BUSINESS 2020

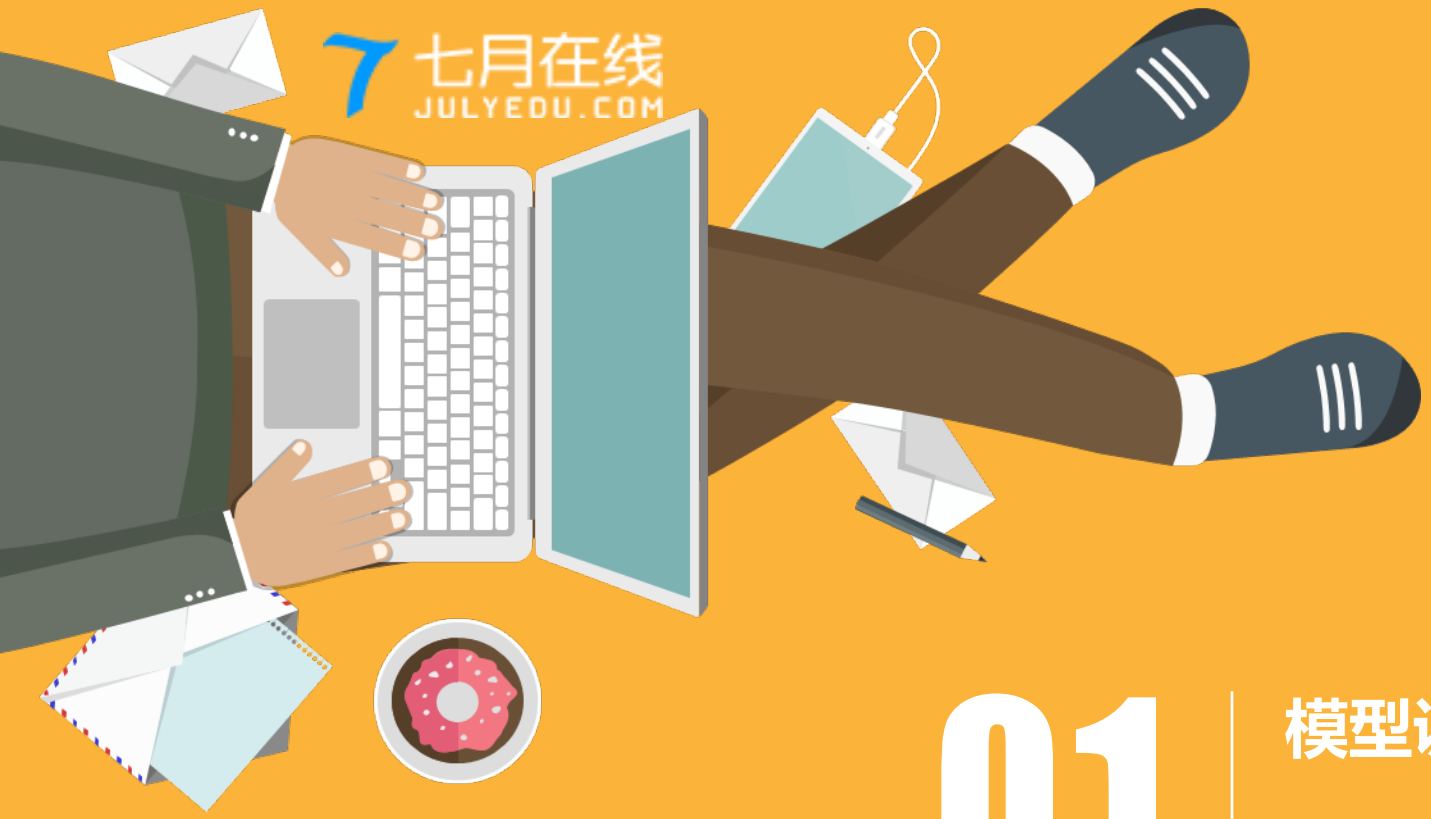
JULYEDU

从头到尾带打Kaggle比赛
模型训练与验证

CV刘老师

https://www.julyedu.com/

CONTENTS



七月在线
JULYEDU.COM

01 模型训练与验证

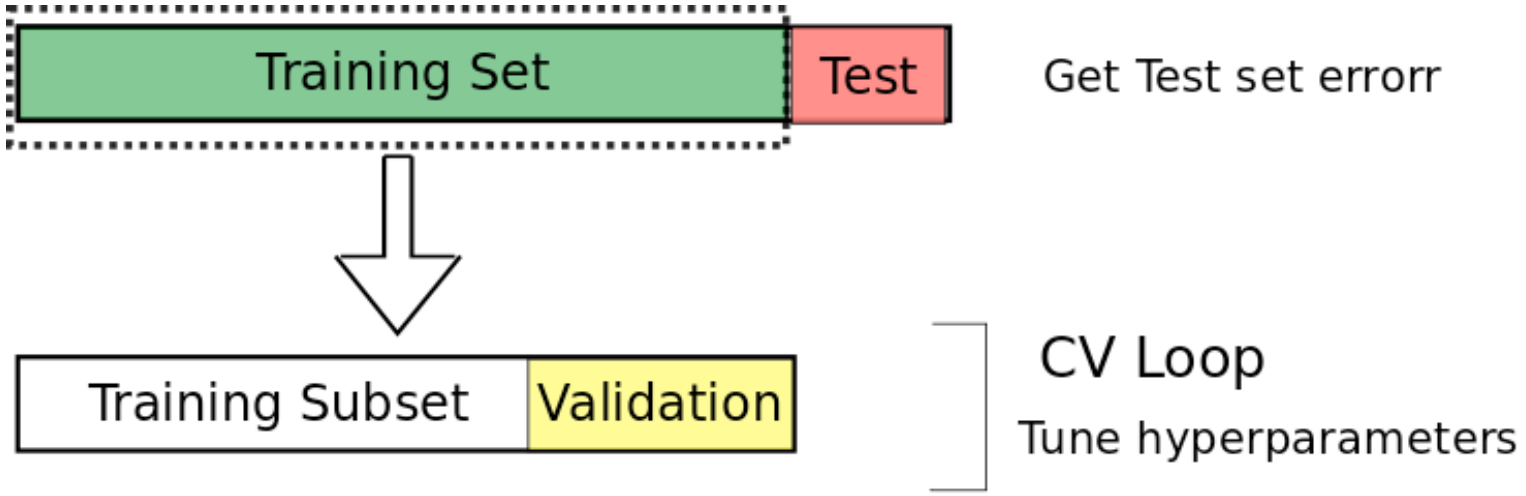
Part 1 模型训练与验证

数据集按照使用用途可以划分为：

✓ 训练集（Training Set）：进行模型训练和参数更新；

✓ 验证集（Validation Set）：进行模型验证集和参数选择；

✓ 测试集（Test Set）：进行验证模型精度；



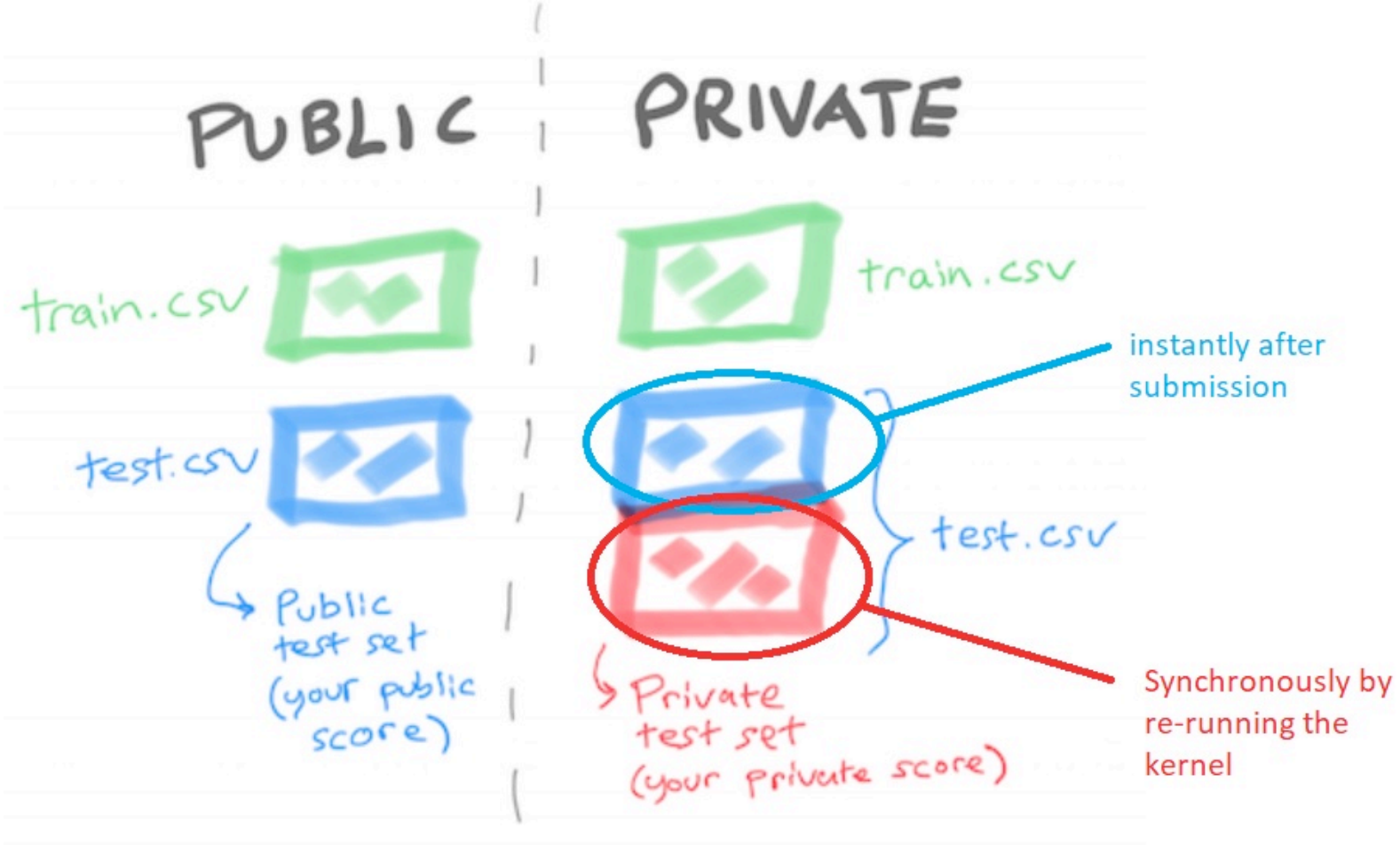
Part 1 模型训练与验证

需要注意：

✓ 测试集一般不能用来训练；

✓ 测试集可能分为AB榜单；

✓ 只要有反馈，就有过拟合；



Part 1 模型训练与验证

在竞赛中测试集在某些情况也可以加入训练，伪标签（Pseudo label）：

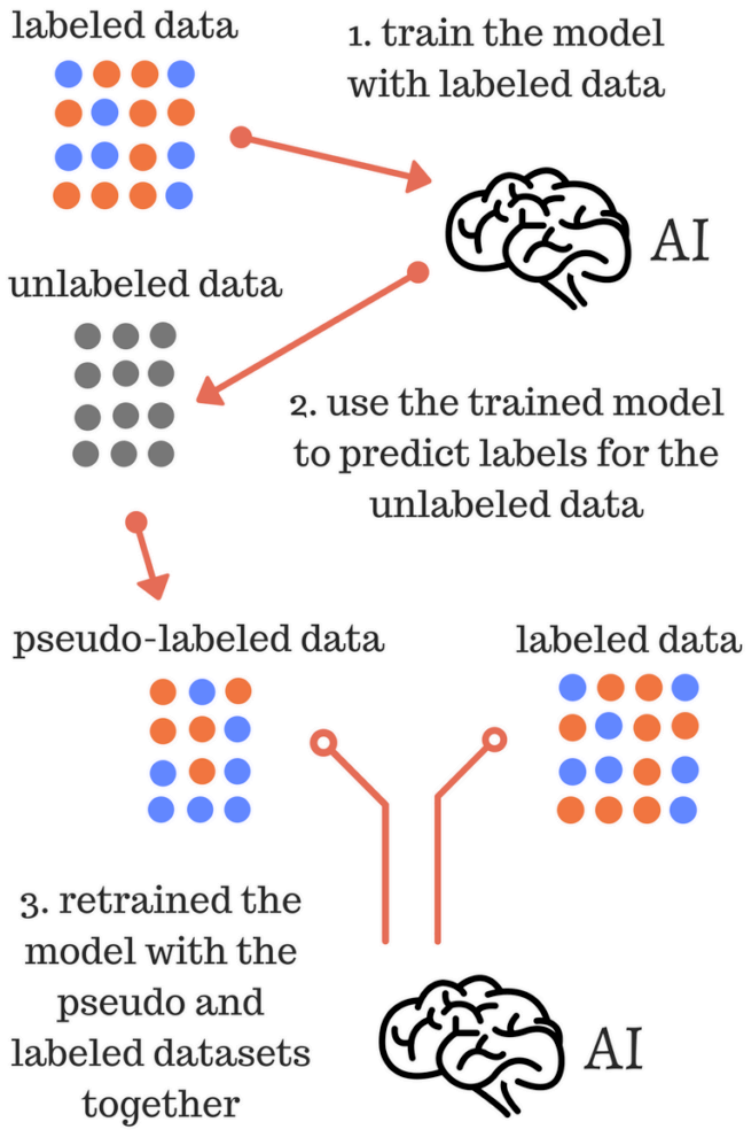
✓ 当模型精度教高时；

✓ 但比赛规则允许时；

模型对测试进行预测，并将预测结果与训练集一起再训练：

https://www.kaggle.com/nvnnghia/yolov5-pseudo-labeling

https://www.kaggle.com/c/kuzushiji-recognition/discussion/112712



