

The background is a grayscale photograph of a historic city, likely Cappadocia, built into a cliffside. The city features numerous cave dwellings and churches. A hot air balloon is visible in the sky on the right side.

机器学习算法讲义

体系结构介绍

石恩名

广州优亿信息科技有限公司

个人简介

石恩名

广州优亿科技有限公司高级技术经理，优亿技术布道师，曾主导公司智能爬虫系统和智能分析平台开发工作，担任过Pycon2016大会讲师，在《地理学报》等多个国内顶级期刊发表论文。现主要负责公司大数据及智能算法相关业务。专注于人工智能、数据采集、机器学习、技术架构设计。



课程目标

- （1）简要地介绍机器学习的整个体系，包括机器学习的问题分析、数据处理、模型选择及优化、算法部署和规模化应用、基础设施、产品方向。
- （2）介绍机器学习的一般流程和步骤，并用案例的形式展示出来。



机器学习体系概要介绍

机器学习结构体系

在处理机器学习相关的问题的时候我们要考虑：

- （1）问题分析。这是一个分类、回归、聚类、异常检测的问题？
- （2）数据收集与处理。如何收集数据？数据具有哪些特点？对数据进行特征清洗、填充、抽取和编码？
- （3）模型选择和优化。如何选择模型？如何优化模型？怎样调整模型的参数？如何评估一个模型的好坏？
- （4）部署与运维。如何将算法应用到生产环境中？如何让算法进行实时处理？如何将机器学习算法整合进其他的系统中，算法的接口采用SDK还是REST API？我们的基础设施能处理算法的规模吗？是使用GPU还是CPU呢？
- （5）产品化。算法产品化的方向有哪些？如何利用好各类工具？

机器学习：问题分析

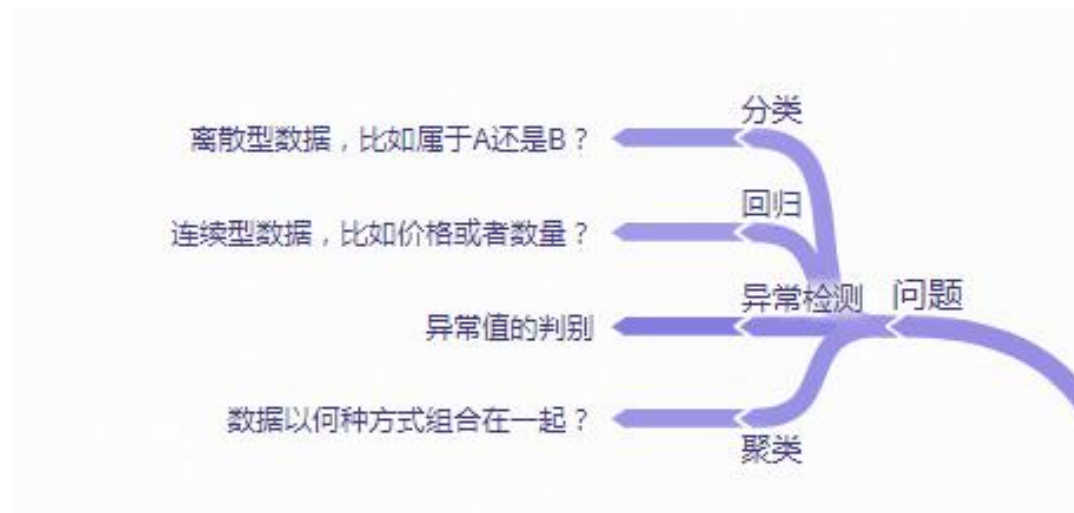
机器学习，首先需要做得是分析问题，这是一个什么类型的问题？

（1）分类问题。预测属于A还是B的问题？特点是预测数据为类别数据，即离散型数据。

（2）回归问题。预测数量和价格之类的问题？预测数据为连续型数据。

（3）异常检测。对异常值进行识别，可以利用一些统计指标进行辨别。

（4）聚类问题。探索数据以何种方式组合。





机器学习：数据与特征处理

数据和特征决定了机器学习的上限
而模型和算法只是逼近这个上限

机器学习：数据及数据处理



问题分析之后，针对问题需要进行数据采集和处理工作：

(1) 数据收集：查询数据库、编写网络爬虫等。

(2) 数据处理：数据探索、特征清洗、填充、特征工程、特征选取、特征编码、特征规范化、构建数据集等。

机器学习：数据收集

Useease Spider Man

获取帮助?

test01

任务列表

启动

停止

删除

全选	任务名称	创建时间	状态	操作
<input type="checkbox"/>	t0830155217873	2017-08-30 15:52	运行中	<div><div>■ 停止</div><div>✕ 删除</div></div>
<input type="checkbox"/>	t0830155216026	2017-08-30 15:52	运行中	<div><div>■ 停止</div><div>✕ 删除</div></div>
<input type="checkbox"/>	t0830155215669	2017-08-30 15:52	运行中	<div><div>■ 停止</div><div>✕ 删除</div></div>
<input type="checkbox"/>	t0830155214272	2017-08-30 15:52	运行中	<div><div>■ 停止</div><div>✕ 删除</div></div>
<input type="checkbox"/>	222	2017-08-29 16:25	运行中	<div><div>■ 停止</div><div>✕ 删除</div></div>
<input type="checkbox"/>	t0829123917126	2017-08-29 12:39	运行中	<div><div>■ 停止</div><div>✕ 删除</div></div>
<input type="checkbox"/>	t0829123914842	2017-08-29 12:39	运行中	<div><div>■ 停止</div><div>✕ 删除</div></div>
<input type="checkbox"/>	anjuke01	2017-08-29 10:56	运行中	<div><div>■ 停止</div><div>✕ 删除</div></div>
<input type="checkbox"/>	test001	2017-08-29 09:54	运行中	<div><div>■ 停止</div><div>✕ 删除</div></div>
<input type="checkbox"/>	t0828230102763	2017-08-28 23:01	运行中	<div><div>■ 停止</div><div>✕ 删除</div></div>
<input type="checkbox"/>	lose	2017-08-25 14:29	已停止	<div><div>▶ 启动</div><div>✕ 删除</div></div>
<input type="checkbox"/>	8yue25ri	2017-08-25 10:24	已停止	<div><div>▶ 启动</div><div>✕ 删除</div></div>
<input type="checkbox"/>	jhgblkhjk	2017-08-24 17:36	已停止	<div><div>▶ 启动</div><div>✕ 删除</div></div>

