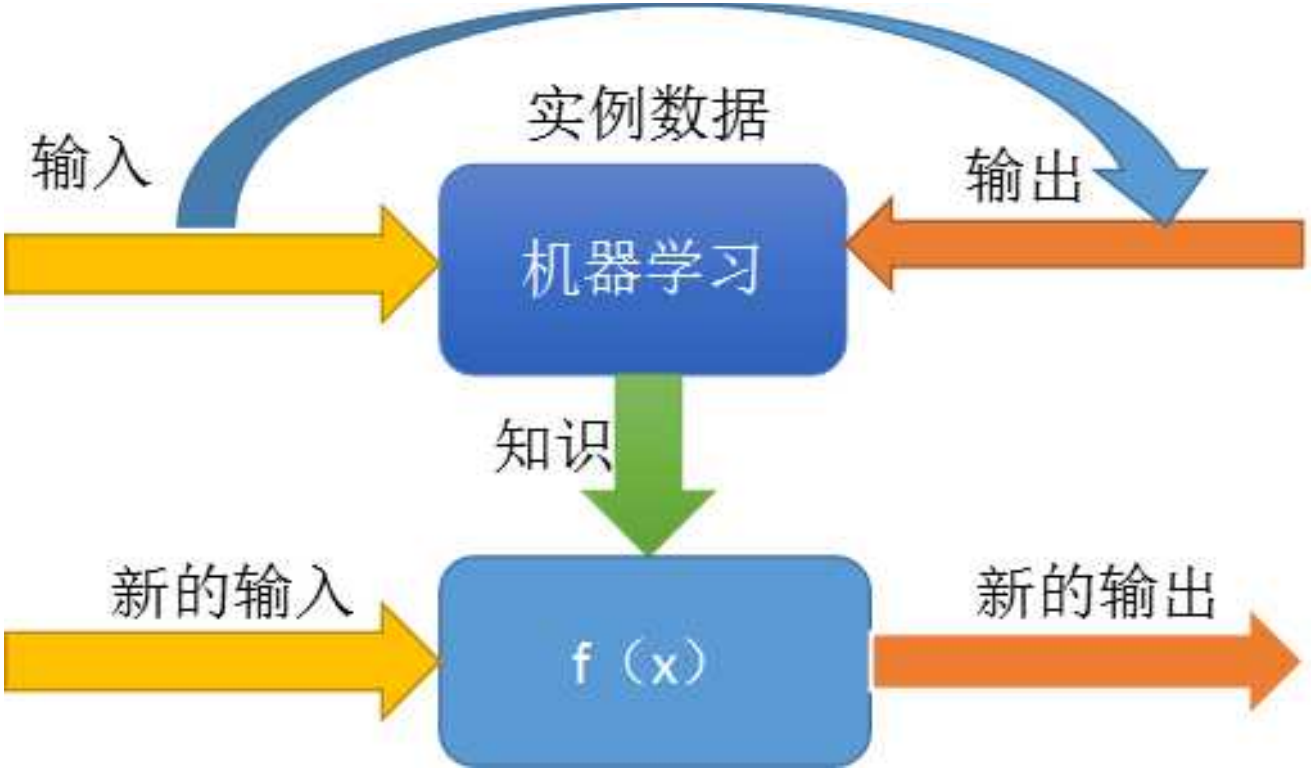
机器学习入门

UNIX2GO.com

目录

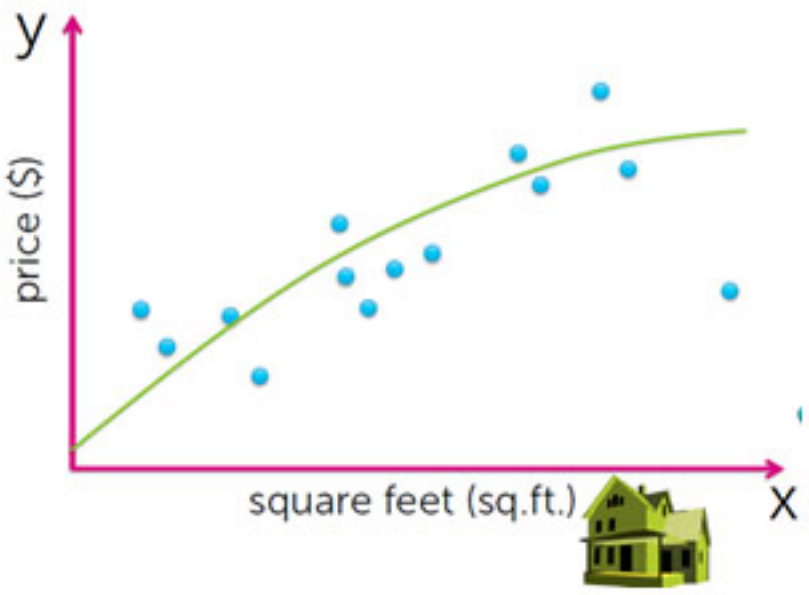
（一）何为机器学习