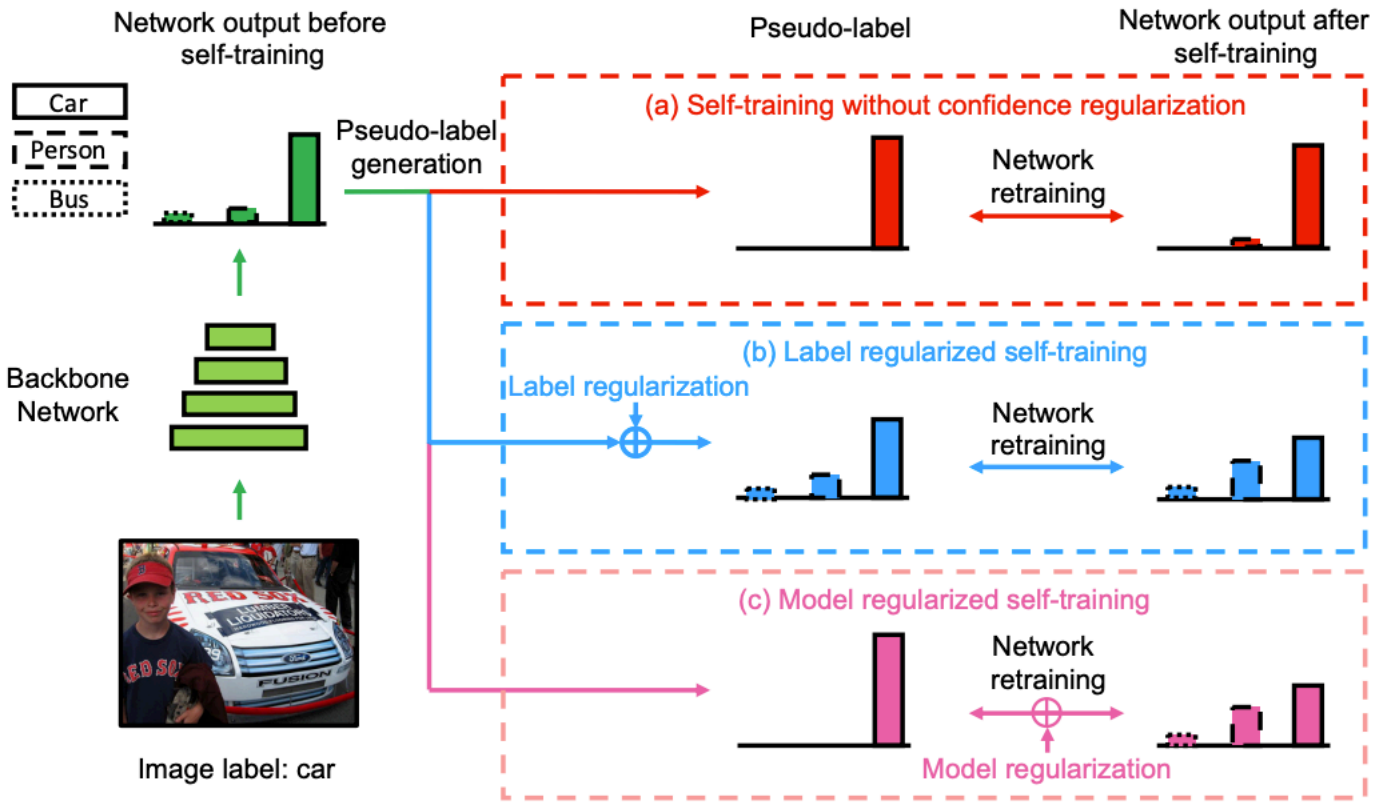
Part 1 模型训练与验证

伪标签（Pseudo label） vs 软标签（Soft label）：

✓ Pseudo label用于测试集打标；

✓ Soft label适用于测试集 和 训练集；

✓ 两者可以同时被使用；



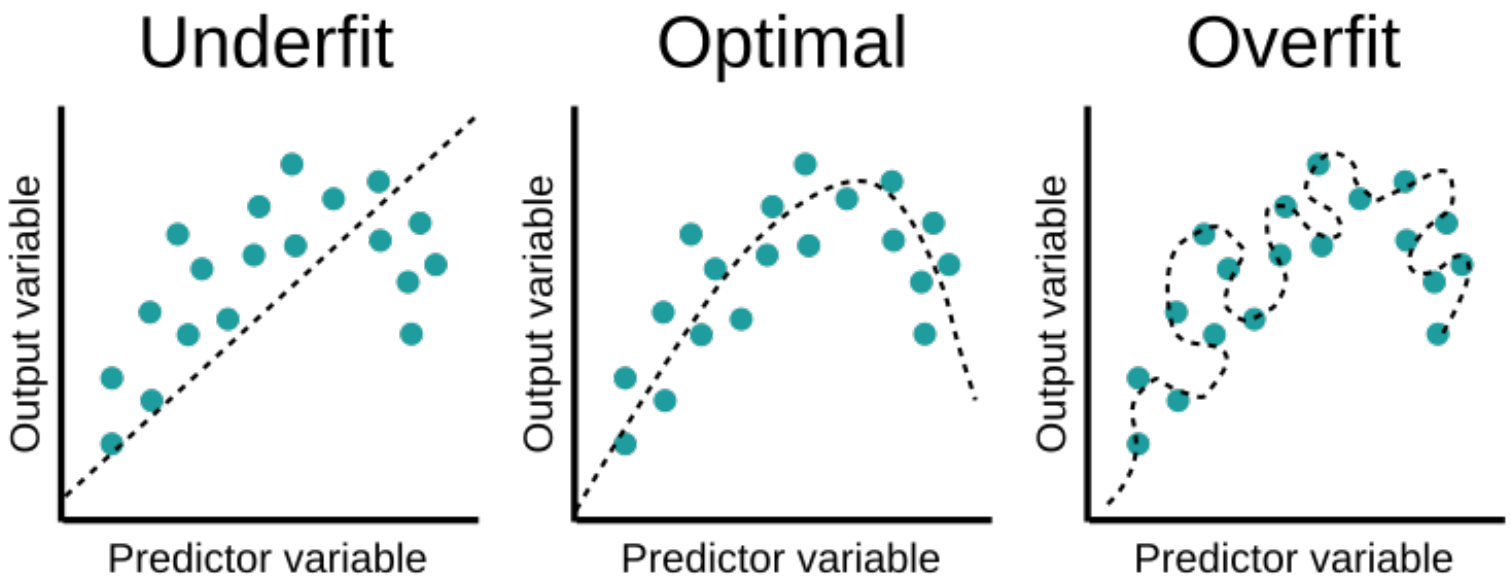
https://arxiv.org/abs/1908.09822

Part 1 模型训练与验证

模型根据训练阶段分为：过拟合 与 欠拟合

✓ 过拟合（Overfit）：模型在训练集精度较好，测试集精度较差；

✓ 欠拟合（Underfit）：模型在训练集精度精度较差，在测试集精度也较差；



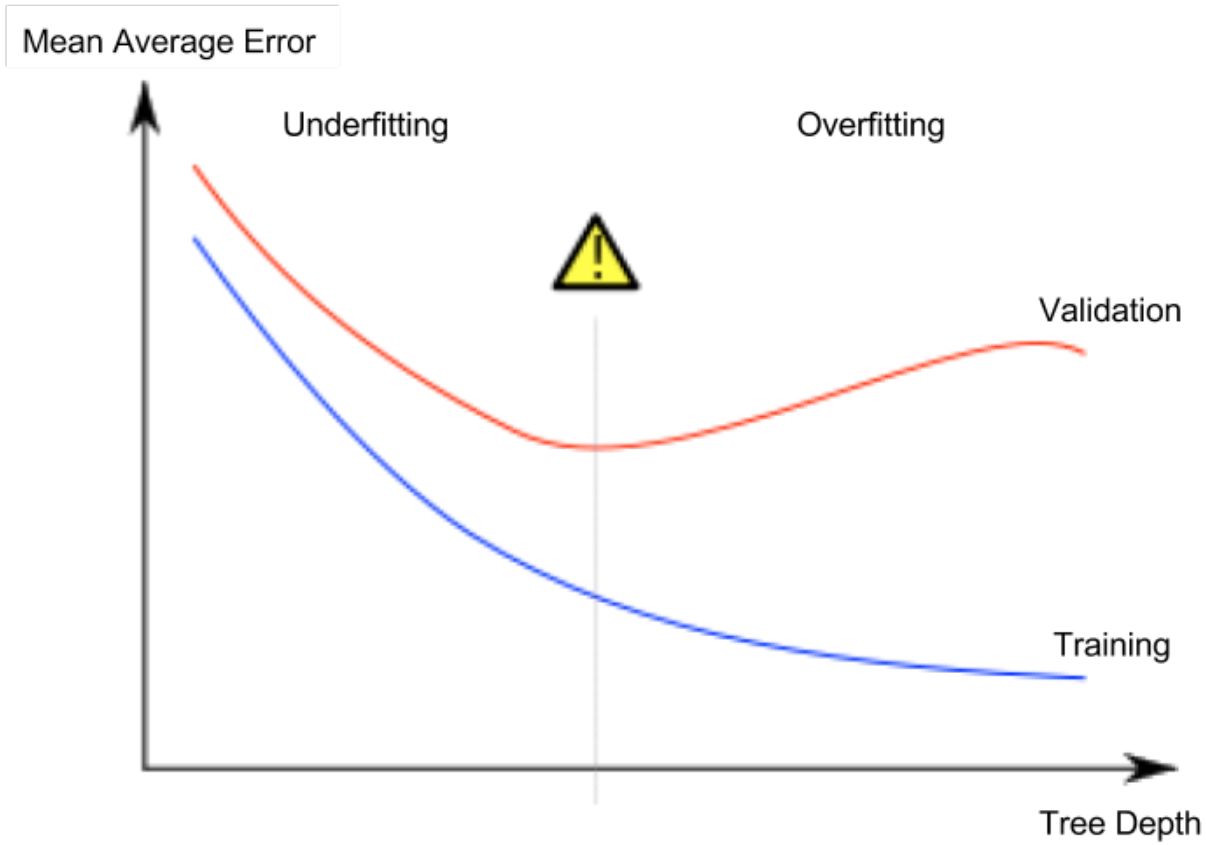
Part 1 模型训练与验证

模型根据训练阶段分为：过拟合 与 欠拟合

✓ 过拟合无法避免，只能缓解；

✓ 缓解过拟合的方法：

- 增加数据量（数据扩增）；
- 做正则化（L1或L2）；
- 做交叉验证（Easy Stopping）；
- 增加随机性（Dropout、样本采样）；



Part 1 模型训练与验证

模型根据训练阶段分为：过拟合 与 欠拟合

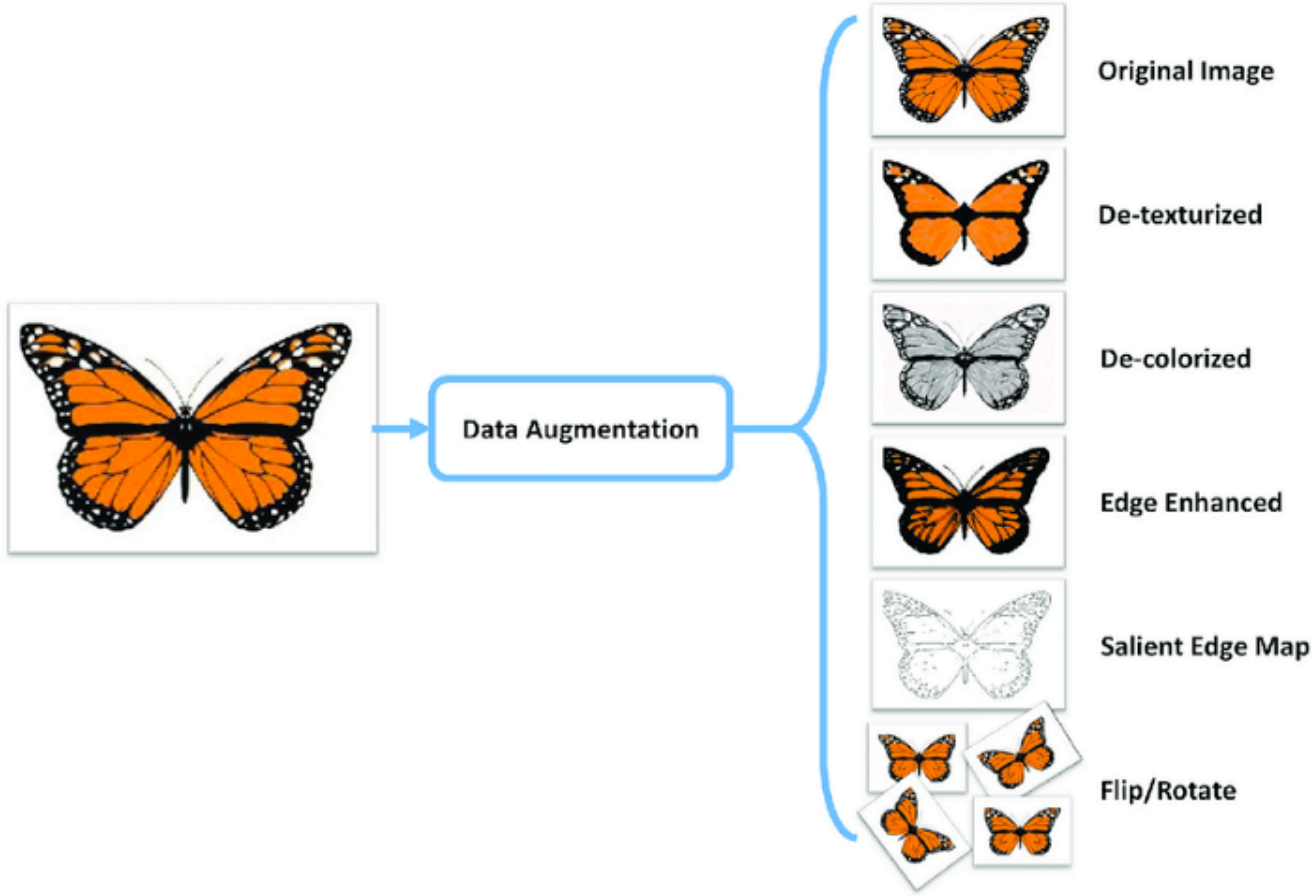
✓ 缓解过拟合的方法：

- 增加数据量（数据扩增）；

数据扩增（Data Augmentation）