对象

pages_link @main (big...

pages_page @main (bi...

citizen_citizen @main (b...

moderna_paper @main...

开始事务

备注

筛选

排序

导入

导出

id	title	pub_date	content	origin_url	doc_vector	pure_content	Source_id	category_id	author_ic
10	Django	2017-02-23	<div id="top	(Null)	(Null)	(Null)	(Null)	4	(Null)
11	职场里传	2017-02-23	<div class="	http://www.jiar	(Null)	前些日子，星彻大	(Null)	8	2
12	当我27岁	2017-02-23	<div class="	http://www.jiar	(Null)	今年，我27岁。虽	(Null)	8	2
13	想努力却	2017-02-23	<div class="	http://www.jiar	(Null)	文/韩大爷的杂货铺	(Null)	8	2
14	我终于不	2017-02-23	<div class="	http://www.jiar	(Null)	拥有一双美腿是每	(Null)	8	2
15	升华自己	2017-02-23	<div class="	http://www.jiar	(Null)	低风险投资是一个	(Null)	8	2
16	八分钟看	2017-02-23	<div class="	http://www.jiar	(Null)	这周收到几位粉丝	(Null)	8	2
17	你那么孤	2017-02-23	<div class="	http://www.jiar	(Null)	文/傲娇哇有人说，	(Null)	8	2
18	一名3年	2017-02-23	<div class="	http://www.jiar	(Null)	来自：链接：前言	(Null)	8	2
19	11本书让	2017-02-23	<div class="	http://www.jiar	(Null)	集悦读 编辑作品集	(Null)	8	2
20	考研那些	2017-02-23	<div class="	http://www.jiar	(Null)	考研的学习是件实	(Null)	8	2
21	我忍你很	2017-02-23	<div class="	http://www.jiar	(Null)	有一种朋友，	(Null)	8	2
22	一个实验	2017-02-23	<div class="	http://www.jiar	(Null)	文字和语言表达都	(Null)	8	2
23	工作六年	2017-02-23	<div class="	http://www.jiar	(Null)	在银行工作了六年，	(Null)	8	2
24	黛玉晚报	2017-02-23	<div class="	http://www.jiar	(Null)	作者：互联网使我	(Null)	8	2
25	就是教你	2017-02-23	<div class="	http://www.jiar	(Null)	相信没有人会否认	(Null)	8	2
26	[下班后怎	2017-02-23	<div class="	http://www.jiar	(Null)	每天我们都会看到	(Null)	8	2
27	又是董卿	2017-02-23	<div class="	http://www.jiar	(Null)	《中国诗词大会第二	(Null)	8	2
28	50 万负	2017-02-23	<div class="	http://www.jiar	(Null)	笔者是个创业者，	(Null)	8	2
29	你不站在	2017-02-23	<div class="	http://www.jiar	(Null)	1.《何以笙箫默》里	(Null)	8	2
30	大一新鲜	2017-02-23	<div class="	http://www.jiar	(Null)	我知道你们刚刚度	(Null)	8	2