简单说，机器通过一系列「任务」从「经验」（数据）中学习，并且评估「效果」如何。

样本、特征、标签

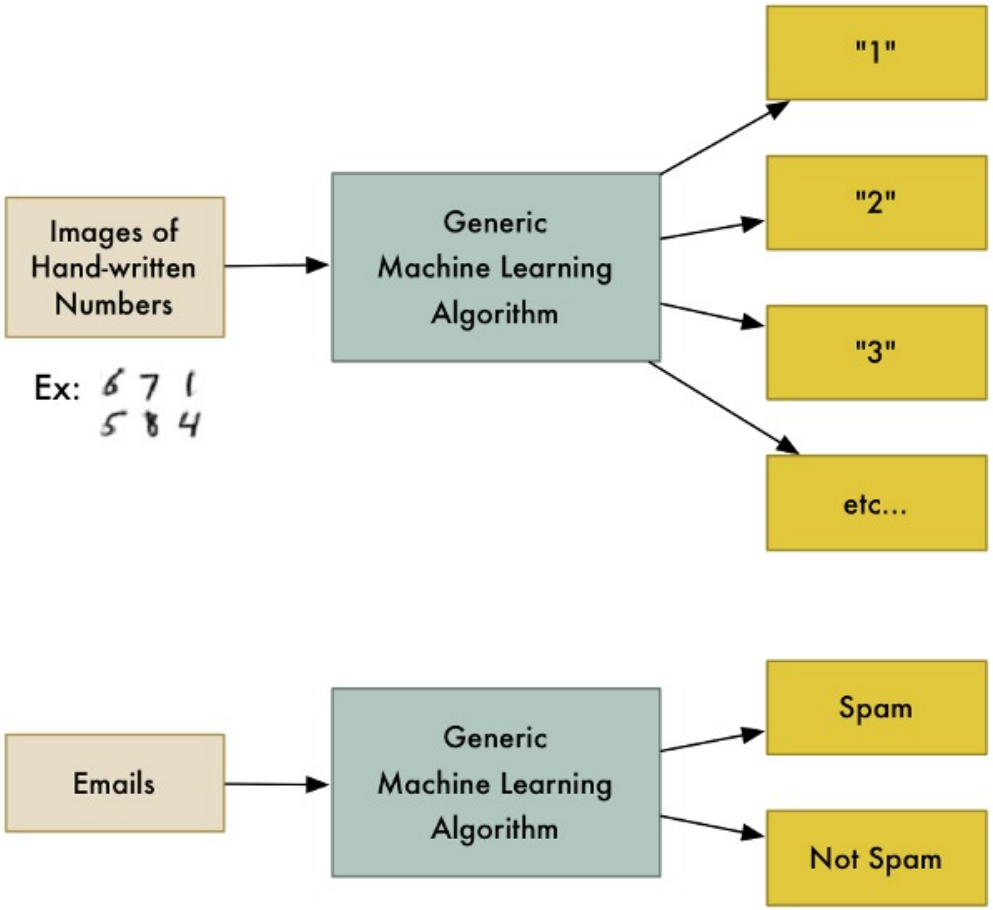
面积（平米）	朝向	居室	价格（w）
105	South	4	300
160	East	6	450