1、数据扩增一般在非结构数据中使用；

2、不同类型的数据有不同的数据扩增方法；



Part 1 模型训练与验证

模型根据训练阶段分为：过拟合 与 欠拟合

✓ 缓解过拟合的方法：

- 做正则化（L1或L2）；

正则化（Regularization）

1、让模型保持简单；

2、对传统机器学习方法和深度学习都适用；

$$\lambda \sum_{j=0}^{M} |W_j|$$

L1 Penalty

$$\lambda \sum_{j=0}^{M} W_j^2$$

L2 Penalty

Part 1 模型训练与验证

模型根据训练阶段分为：过拟合 与 欠拟合

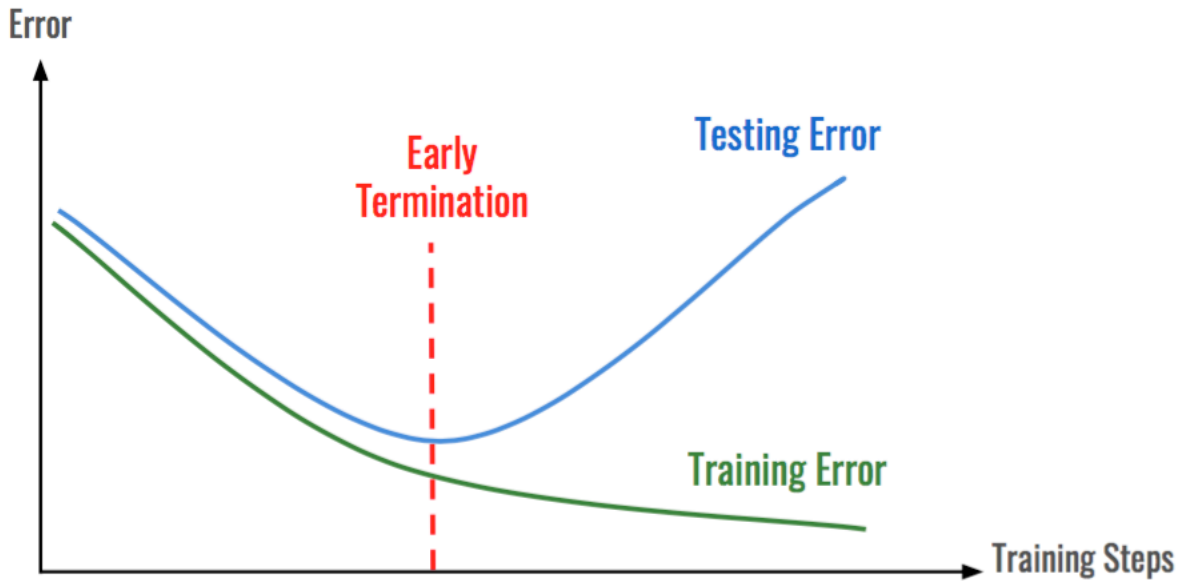
✓ 缓解过拟合的方法：

- 做交叉验证（Early Stopping）；

在划分有验证集后，可以进行控制停止训练：

✓ 对传统机器学习方法和深度学习都适用；

✓ 数据、超参数不同会影响训练轮数；



Part 1 模型训练与验证

模型根据训练阶段分为：过拟合 与 欠拟合

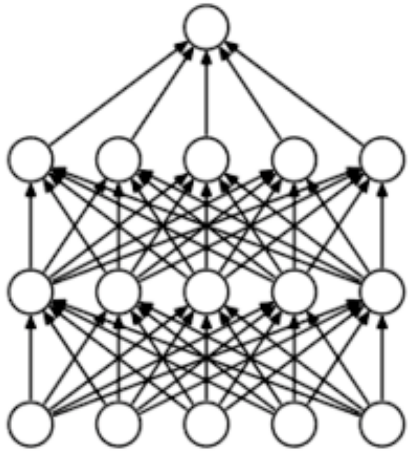
✓ 缓解过拟合的方法：

- 增加随机性（Dropout、样本采样）；

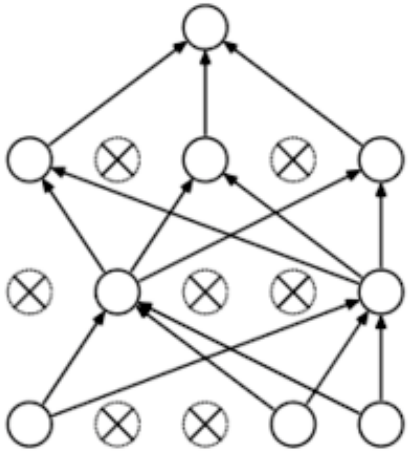
Dropout让模型加入随机性：

1、一般用于深度学习中；

2、有模型集成的思想；



(a) Standard Neural Net



(b) After applying dropout.