SELECT *, _ROWID_ "NAVICAT_ROWID" FROM "moderna_paper" LIMIT 0, 1000

第 10 条记录 (共 30 条) 于

机器学习：数据处理

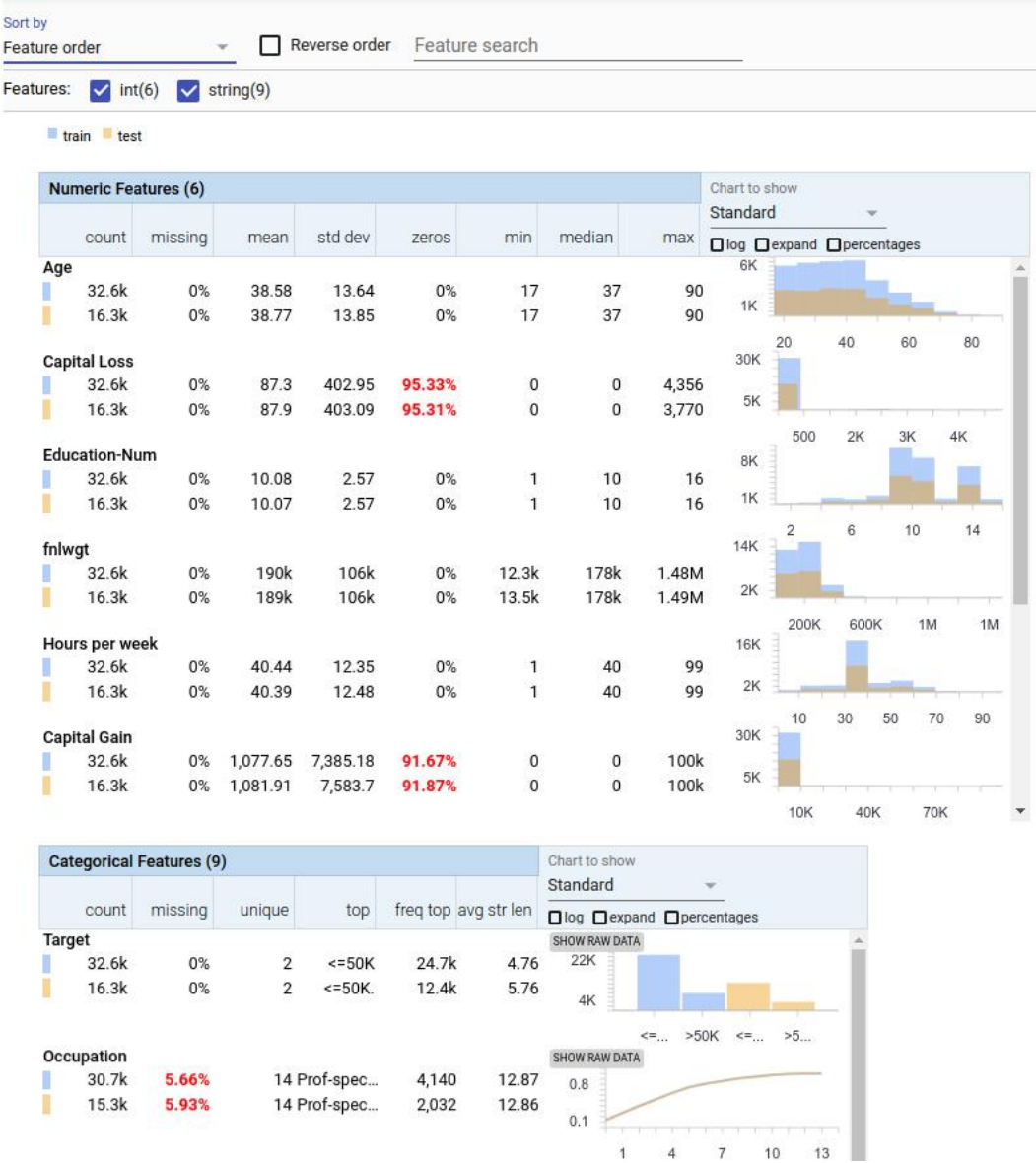
- (1) 数据探索：分析变量之间的关系、数据分布特征
- (2) 特征清洗：对数据中的特殊值、异常值进行处理
- (3) 特征填充：对缺失数据进行填充，比如均值填充、平滑填充、热启动、冷启动
- (4) 特征工程：通过特定的方法构建新的特征变量，比如数据变换、交叉合并。
- (5) 特征选取：通过对所有的变量进行筛选和评估，选择或构建有效特征。
- (6) 特征编码：由于机器学习的模型都是进行数学计算，因此需要将非数值型变量进行编码和处理变成数值型数据，常用的特征编码包括标签编码、one-hot编码等。
- (7) 特征标准化与归一化：。
- (8) 构建数据集：训练数据、测试数据、验证数据、交叉验证

机器学习：数据探索性分析

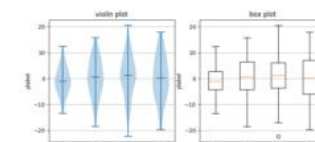
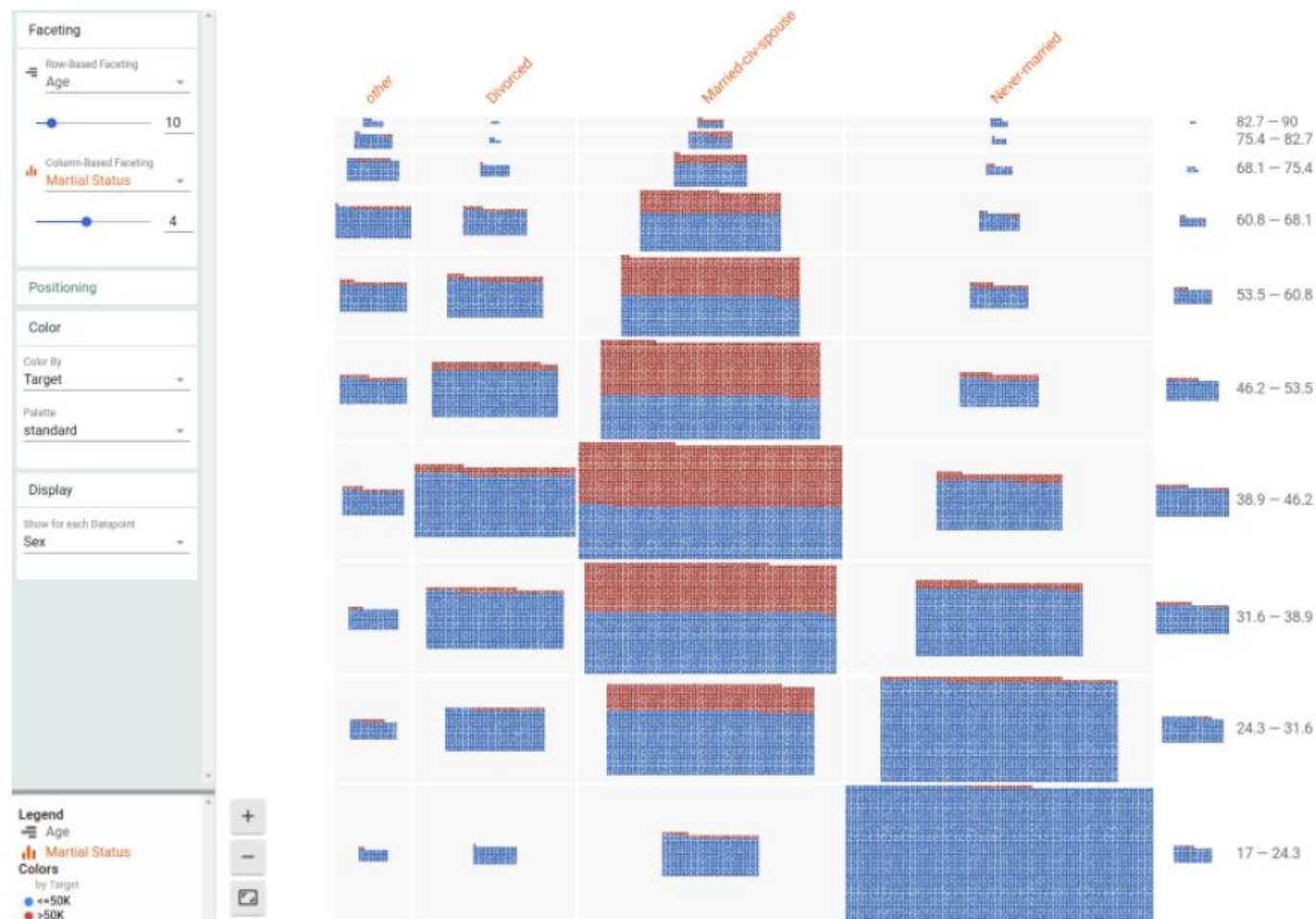
探索性数据分析，主要通过一系列可视化手段的辅助让数据分析师最大程度得到数据的直觉、理解数据、发掘潜在的结构。

