经网络模型的 layer.py 文件中，代码行数接近 4000。

这，应该是目前用 NumPy 手写机器学习模型的「最高境界」吧。

李航：《统计学习方法》代码

https://github.com/fengdu78/lihang-code

https://github.com/Dod-o/Statistical-Learning-Method_Code

周志华：《机器学习》代码

https://github.com/datawhalechina/pumpkin-book

Sklearn源码

https://scikit-learn.org/stable/index.html

sklearn.cluster.KMeans

class sklearn.cluster.KMeans(*n_clusters=8, *, init='k-means++', n_init=10, max_iter=300, tol=0.0001, verbose=0, random_state=None, copy_x=True, algorithm='auto'*) [source]

大纲

- 机器学习简介
- 学习目标
- 撸起袖子加油干
- 网络课程推荐

一、工具包

- Jupyter notebooks： 交互式编程、输出可视化
- NumPy, SciPy, Pandas： 数值计算
- Matplotlib, Seaborn：数据可视化
- Scikit-learn：机器学习

Code

```
# The location of the data file
filepath = 'data/Iris_Data/Iris_Data.csv'

# Import the data
data = pd.read_csv(filepath)

# Print a few rows
print(data.iloc[:5])
```

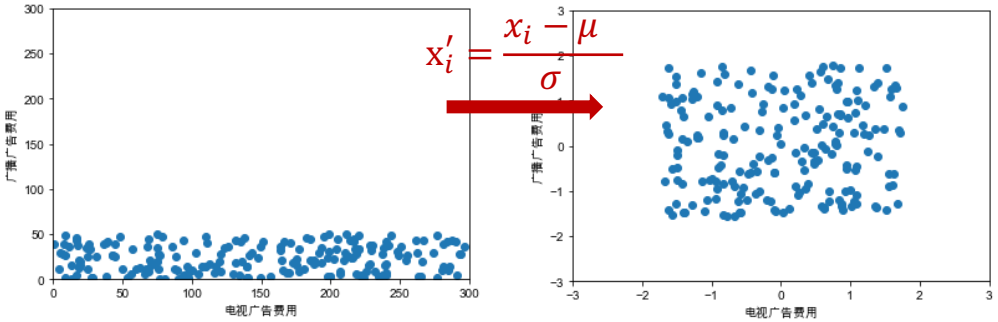
Output

```
>>>
```

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa

二、数学+代码+图示

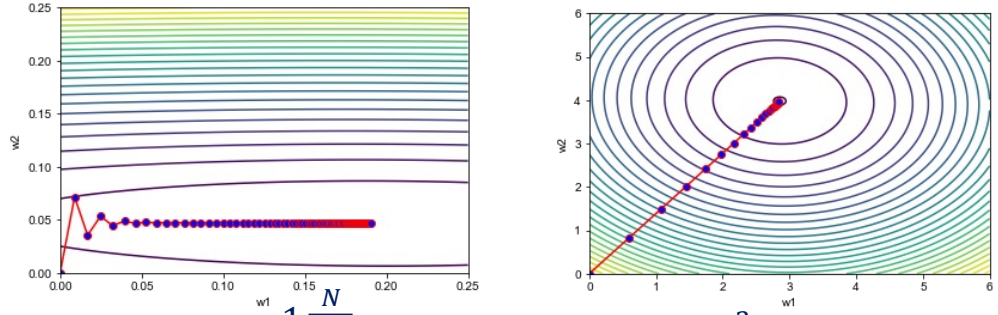
- 数学原理的代码实现
- 适当图示化输出结果，加深理解码
  - 例：梯度下降的特征缩放




```
# 数据标准化
from sklearn.preprocessing import StandardScaler

# 构造输入特征的标准化器
ss_X = StandardScaler()

# 分别对训练和测试数据的特征进行标准化处理
X_train = ss_X.fit_transform(X_train)
X_test = ss_X.transform(X_test)
```




```
#目标函数 等高线
plt.contour(W1, W2, Js_grid, 30)

#梯度下降
ws_bgd, Js_bgd = bgd_ols(X_train, y_train, w_start, grad_OLS, lr =  0.1)

#梯度下降轨迹
ws_np_bgd = np.mat(ws_bgd)
plt.plot(ws_np_bgd[:,0], ws_np_bgd[:,1],
color='r',markerfacecolor='blue',marker='o')
```

$$L(w_1\ ,\ w_2) = \frac{1}{2}\sum_{i=1}^{N}\left(y_i - \left(w_1 x_{i,1} + w_2 x_{i,2}\right)\right)^2$$