Part 1 模型训练与验证

模型根据训练阶段分为：过拟合 与 欠拟合

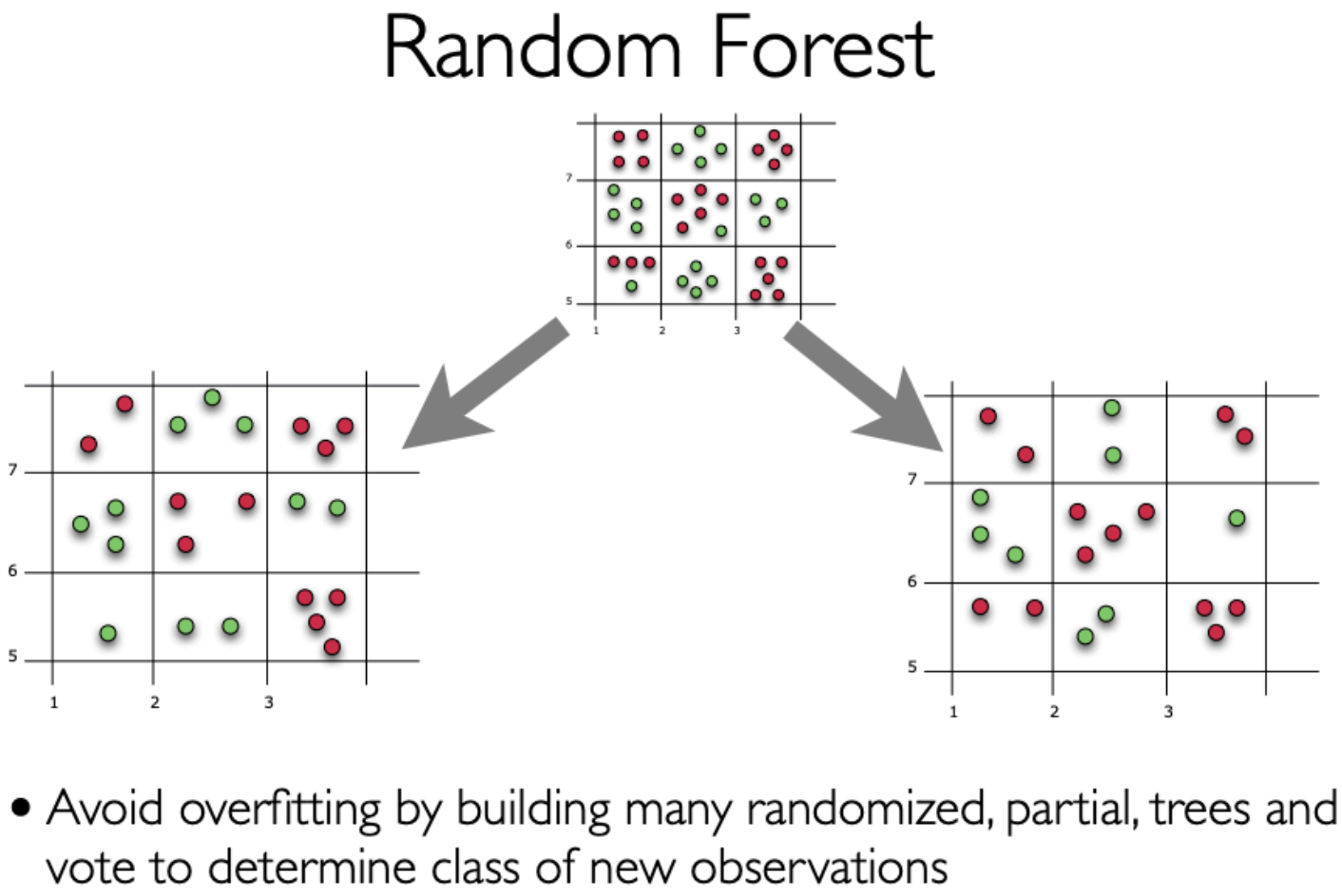
✓ 缓解过拟合的方法：

- 增加随机性（Dropout、样本采样）

随机森林（Random Forest）：

1、样本特征随机输入；

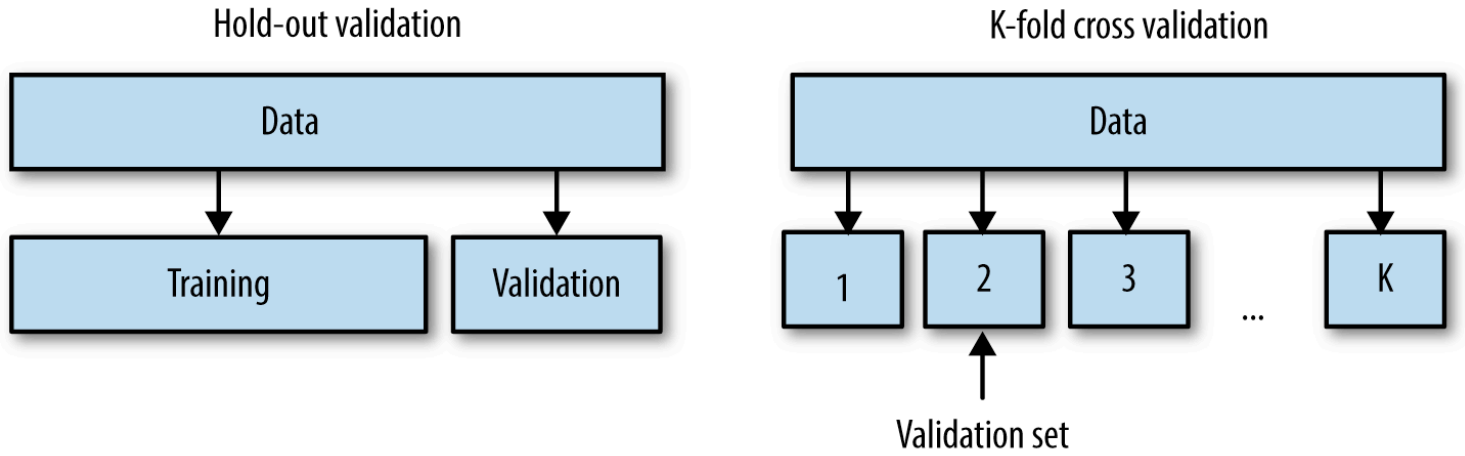
2、树模型集成；

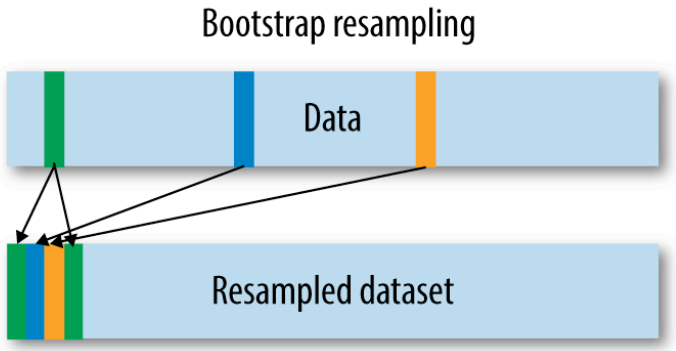


Part 1 模型训练与验证

数据划分（模型评估）方法：

- ✓ 留出法（Hold-out）
- ✓ K折交叉验证（K-fold CV）
- ✓ 自助采样（Bootstrap）





Part 1 模型训练与验证

数据划分（模型评估）方法：

✓ 所有的数据划分方法都需要保证训练集与验证集分布一致；

✓ 思考：

- ✓ 如何判断两个数据集分布是否一致？
- ✓ 不同类型数据（例如图片数据 或 时序数据）如何划分？

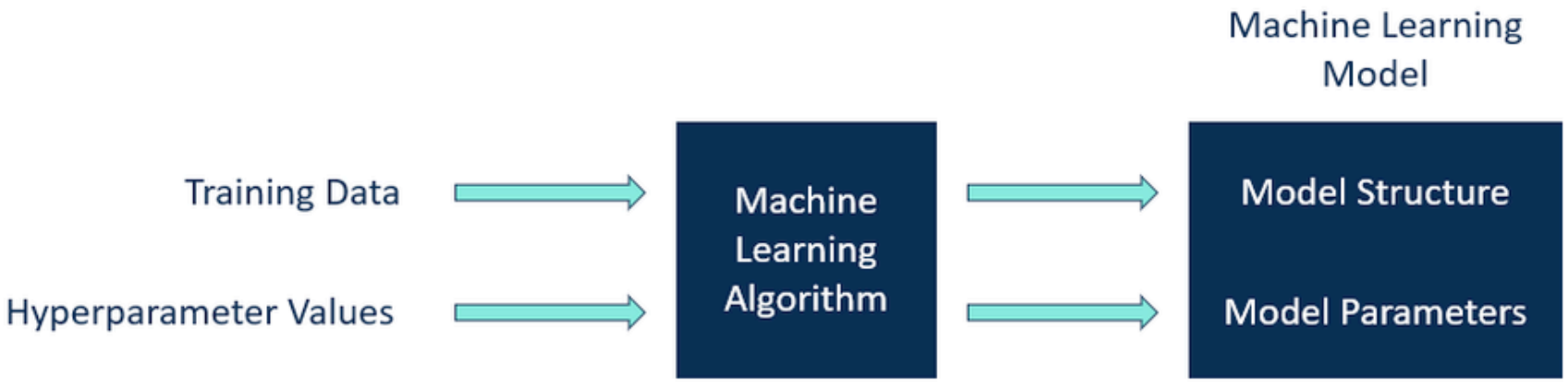
	留出法	K折交叉验证	自助采样
优点	简单、只需要训练一次	所有样本参与训练与验证，模型有多样性	采样得到多个数据集
缺点	验证集没参与训练	时间复杂度高	会改变数据集分布
适用场景	数据量大、通用	竞赛、追求精度	数据量小

Part 1 模型训练与验证

模型参数 vs 模型超参数：

✓ 模型参数（Model parameter）：通过数据可以学习到的参数；

✓ 模型超参数（Model hyperparameters）：需要人为设定的参数，无法通过数据进行学习；



(Model Design + Hyperparameters) → Model Parameters

The building blocks:

- # Layers
- Activations
- Optimizers

...

The knobs that you can turn:

- Learning Rate
- Dropout

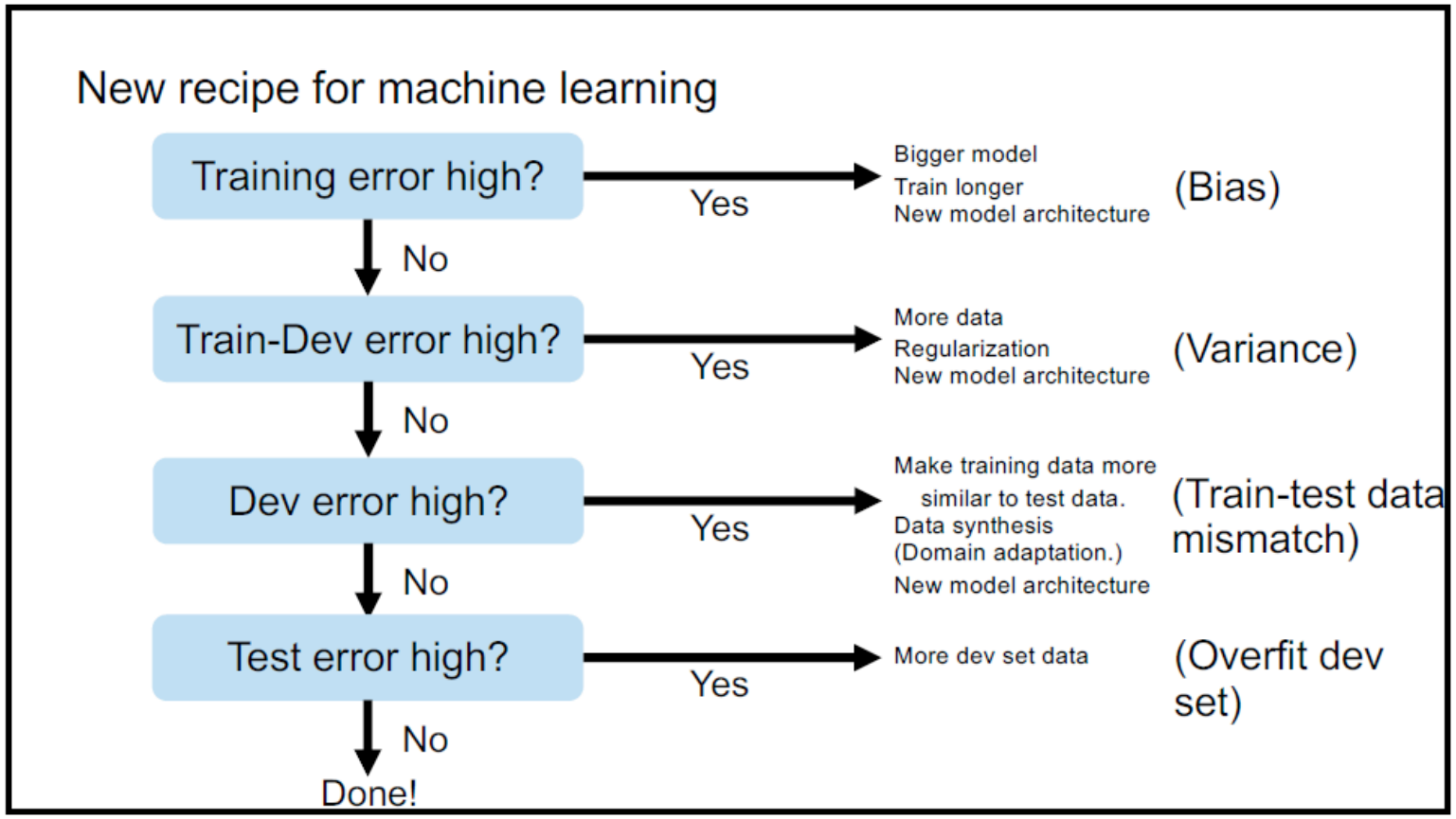
...

The variables learned from the data:

- weights

...