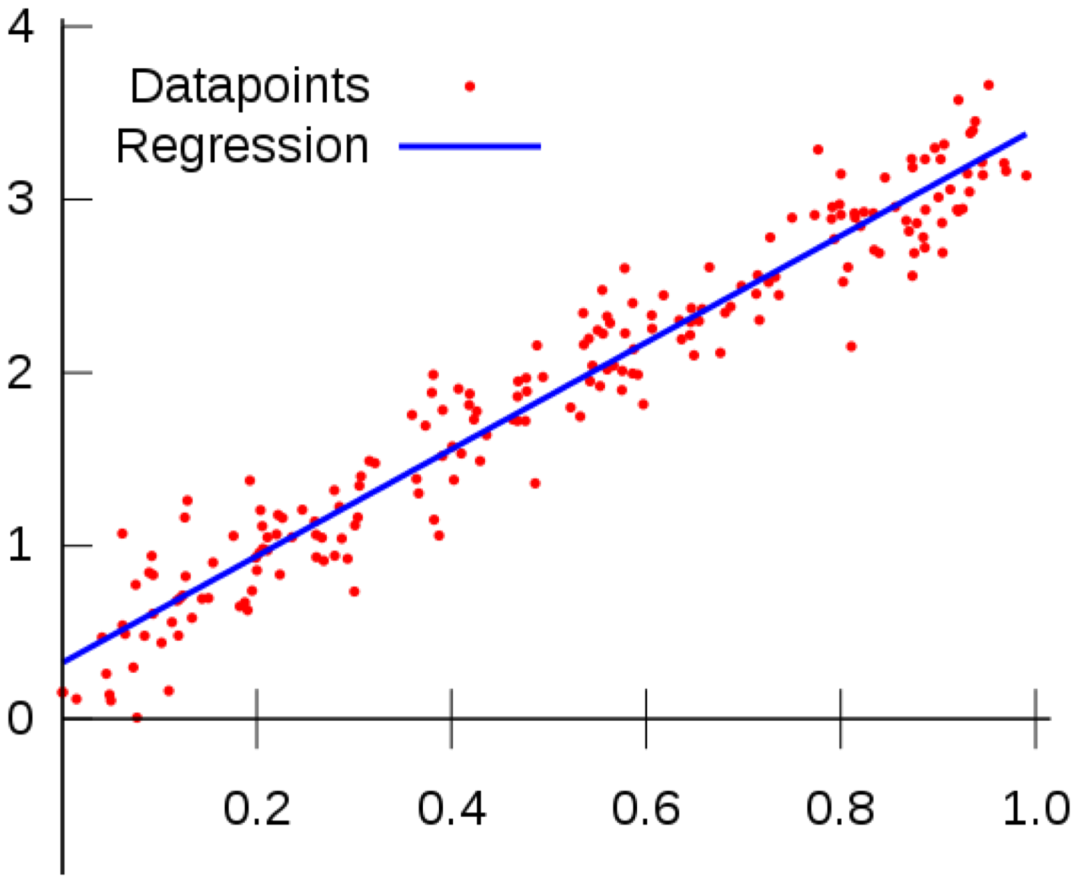
（二）机器学习用途

- Classification 分类，如垃圾邮件识别(detection, ranking)
- Regression 回归，例如股市预测
- Clustering 聚类，如 iPhoto 按人分组

分类


Images of Hand-written Numbers
Ex:
Generic Machine Learning Algorithm
"1"
"2"
"3"
etc...
Emails
Generic Machine Learning Algorithm
Spam
Not Spam

回归



聚类

Clustering:

- Clustering is the task of gathering samples into groups of similar samples according to some predefined similarity or dissimilarity measure

sample

Cluster/group

（三）机器学习类型

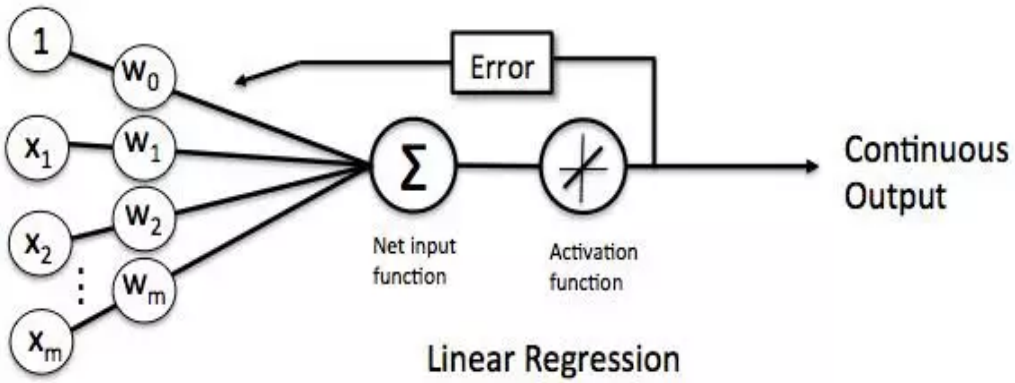
- 监督学习：给出定义好的标签，程序「学习」标签和数据之间的映射关系
- 无监督学习：在没有标签的数据集上进行学习