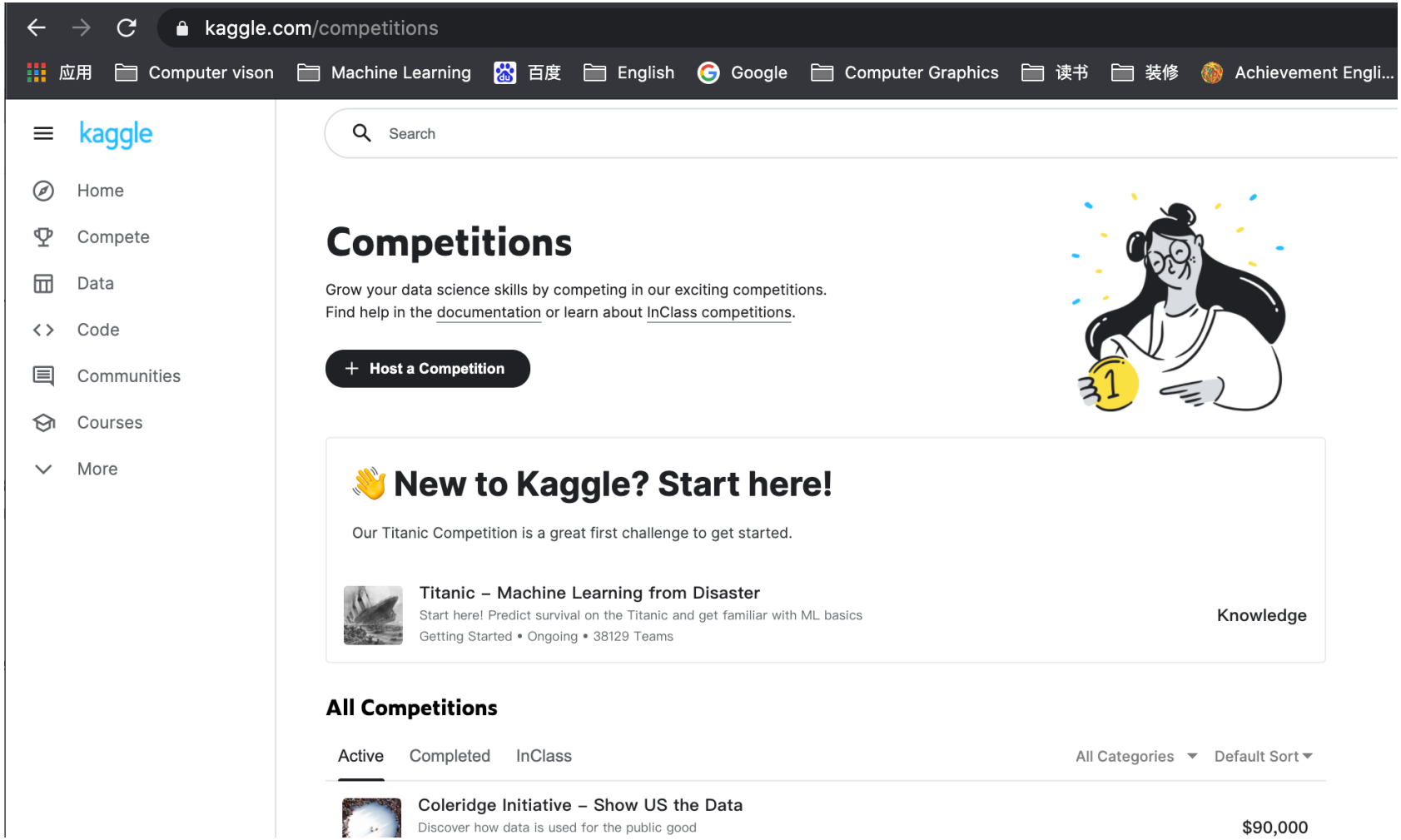
三、以赛代练

- 简单数据/玩具数据：体会算法精髓
- 实际案例数据：体会真实场景中各个组成部分的应用

- Kaggle：https://www.kaggle.com/competitions
- 天池： https://tianchi.aliyun.com/competition/gameList/activeList
- Kesci
- DataCastle
- FlyAI
- 会议、协会、企业、…
  - CCF、中国人工智能学会、阿里、腾讯、京东、…

实践出真知！

例：Kaggle


kaggle.com/competitions
应用
Computer vison
Machine Learning
百度
English
Google
Computer Graphics
读书
装修
Achievement Engli...
kaggle
Search
Home
Compete
Data
Code
Communities
Courses
More
Competitions
Grow your data science skills by competing in our exciting competitions.
Find help in the documentation or learn about InClass competitions.
Host a Competition
New to Kaggle? Start here!
Our Titanic Competition is a great first challenge to get started.
Titanic – Machine Learning from Disaster
Start here! Predict survival on the Titanic and get familiar with ML basics
Getting Started • Ongoing • 38129 Teams
Knowledge
All Competitions
Active
Completed
InClass
All Categories
Default Sort
Coleridge Initiative – Show US the Data
Discover how data is used for the public good
$90,000