Part 1 模型训练与验证



Part 1 模型训练与验证

学习机器学习：

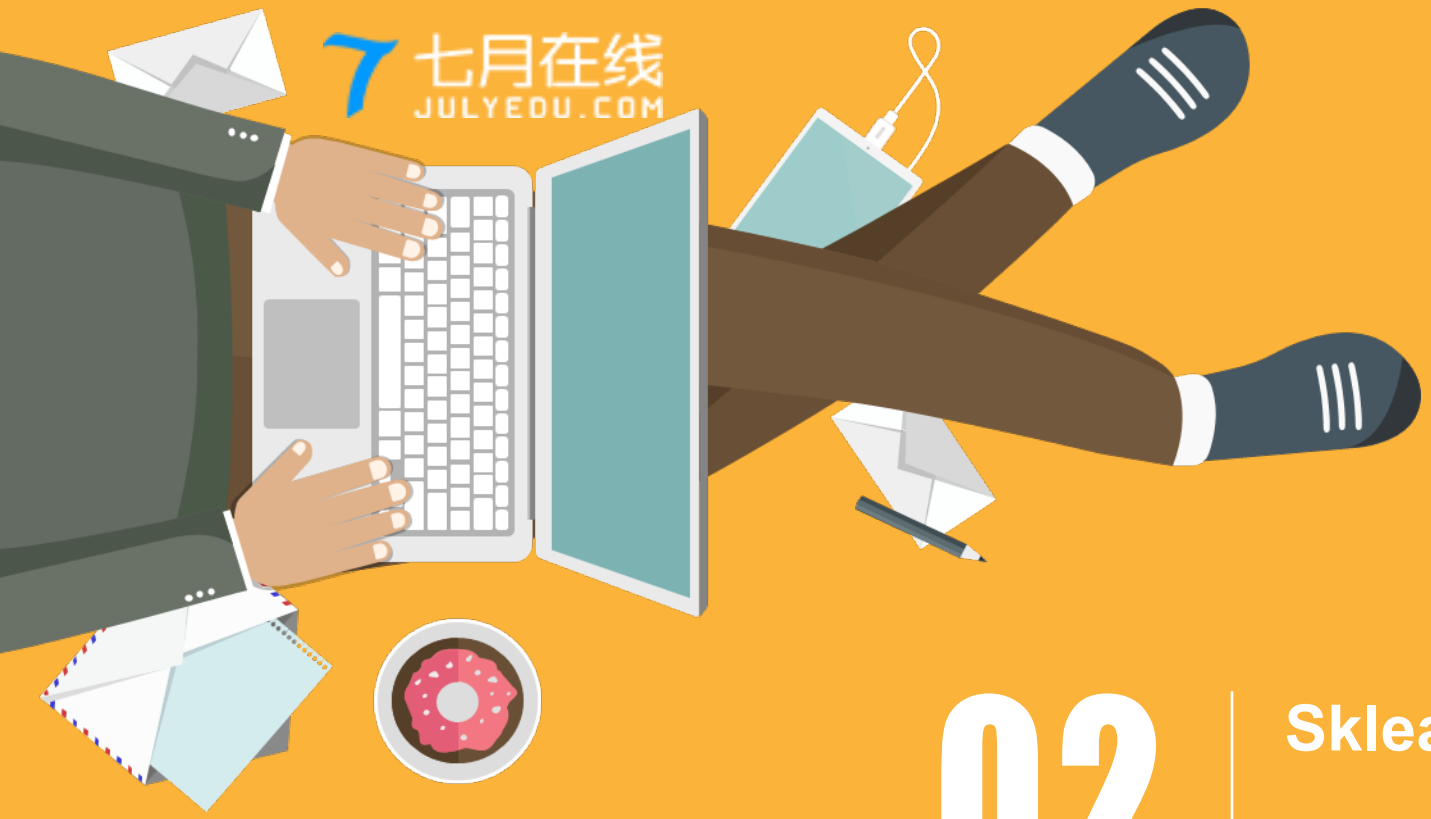
✓ 思路1：从每种算法的原理、公式和推导，从底向上；

✓ 思路2：从算法的使用、超参数开始，从上向下；

思考：

✓ 你是哪种学习方法呢？🤔

✓ 你适合用哪种学习方法呢？



七月在线
JULYEDU.COM

02 Sklearn学习技巧

Part 2 Sklearn使用技巧

scikit-learning是Python最通用的机器学习库：

✓ 提供丰富的机器学习模块；

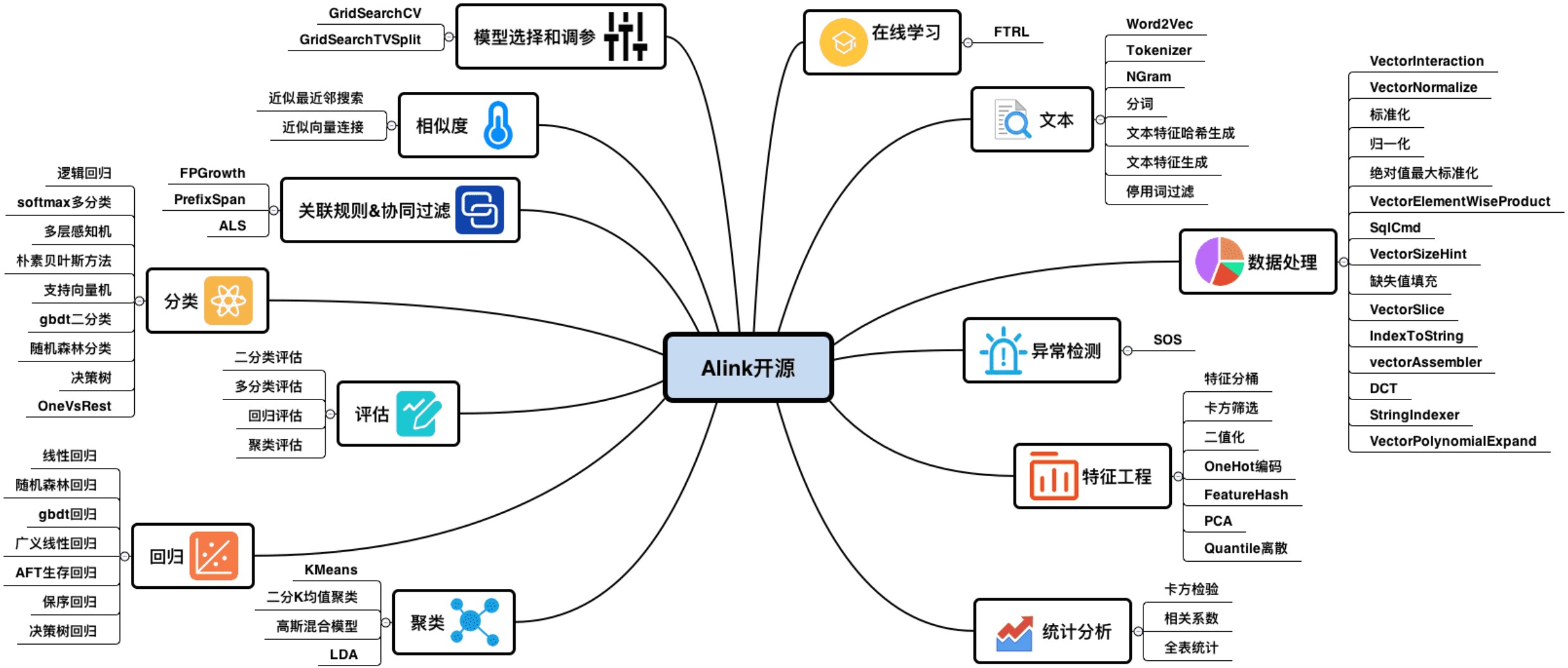
✓ 定义了完善的机器学习API；

```
from sklearn import tree
t = tree.DecisionTreeClassifier(criterion="entropy")

t = t.fit(train_attributes, train_labels)          Build the decision tree

t.score(test_attributes, test_labels)              Build the decision tree

t.predict(example_attributes)                      Predict a new example

cross_val_score(t, all_attributes, all_labels)     Average scores with
                                                   cross-validation
```

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

Part 2 Sklearn使用技巧

https://github.com/alibaba/Alink



Part 2 Sklearn使用技巧

scikit-learning是Python最通用的机器学习库：

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

✓ 十分钟入门：https://scikit-learn.org/stable/getting_started.html

✓ 入门教程：https://scikit-learn.org/stable/tutorial/index.html

✓ 用户指南：https://scikit-learn.org/stable/user_guide.html#

✓ 专业术语表：https://scikit-learn.org/stable/glossary.html

✓ API文档：https://scikit-learn.org/stable/modules/classes.html

Part 2 Sklearn使用技巧

scikit-learning是Python最通用的机器学习库：

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

✓ 特征选择：https://scikit-learn.org/stable/modules/classes.html#module-sklearn.feature_selection

✓ 性能度量：https://scikit-learn.org/stable/modules/classes.html#sklearn-metrics-metrics

✓ 模型选择：https://scikit-learn.org/stable/modules/classes.html#module-sklearn.model_selection

Part 2 Sklearn使用技巧

scikit-learning是Python最通用的机器学习库：

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

sklearn中每个函数和模型：

✓ 文字介绍；

✓ 调用方法；

✓ 参考文献；

References

R33e4ec8c4ad5-1 Y. Freund, R. Schapire, "A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting", 1995.

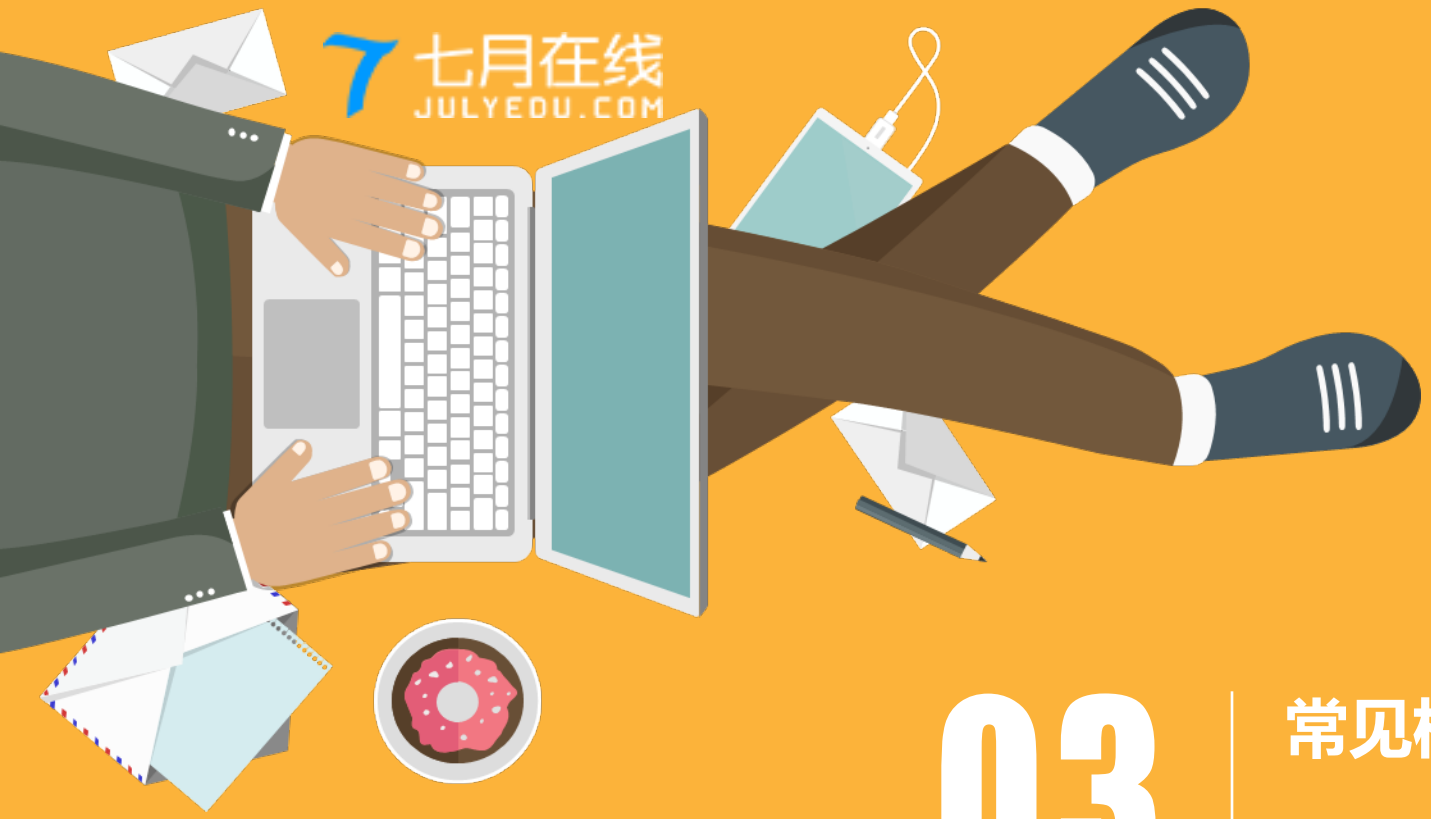
R33e4ec8c4ad5-2 Zhu, H. Zou, S. Rosset, T. Hastie, "Multi-class AdaBoost", 2009.