（1）单变量分析：均值，中位数，绝对值，最小最大值，区域，分位数，IQR,方差，标准差，斜率，直方图，箱图。

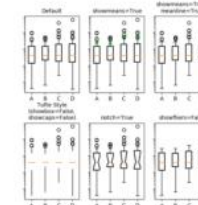
（2）双变量分析：散点图、相关热力图、堆叠柱状图。



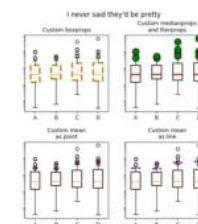
机器学习：数据探索性分析



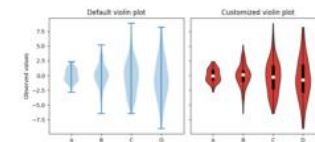
boxplot_vs_violin_demo



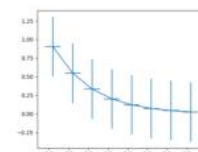
bxp_demo



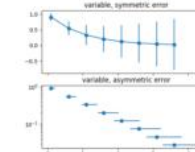
bxp_demo



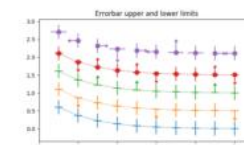
customized_violin_demo



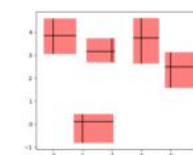
errorbar_demo



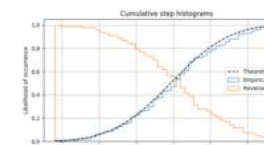
errorbar_demo_features



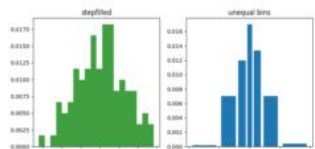
errorbar_limits



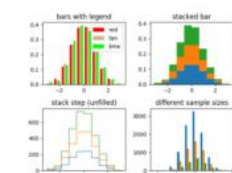
errorbars_and_boxes



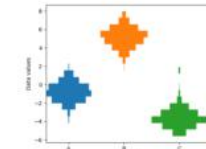
histogram_demo_cumulative



histogram_demo_histtypes



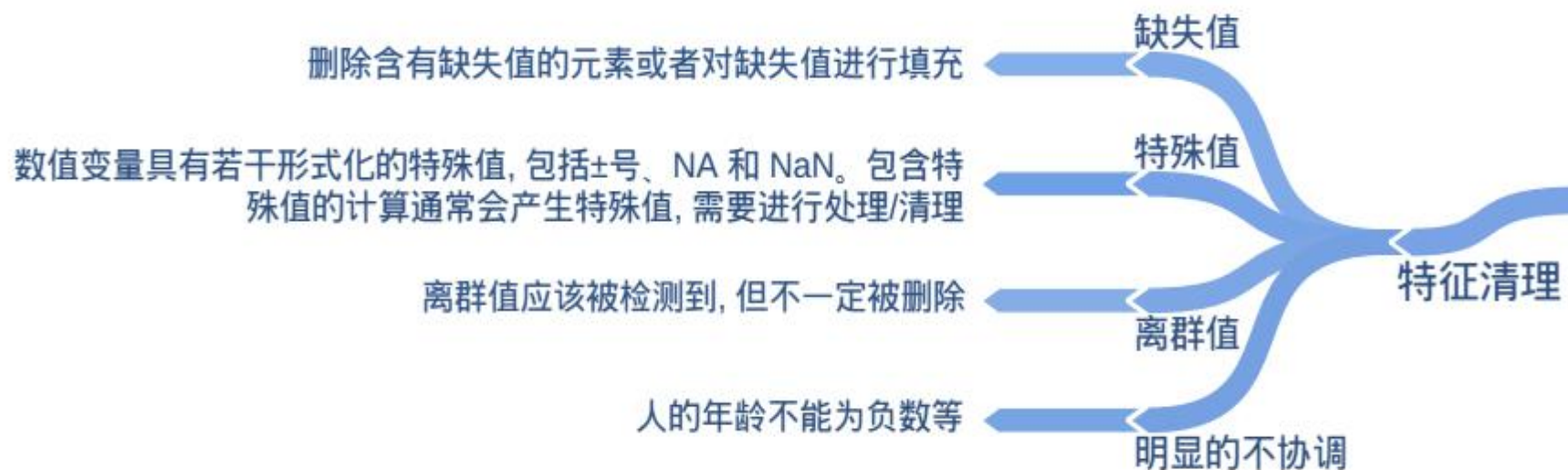
histogram_demo_multihist



multiple_histograms_side_by_side

机器学习：特征清理

- (1) 缺失值：删除含有缺失值的元素或者对缺失值进行填充。