监督学习案例：逻辑回归

- 目标：预测美女是否有男友
- 样本：

身高（厘米）	外貌（分数）	独自逛街次数	是否有男友
165	80	5	1
158	65	20	0

监督学习案例：逻辑回归



- 计算公式： $y = f\left(h(x)\right) = f\left(w^T \cdot x + b\right)$，其中f(z)为激活函数。
- y是函数输出结果，w是权重参数集，x是特征向量，b是偏移量。
- b在实际中也当作一项特殊参数处理（w_0），它对应的输入值x始终是1。

监督学习案例：逻辑回归

接上，得到公式：

$$y = sigmoid(w_1 * 身高 + w_2 * 外貌 + w_3 * 独自逛街次数 + b)$$

sigmoid是激活函数，返回值在[0,1]。

假设模型训练后，得到的参数集：

W1	W2	W3	b
0.6	0.7	-0.9	-140

监督学习案例：逻辑回归

那么，开始预测：

假设某美女身高165，外貌75分，独自逛街次数15，套用公式一算：

- 165*0.6 + 75*0.7 + 15*(-0.9) + (-140) = -2.0
- sigmoid(-2.0) = 0.1192 （大概率没有男友）

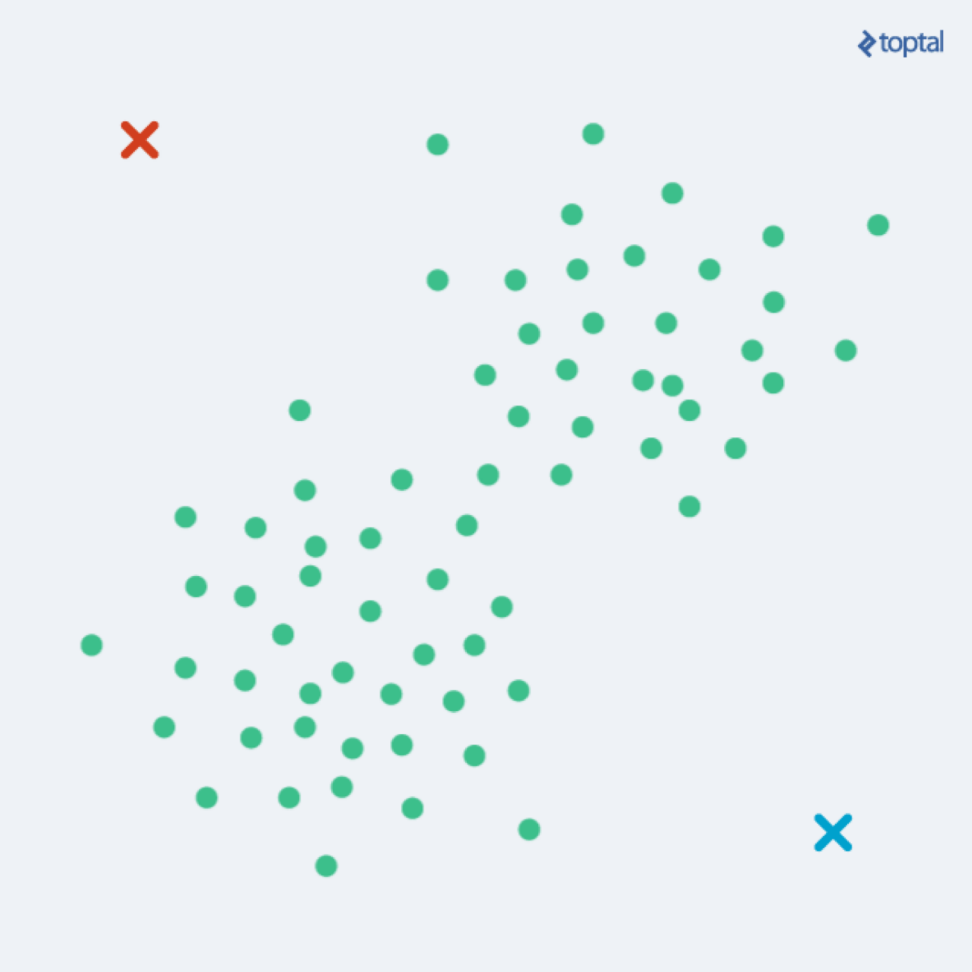
上述关键过程，在于通过大量的样本训练，得到模型的参数集（w1, w2, w3, b）。训练有2个关键要素：