Examples

```
>>> from sklearn.ensemble import AdaBoostClassifier
>>> from sklearn.datasets import make_classification
>>> X, y = make_classification(n_samples=1000, n_features=4,
...                            n_informative=2, n_redundant=0,
...                            random_state=0, shuffle=False)
>>> clf = AdaBoostClassifier(n_estimators=100, random_state=0)
>>> clf.fit(X, y)
AdaBoostClassifier(n_estimators=100, random_state=0)
>>> clf.predict([[0, 0, 0, 0]])
array([1])
>>> clf.score(X, y)
0.983...
```

Methods

decision_function(self, X)	Compute the decision function of X.
fit(self, X, y[, sample_weight])	Build a boosted classifier from the training set (X, y).
get_params(self[, deep])	Get parameters for this estimator.
predict(self, X)	Predict classes for X.
predict_log_proba(self, X)	Predict class log-probabilities for X.
predict_proba(self, X)	Predict class probabilities for X.



七月在线
JULYEDU.COM

03 常见机器学习模型

Part 2 常见机器学习模型

- 为什么需要这么多的机器学习模型？
  - ***没有绝对最优的机器学习模型，在特定场景下只有比较合适的模型。***
  - 不同模型有特定的偏好和应用场景；
  - 数据量会影响模型的泛化精度；
- 有哪些机器学习模型？
- 每种机器学习模型有什么用处和注意事项？

Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?

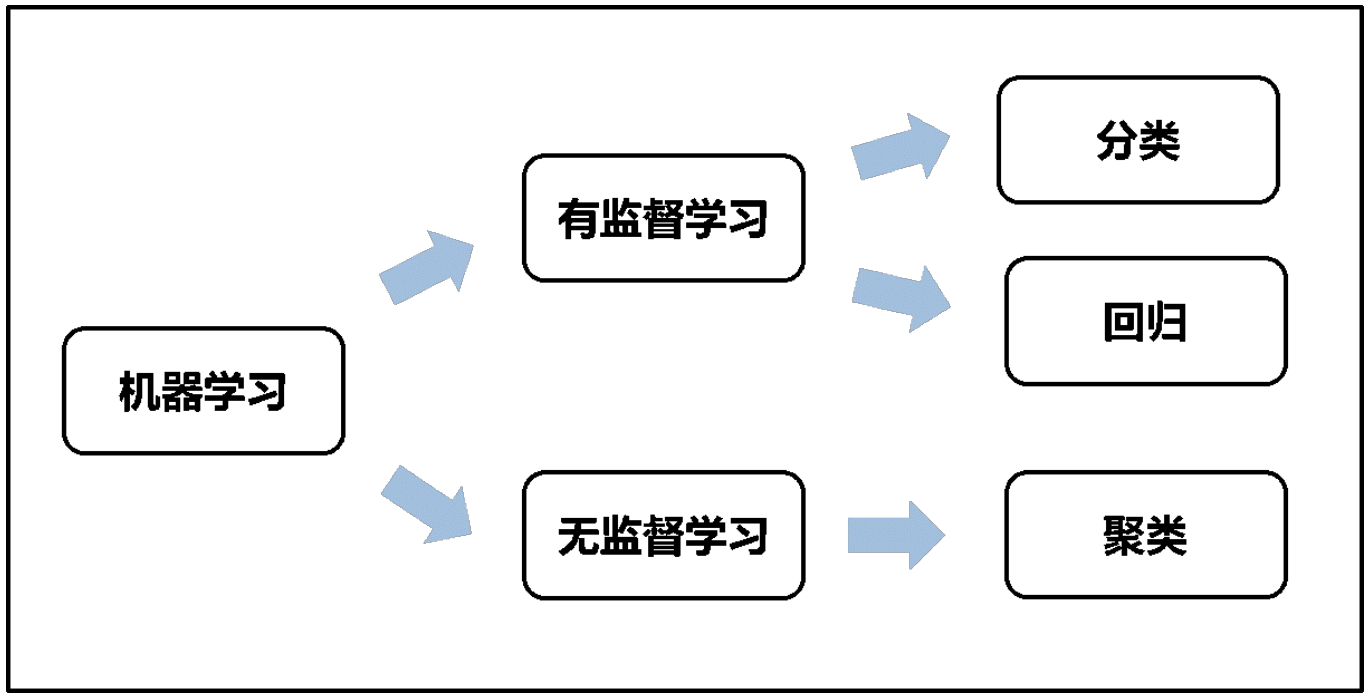
http://jmlr.csail.mit.edu/papers/volume15/delgado14a/delgado14a.pdf