- (2) 特殊值：数值变量具有若干形式化的特殊值, 包括±号、NA 和 NaN。包含特殊值的计算通常会产生特殊值, 需要进行处理/清理。
- (3) 离群值：离群值应该被检测到, 但不一定被删除。
- (4) 明显的不协调：人的年龄不能为负数等



机器学习：特征填充

- (1) 均值填充：计算出该类型无缺失值案例的均值进行填充。
- (2) 回归填充：基于一个变量对另一个变量进行线性回归，通过线性关系填充缺失值。
- (3) 热启动：对于缺失值随机填充其他案例中的值或者寻找相似的案例的值。（同一数据集）
- (4) 冷启动：从另一个数据集选择案例的值来填补。（不同数据集）

机器学习：特征工程

- (1) 离散化：将连续特征进行分段，如将一天几部分（早中晚）。
- (2) 少数类别合并：有些类别特征可以合并，特别是有些类别只有很少的样本。
- (3) 交叉合并：从已有的特征中新建新的特征，可以是乘上数值型特征或者合并类别特征。
- (4) 数据变换：通过特定的函数（log、指数函数）对现有特征进行处理，构建新的衍生变量。

机器学习：特征选择

特征选择主要是对数据属性进行选择 and 抽取，主要方法包括降维、重要性排序。

(1) 降维：主成分分析（PCA）、奇异值分解（SVD）等。

(2) 重要性排序：过滤式方法(Filter)、包裹式方法(Wrapper)、嵌入式方法(Embedded)。

- 过滤式方法：通过预测变量计算相关度的矩阵中选择特征，即计算自变量和预测变量的相关性对变量进行筛选。常用的方法包括：相关系数、卡方检验、ANOVA方差分析等。
- 包裹式方法：通过目标函数来决定是否加入变量。常用的方法包括：启发式搜索的前向选择法、后向选择法、随机搜索遗传算法等。
- 嵌入式方法：特征选择算法本身作为组成部分嵌入到学习算法中，比如常用的L1正则化和L2正则化。

机器学习：特征标准化与归一化

为什么需要标准化和归一化？

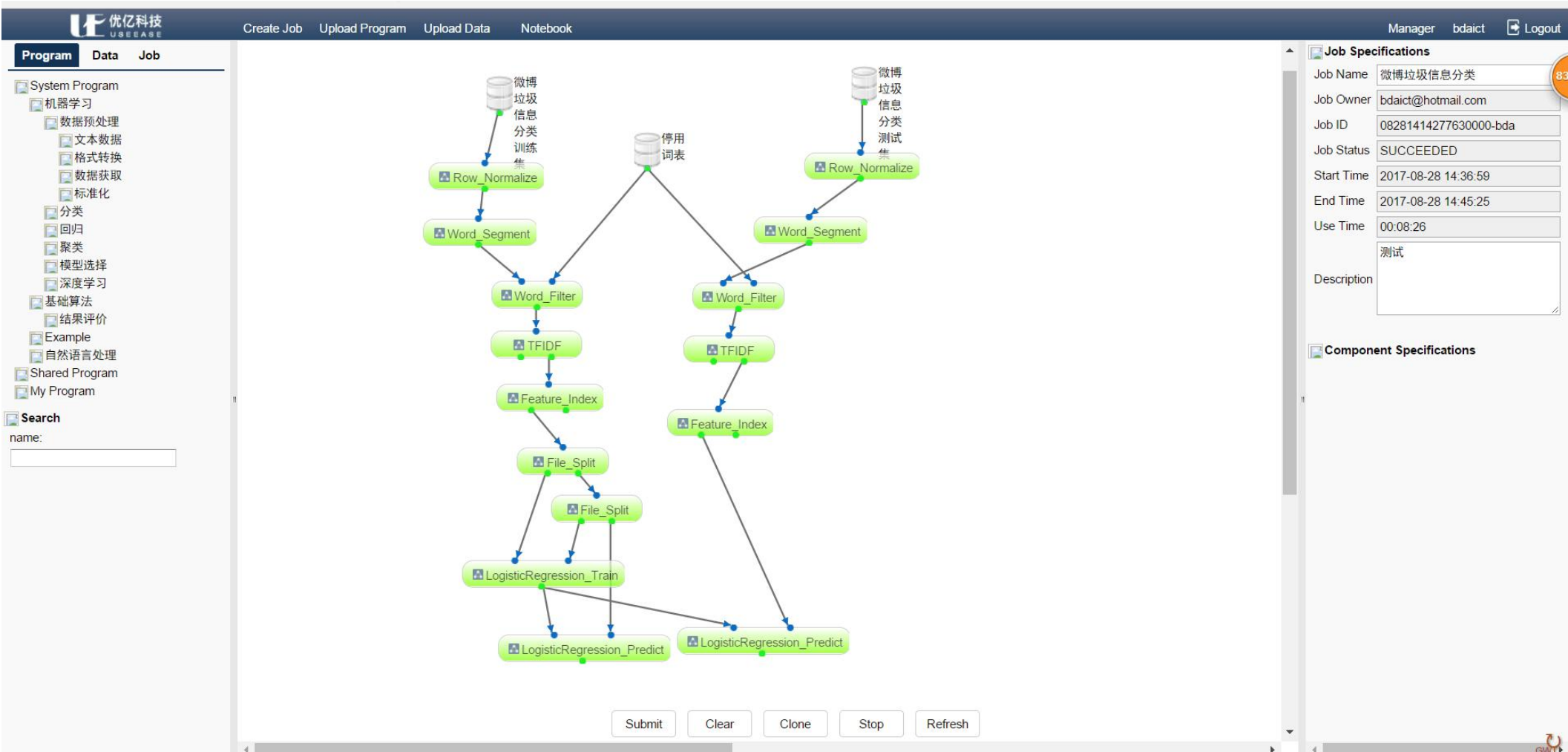
（1）对于一些机器学习算法中，在数据没有规范化的情况下，其目标函数不能较好地发挥作用，例如SVM，在各个维度进行不均匀伸缩后，最优解与原来不等价。