•损失函数

•优化方式

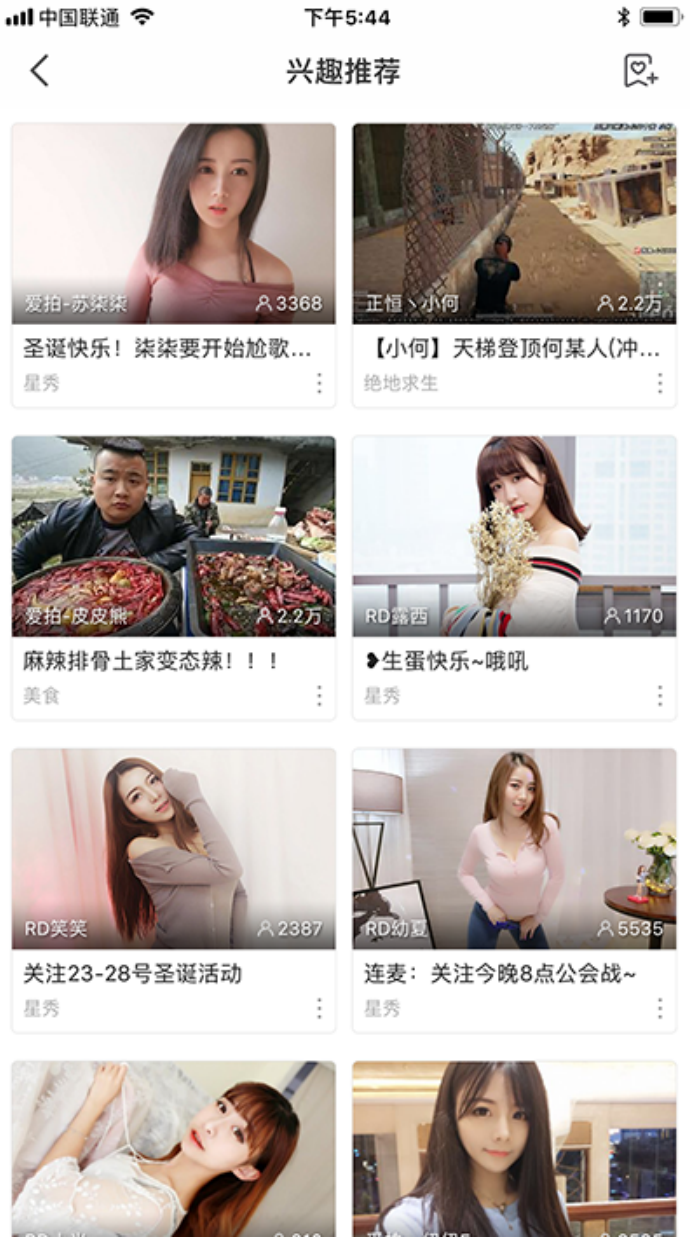
详细过程，不在此描述，请见本人撰文：《如何写一个模型训练器》

无监督学习：k-means

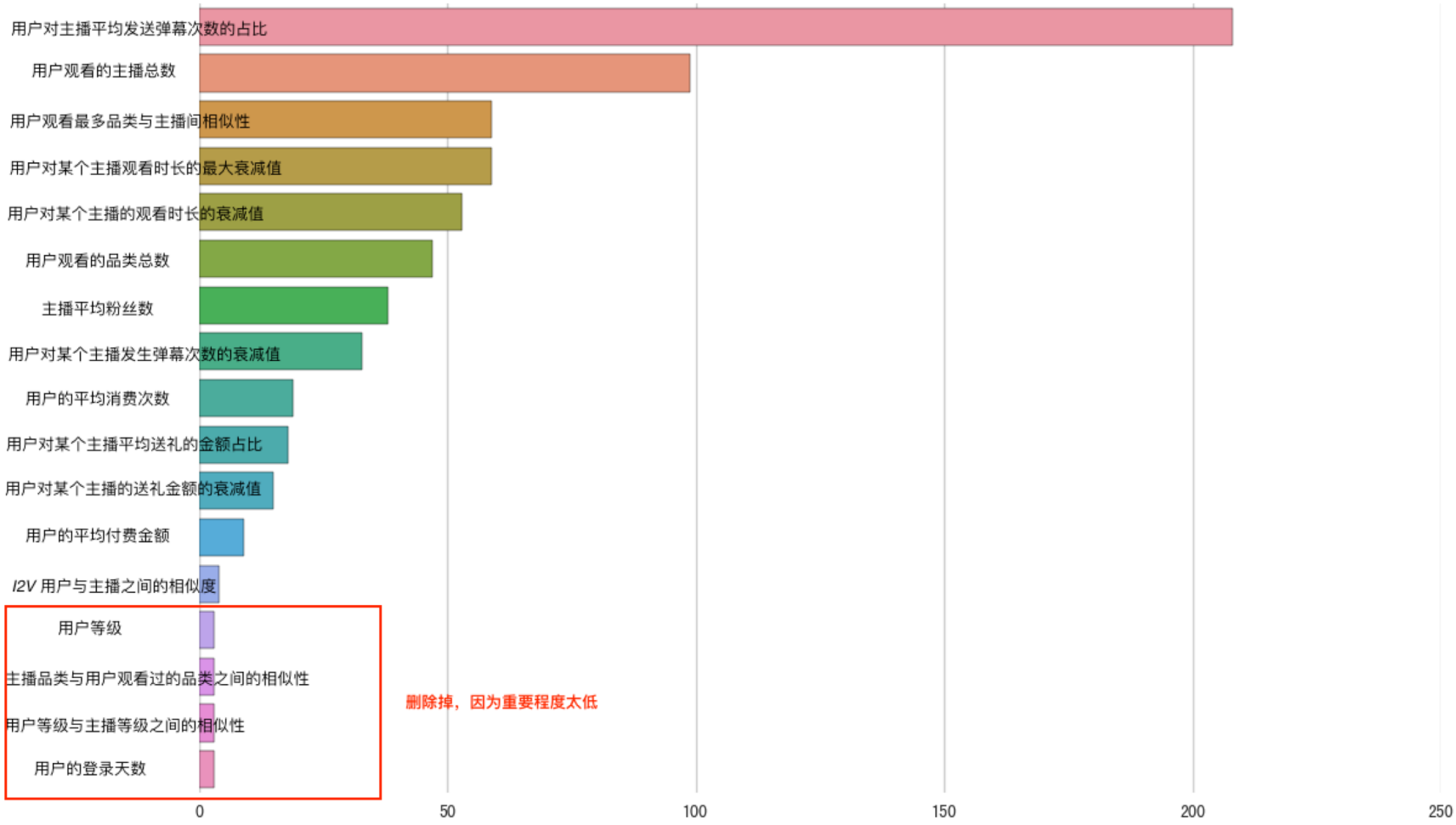


（四）机器学习实践

- 虎牙直播App个性推荐系统
- 在“全部直播”和“发现”栏目，接入个性化推荐引擎



个性推荐特征工程


用户对主播平均发送弹幕次数的占比
用户观看的主播总数
用户观看最多品类与主播间相似性
用户对某个主播观看时长的最大衰减值
用户对某个主播的观看时长的衰减值
用户观看的品类总数
主播平均粉丝数
用户对某个主播发生弹幕次数的衰减值
用户的平均消费次数
用户对某个主播平均送礼的金额占比
用户对某个主播的送礼金额的衰减值
用户的平均付费金额
I2V 用户与主播之间的相似度
用户等级
主播品类与用户观看过的品类之间的相似性
用户等级与主播等级之间的相似性
用户的登录天数
删除掉，因为重要程度太低
0
50
100
150
200
250

个性推荐原理

- 采用简单的逻辑回归模型（可解释性）
- 业务流程上分为召回服务（matching）、排序服务（ranking），ranking部分用到逻辑回归
- 模型的训练过程，请见陈聪撰文《生产模型的训练过程》

（五）参考书籍

- 李航：统计学习方法
- 周志华：机器学习
- 吴军：数学之美
- 开源书籍：Deep Learning （中译版）
- 开源书籍：THE LION WAY, Machine Learning plus Intelligent Optimization