Part 2 常见机器学习模型

模型选择问题1：遇到的问题是什么？

✓有监督 & 无监督？

✓是否可以转化 或者 简化？



Part 2 常见机器学习模型

模型选择问题2：选择什么模型？

✓ 根据数据类型选择模型；

✓ 根据标签选择模型；

Part 2 常见机器学习模型

机器学习模型：线性模型（Linear Model）

✓ Logistic Regression、Ridge regression

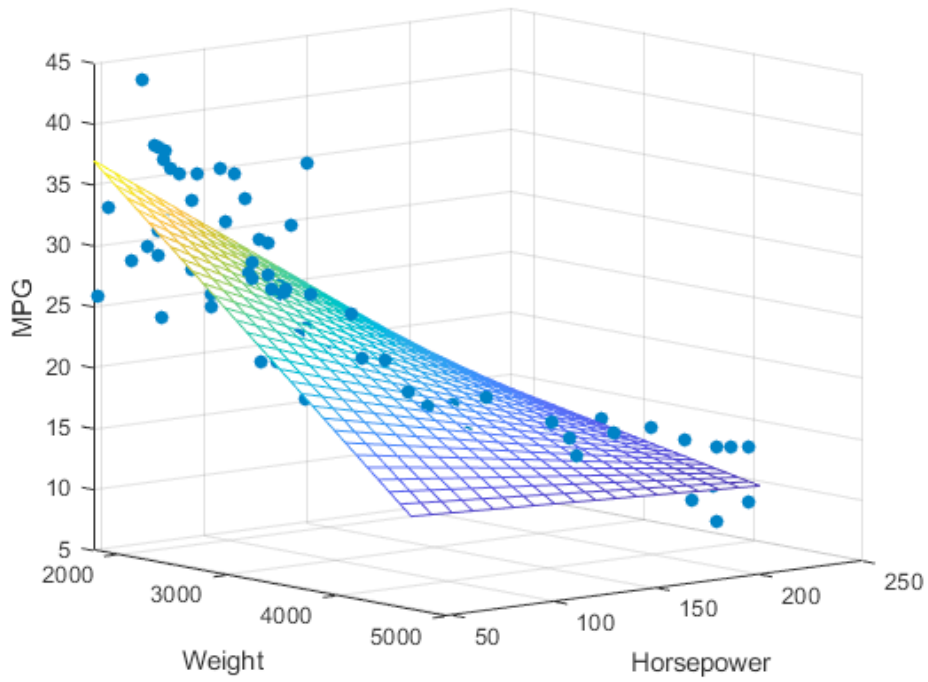
机器学习模型：SVM

✓ SVM、SVC

优点：模型简单，对可分和不可分都支持，模型可解释性强；

缺点：需要对输入进行转换和缩放，学习能力有限；

适用范围：范围较小，偏向数值类型数据；



Part 2 常见机器学习模型

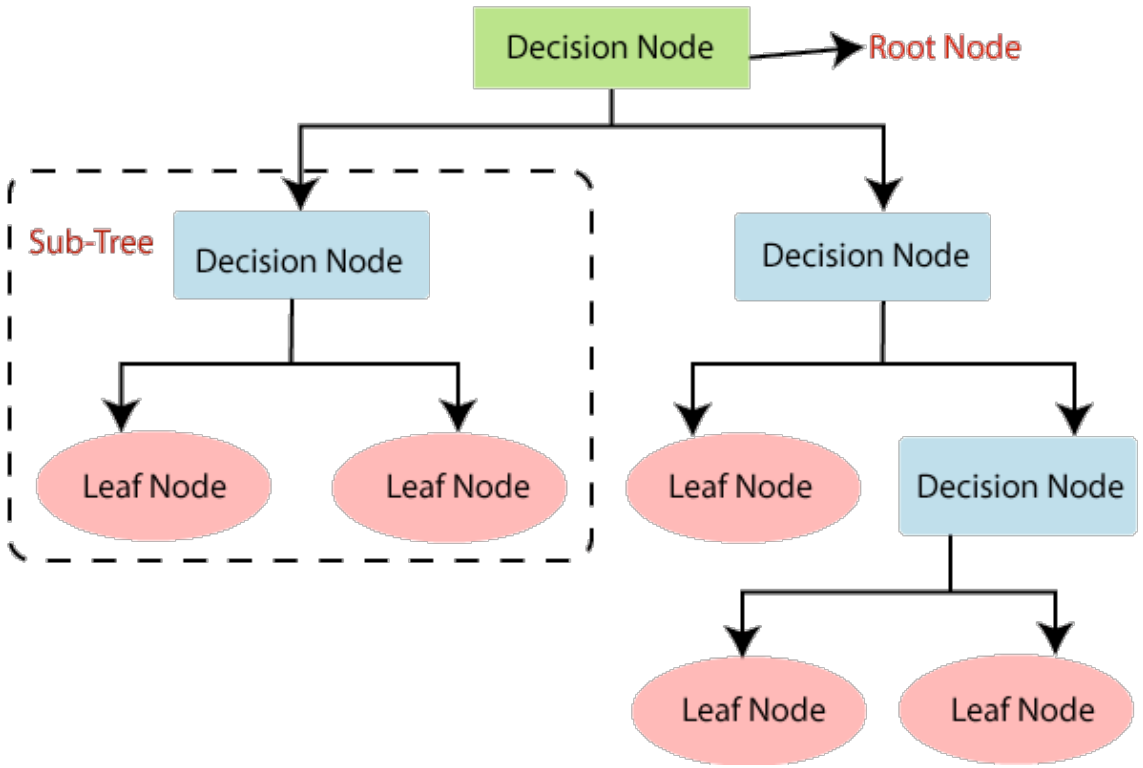
机器学习模型：树模型

✓ Decision Tree

✓ Random Forest

✓ LightGBM

✓ XGBoost



优点：对类别特征友好，精度高；

缺点：可解性差；

适用范围：范围较大，特别是类别类型的数据；

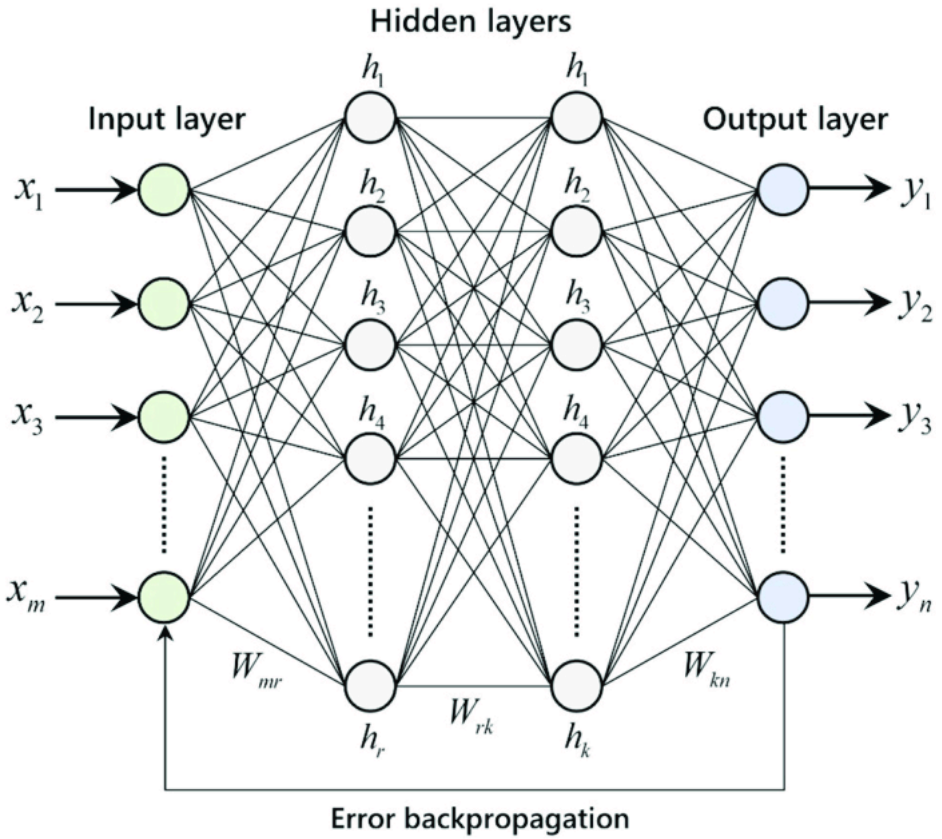
Part 2 常见机器学习模型

机器学习模型：神经网络

✓ 全连接网络

✓ 卷积神经网络

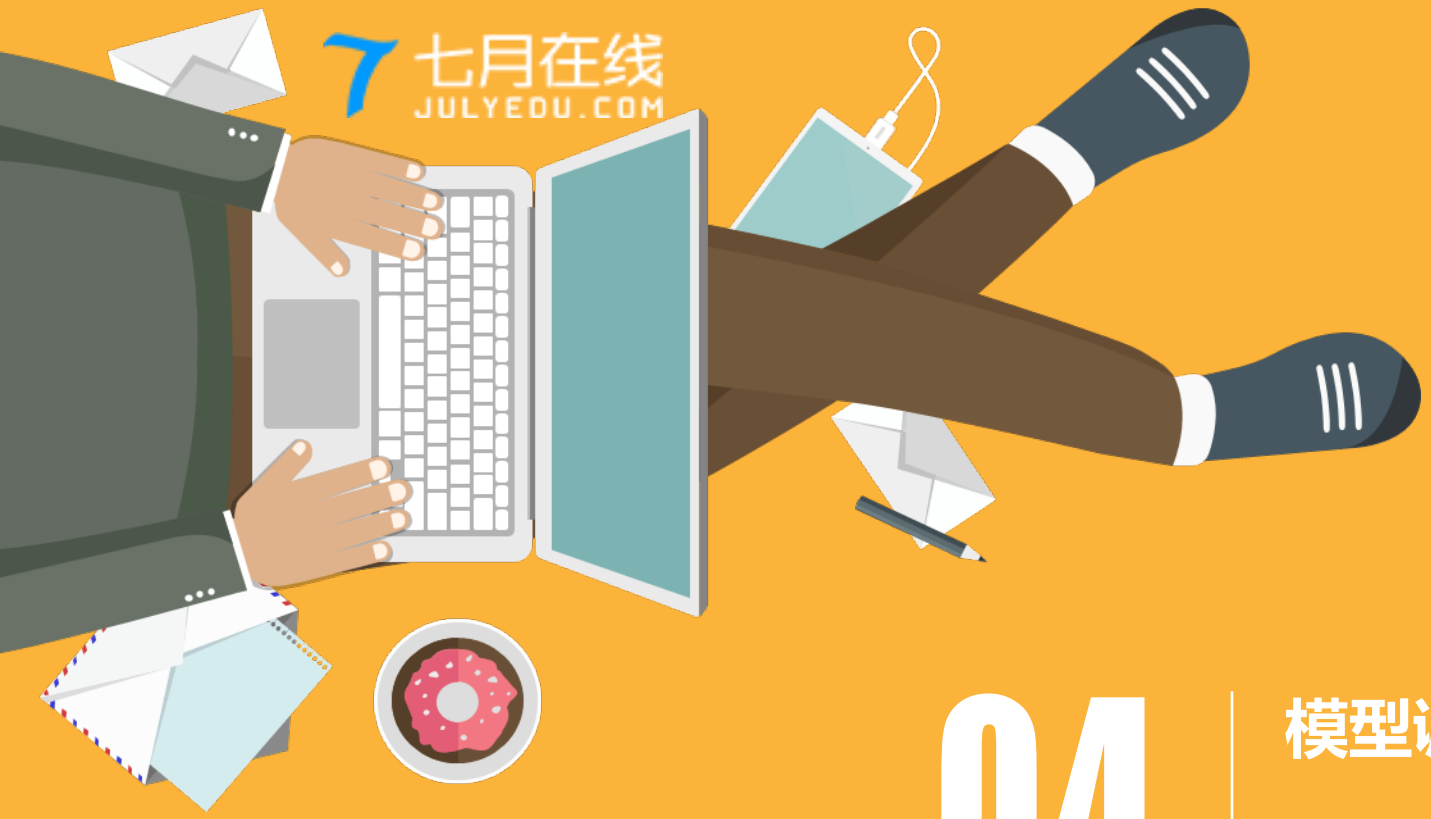
✓ 循环神经网络



优点：精度上限高，模型拟合能力强；

缺点：可解性差，容易过拟合；

适用范围：大数据及模型情况



七月在线
JULYEDU.COM

04 模型调参方法

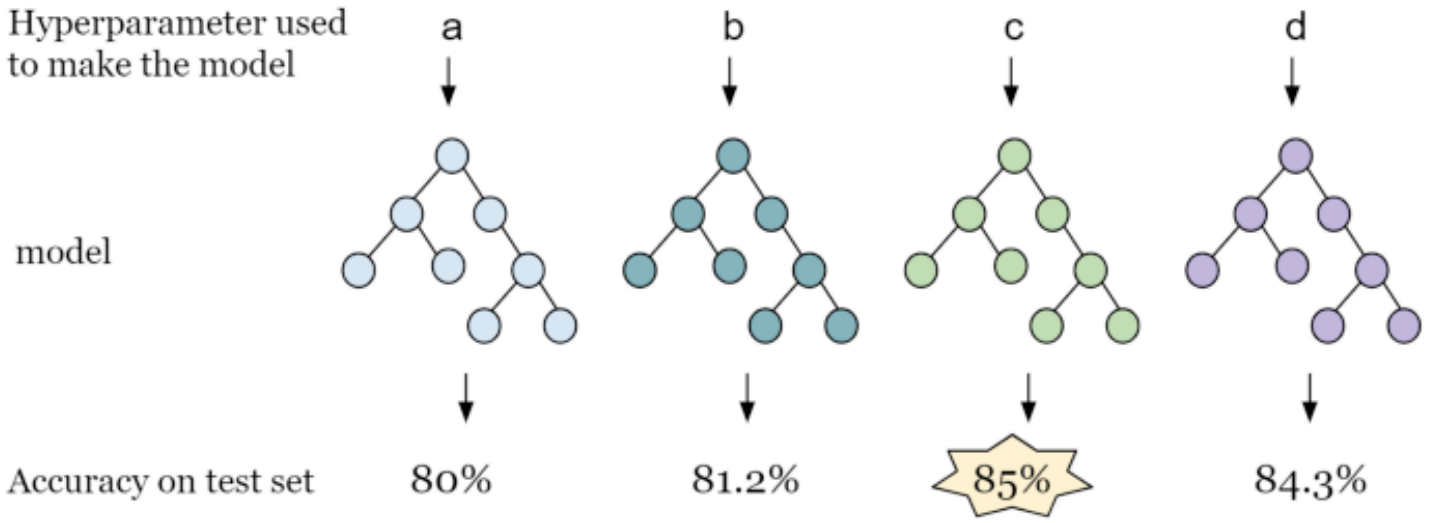
Part 4 模型调参方法

模型超参数选择：通过验证集精度选择模型参数，类似人工筛选；

✓ 优点：靠谱的方法，需要较少的计算资源；

✓ 缺点：需要人工参与和人工知识；

```
Test_Hyperparameters = [a, b, c, d]
```

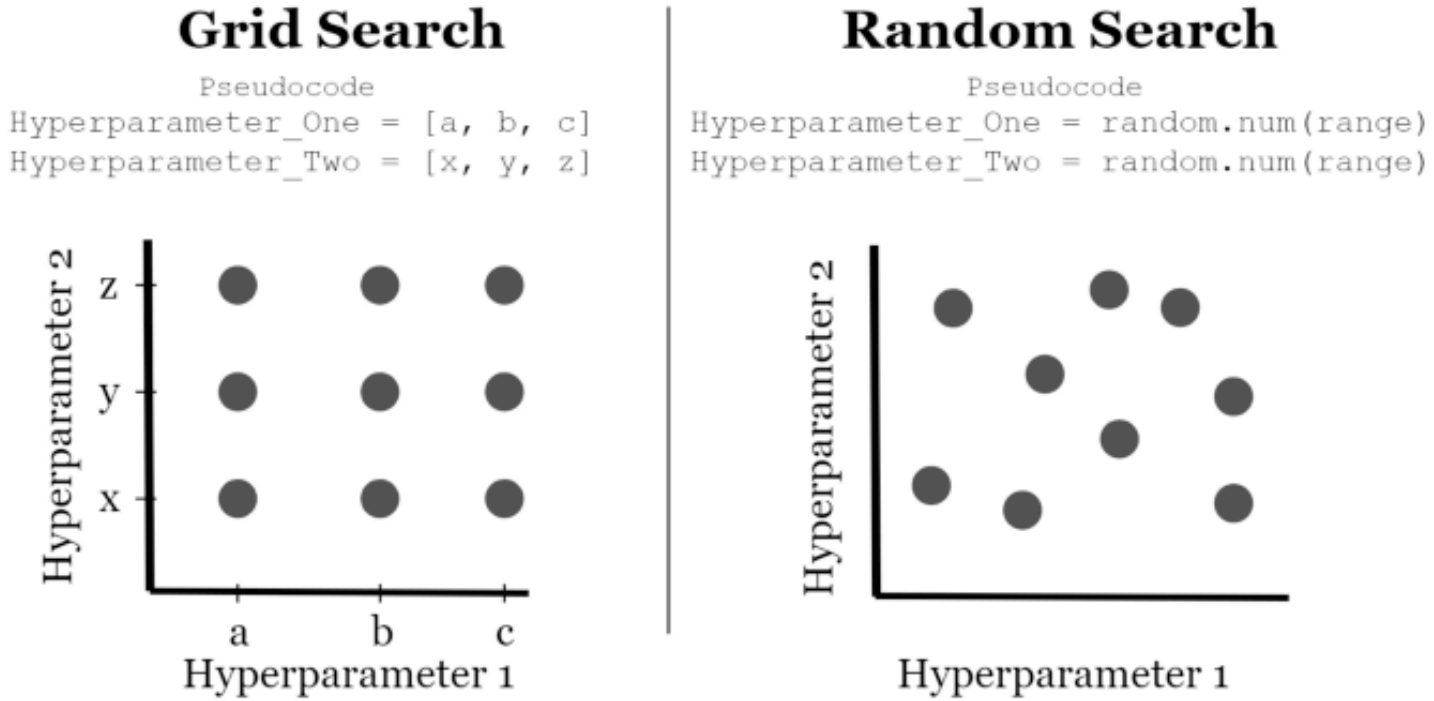


Part 4 模型调参方法

模型超参数选择：通过网格搜索和随机搜索选择参数；

✓ 优点：对参数空间进行完备的搜索；

✓ 缺点：计算量比较大；

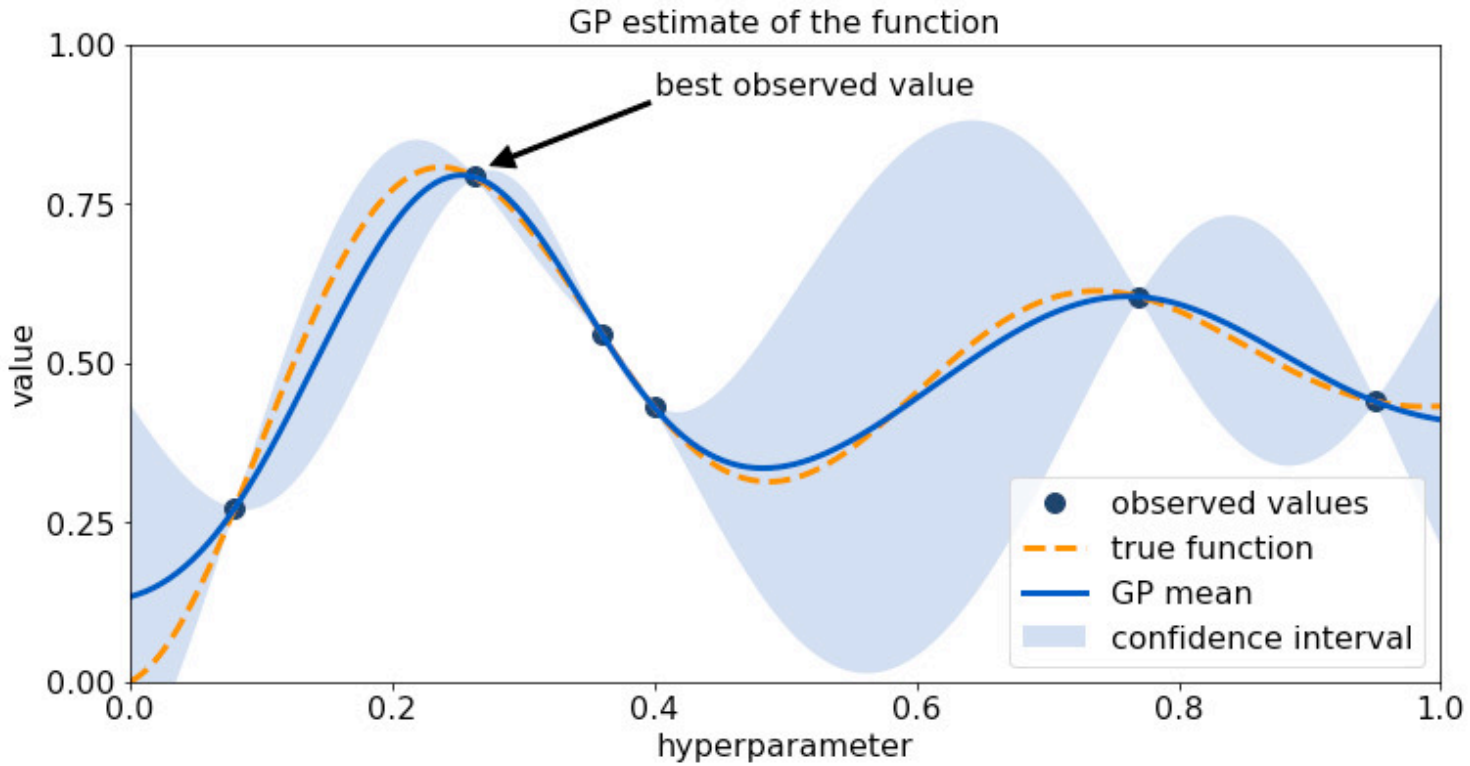


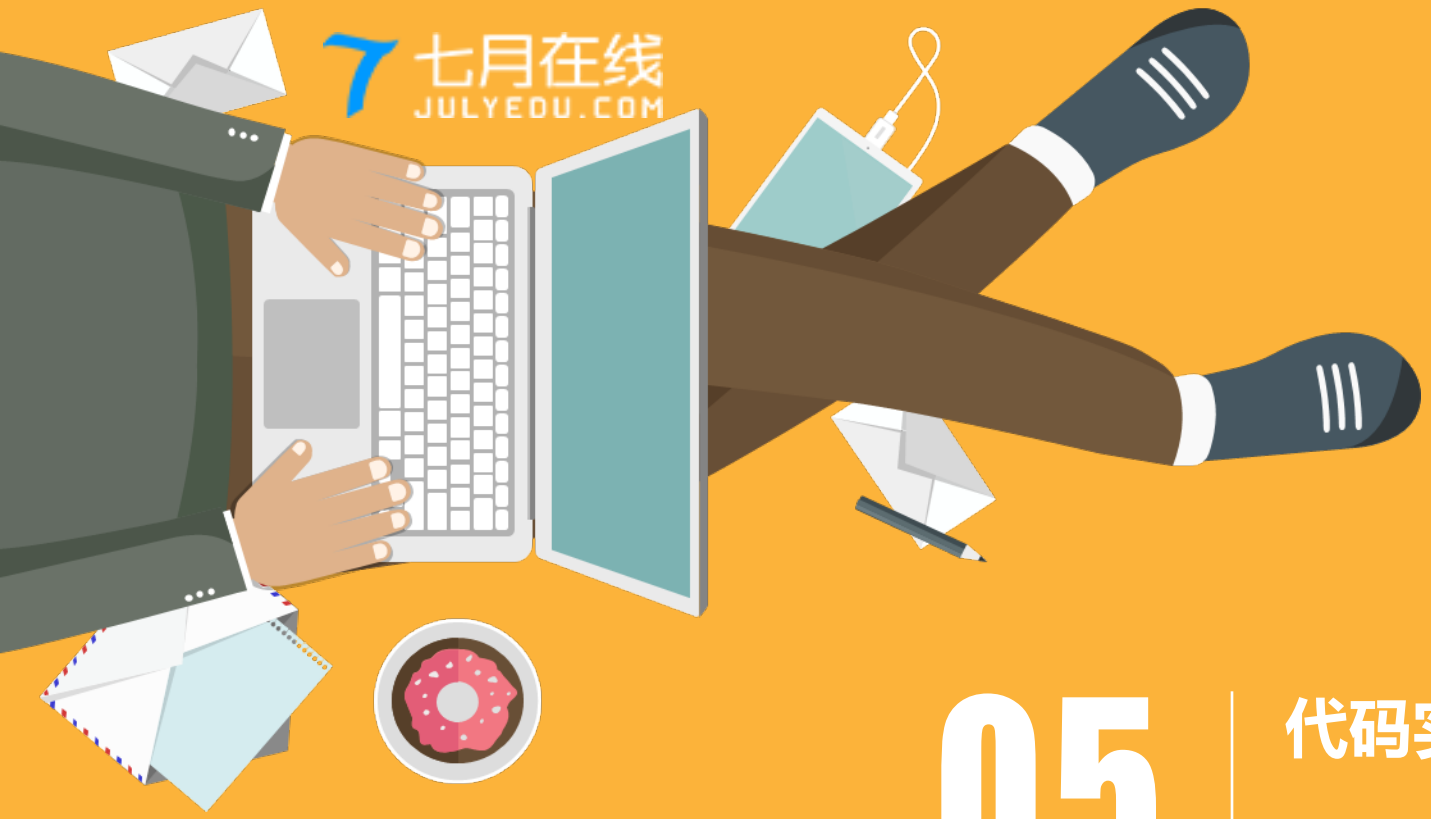
Part 4 模型调参方法

模型超参数选择：通过贝叶斯优化或遗传算法

✓ 优点：能够减少参数搜索空间；

✓ 缺点：计算量较大；





七月在线
JULYEDU.COM

05 代码实践

课堂作业

1、下面赛题适合选择哪种机器学习模型？