（2）使用梯度下降的算法，在规范化之后梯度下降法的收敛速度会更快。

机器学习：构建数据集

- (1) 训练数据集。主要用于学习模型参数的数据集集合。
- (2) 验证数据集。主要用于调整分类器参数的数据集集合。
- (3) 测试数据集。主要用于评估训练好的分类器的数据集集合。

案例：文本分类



The background image is a dark, atmospheric photograph of a city, likely Cappadocia, featuring a prominent rock formation on the left and a hot air balloon floating in the sky on the right. A horizontal green band is overlaid across the middle of the image, containing the title text.

机器学习：模型选择与调优

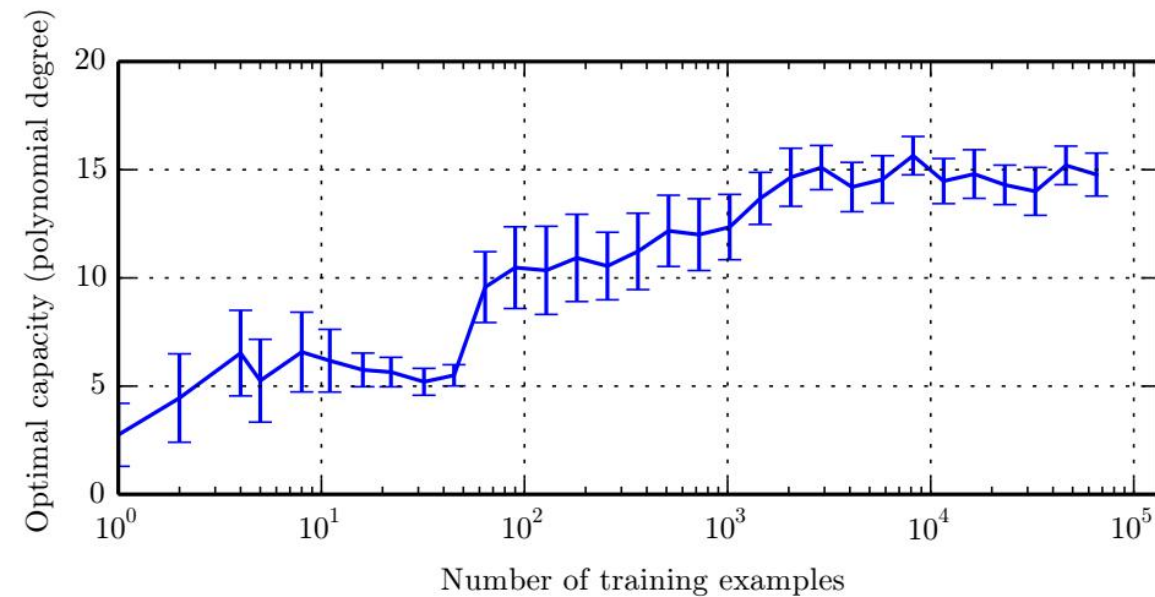
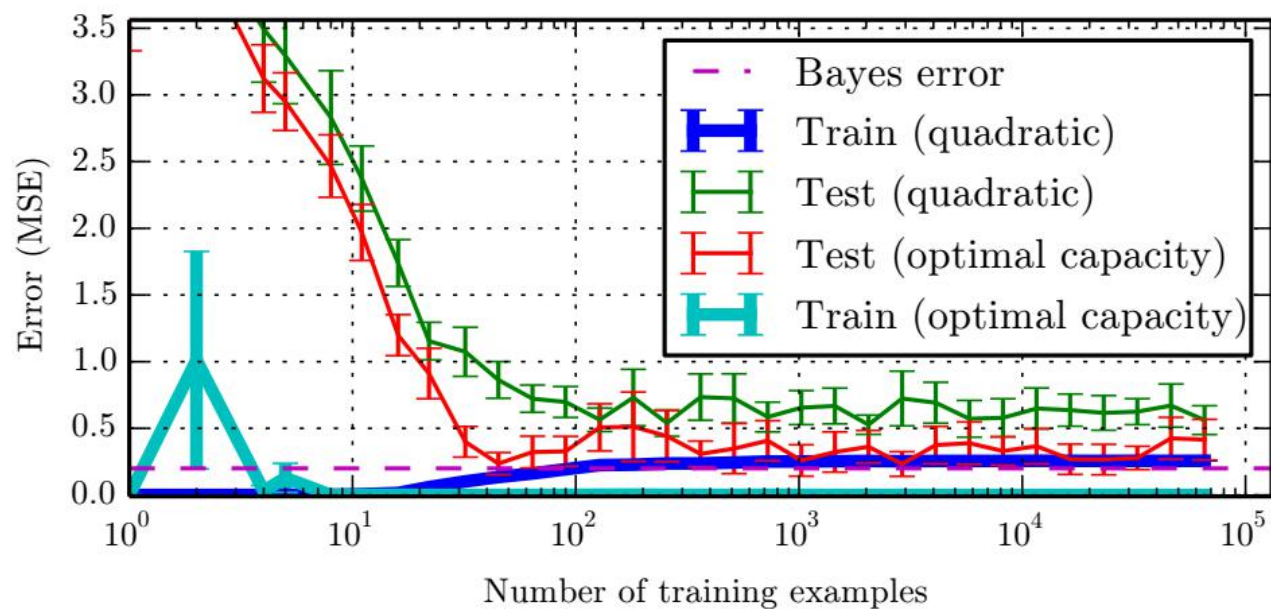
机器学习：常用的模型