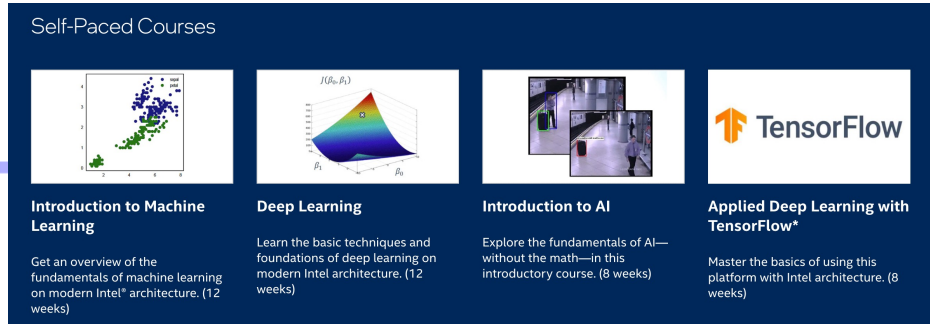
例：天池大数据竞赛


tianchi.aliyun.com/competition/gameList/activeList
应用
Computer vison
Machine Learning
百度
English
Google
Computer Graphics
读书
装修
Achievement Engli...
教育
其他书签
天池大数据竞赛
打造国际高端算法竞赛，让选手用算法解决社会或业务问题
天池出版
赛题解析图书啦
点此购买
Active
算法大赛
创新应用大赛
程序设计大赛
学习赛
可视化大赛
诸神之战
第三届数据库大赛创新上云性能挑战赛--高性能分析性查询引擎赛道
程序设计大赛
赛事简要："数据库性能大赛"自2018年成功举办第一届，已经连续成功举办两届，吸引了国内外数上千支队伍参加，活跃队伍不仅覆盖了国内所有大型互联网公司，各大知名院校组...的队伍亦参与...
举办方:
阿里云 intel
奖金
￥400000
团队
156
赛季1
2021-06-10
进行中

大纲

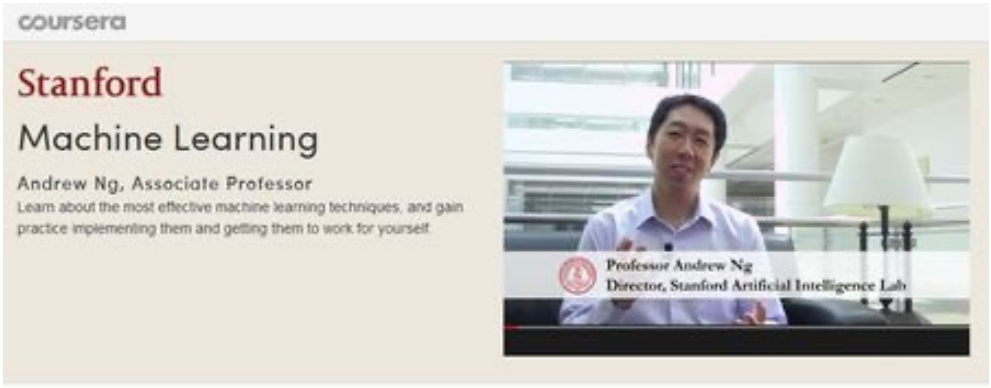
- 机器学习包含的主要内容
- 学习目标
- 撸起袖子加油干
- 网络课程推荐

参考课程



- Intel提供了一些很实用的课程：课件+数据+代码
  - https://www.intel.com/content/www/us/en/developer/topic-technology/artificial-intelligence/training/courses.html
- Introduction to Machine Learning
  - https://www.intel.com/content/www/us/en/developer/learn/course-machine-learning.html
- Deep Learning：
  - https://www.intel.com/content/www/us/en/developer/learn/course-deep-learning.html

参考课程



吴恩达：机器学习（入门级）

https://www.coursera.org/learn/machine-learning
http://open.163.com/newview/movie/courseintro?newurl=M6SGF6VB4

https://github.com/fengdu78/Coursera-ML-AndrewNg-Notes

李宏毅：深度学习（高级）

http://speech.ee.ntu.edu.tw/~tlkagk/courses_ML20.html
https://www.bilibili.com/video/BV1JE411g7XF?p=3
https://datawhalechina.github.io/leeml-notes/#/

感谢各位聆听，欢迎批评指正！

祝大家学有所成！