Severstal: Steel Defect Detection

Can you detect and classify defects in steel?

Featured • 9 months ago • Code Competition • 2431 Teams

https://www.kaggle.com/c/severstal-steel-defect-detection

Mercari Price Suggestion Challenge

Can you automatically suggest product prices to online sellers?

Featured • 2 years ago • Code Competition • 2382 Teams

https://www.kaggle.com/c/mercari-price-suggestion-challenge

课堂作业

1、下面赛题适合选择哪种机器学习模型？

Severstal: Steel Defect Detection

Can you detect and classify defects in steel?

Featured • 9 months ago • Code Competition • 2431 Teams

深度学习

https://www.kaggle.com/c/severstal-steel-defect-detection

Mercari Price Suggestion Challenge

Can you automatically suggest product prices to online sellers?

Featured • 2 years ago • Code Competition • 2382 Teams

深度学习 或 机器学习

https://www.kaggle.com/c/mercari-price-suggestion-challenge

课堂作业

2、如果模型在本地验证集精度提高了，但提交精度却下降了，是什么原因？

课堂作业

2、如果模型在本地验证集精度提高了，但提交精度却下降了，是什么原因？

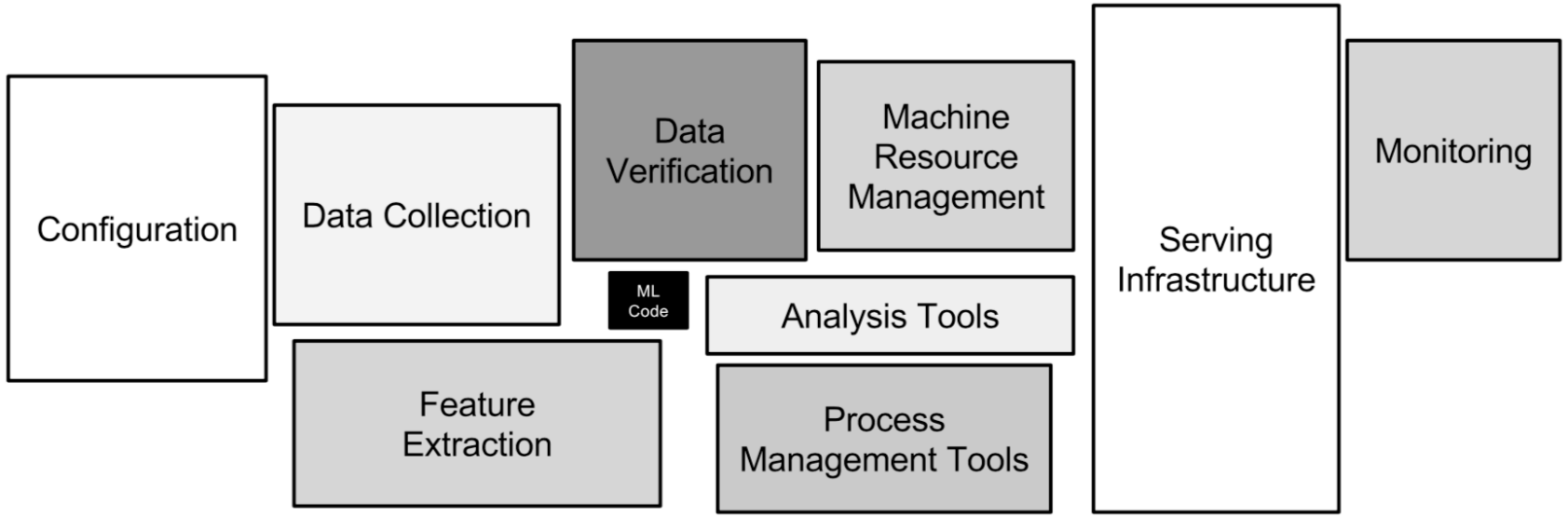
✓ 由于是通过验证集进行调参、提取特征，所以导致过拟合验证集；

✓ 验证集分布于测试集不一致；

课后作业

1、阅读下面论文，并总结你对工业届中机器学习算法的感受；

Hidden Technical Debt in Machine Learning Systems, Google



2、复现课程提供的notebook，并将最优模型的结果进行提交，截图发在群里。

https://www.kaggle.com/finlay/kaggle-ch3/

微信扫一扫关注我们

THANKS

刘老师

https://www.julyedu.com/