- (1) 贝叶斯：朴素贝叶斯、贝叶斯网络。
- (2) 决策树：ID3、C4.5、随机森林。
- (3) 回归：逻辑回归、岭回归(L2)、LOSS回归（L1）。
- (4) 聚类：k均值、层次聚类、高斯混合模型(GMM)。
- (5) 关联分析。
- (6) 表示学习：word2vec。
- (7) 支持向量机。
- (8) 神经网络。

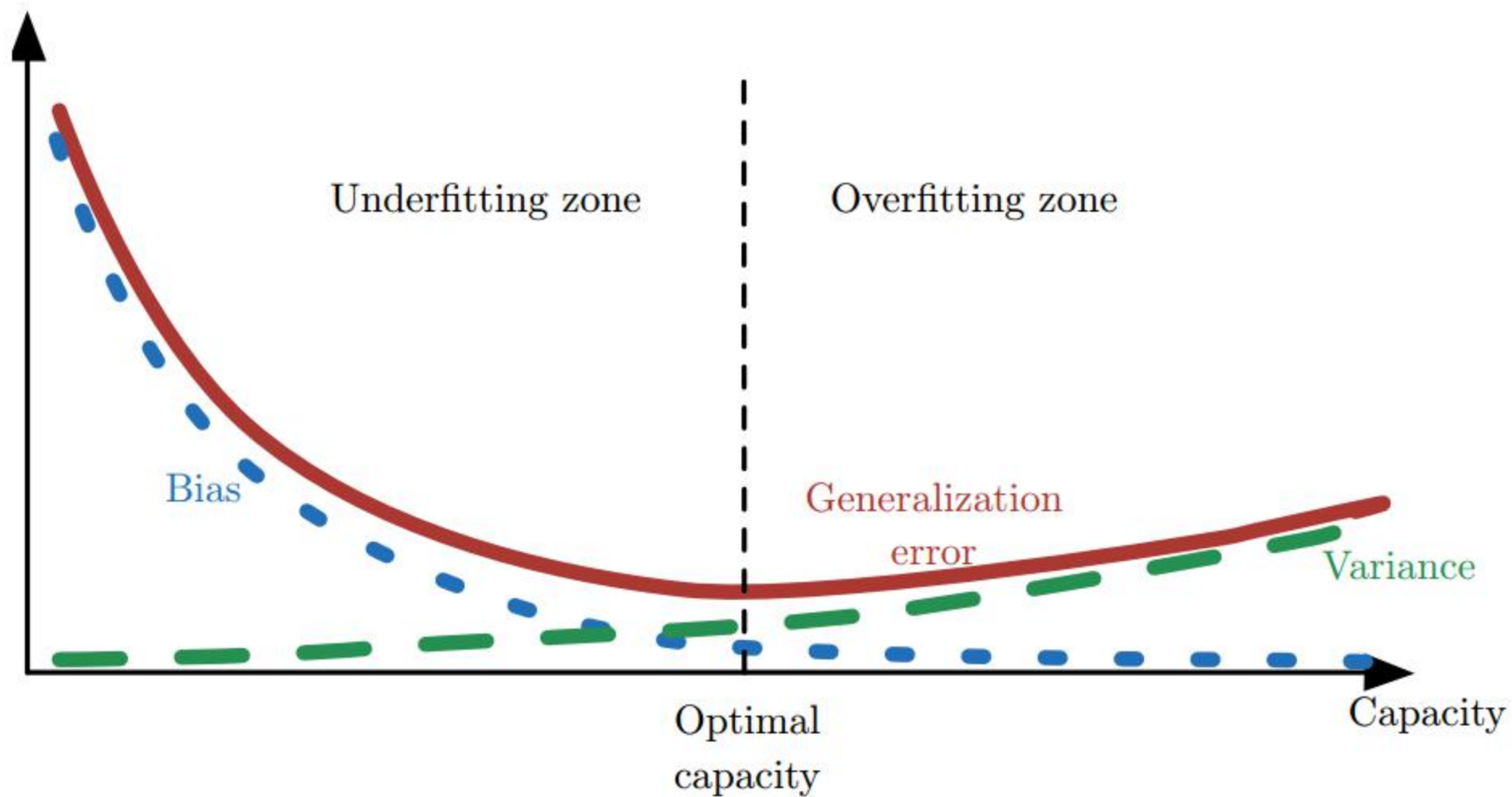
机器学习：没有免费午餐定理

机器学习的 没有免费午餐定理（no free lunch theorem）表明，在所有可能的数据生成分布上平均之后，每一个分类算法在未事先观测的点上都有相同的错误率。换言之，在某种意义上，没有一个机器学习算法总是比其他的要好。我们能够设想的最先进的算法和简单地将所有点归为同一类的简单算法有着相同的平均性能（在所有可能的任务上）。

机器学习：没有免费午餐定理



机器学习：过拟合与欠拟合




机器学习：调优方法

- (1) 交叉验证法。
- (2) 超参数搜索。
- (3) 提前中止。
- (4) Bootstrap。
- (5) Bagging。
- (6) 数据集增强。

在交叉验证法的一个回合中，将数据分成几个子集，使用一个子集来分析模型（训练集），另一个用来验证模型优劣（验证集或测试集）。为了减少模型的变化性，在交叉验证的多个回合中，其子集的分组都是不一样的，最后的验证结果取多个回合的平均。



机器学习：模型评估

		True condition			
Total population		Condition positive	Condition negative	Prevalence $= \frac{\Sigma \text{Condition positive}}{\Sigma \text{Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{True positive} + \Sigma \text{True negative}}{\Sigma \text{Total population}}$
Predicted condition	Predicted condition positive	True positive	False positive (Type I error)	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{True positive}}{\Sigma \text{Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{False positive}}{\Sigma \text{Predicted condition positive}}$
	Predicted condition negative	False negative (Type II error)	True negative	False omission rate (FOR) = $\frac{\Sigma \text{False negative}}{\Sigma \text{Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{True negative}}{\Sigma \text{Predicted condition negative}}$
		True positive rate (TPR), Recall, Sensitivity, probability of detection $= \frac{\Sigma \text{True positive}}{\Sigma \text{Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{False positive}}{\Sigma \text{Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR-}}$ $F_1 \text{ score} = \frac{2}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{False negative}}{\Sigma \text{Condition positive}}$	True negative rate (TNR), Specificity (SPC) $= \frac{\Sigma \text{True negative}}{\Sigma \text{Condition negative}}$	Negative likelihood ratio (LR-) $= \frac{\text{FNR}}{\text{TNR}}$	



机器学习：部署与运维

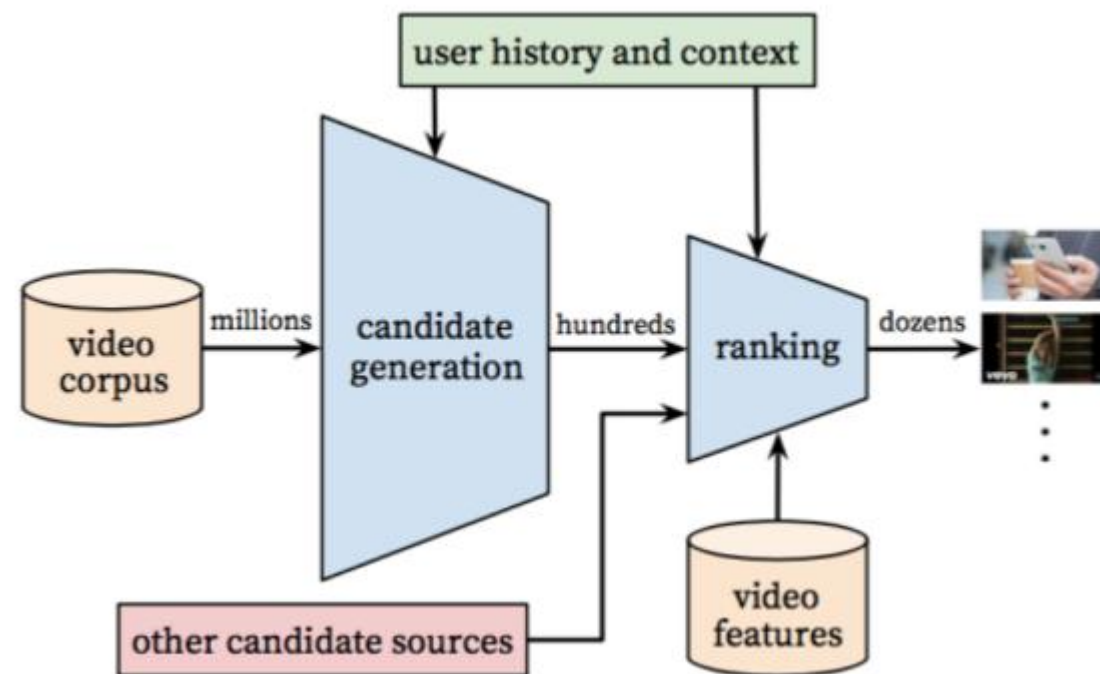
机器学习：模型的部署

部署到实际生产环境我们需要考虑的问题：

- （1）如何将算法应用到生产环境中？（生产环境的特殊性）
- （2）如何让算法进行实时处理？
- （3）如何将机器学习算法整合进其他的系统中，算法的接口采用SDK还是REST API？
- （4）我们的基础设施能处理算法的规模吗？是使用GPU还是CPU呢？

机器学习：生产环境模型设计

youtube的推荐系统设计（10亿量级的用户），整个推荐系统分为Candidate Generation和Ranking两个阶段(和淘宝的推荐系统架构类似，淘宝内部叫做Matching和Ranking)。Candidate Generation阶段通过i2i/u2i/u2u/user profile等方式“粗糙”的召回候选商品，Candidate Generation阶段视频的数量是百级别了；Ranking阶段对Candidate Generation后的视频采用更精细的特征计算user-item之间的排序分，作为最终输出推荐结果的依据。





Thanks for
your attention!