[谷歌机器学习速成课笔记 2](#_Toc908180880)

[一、机器学习主要术语 2](#_Toc1903229816)

[1、标签 2](#_Toc1623687827)

[2、特征 2](#_Toc666411043)

[3、样本 3](#_Toc1181439423)

[4、模型 4](#_Toc672128946)

[5、回归与分类 4](#_Toc346956450)

# 谷歌机器学习速成课笔记

## 一、机器学习主要术语

什么是（监督式）机器学习？简单来说，它的定义如下：

机器学习系统通过学习如何组合输入信息来对从未见过的数据做出有用的预测。

下面我们来了解一下机器学习的基本术语。

## 1、标签

在简单线性回归中，标签是我们要预测的事物，即 y 变量。标签可以是小麦未来的价格、图片中显示的动物品种、音频剪辑的含义或任何事物。

## 2、特征

在简单线性回归中，特征是输入变量，即 x 变量。简单的机器学习项目可能会使用单个特征，而比较复杂的机器学习项目可能会使用数百万个特征，按如下方式指定：

2018-03-06 08-20-24 的屏幕截图

在垃圾邮件检测器示例中，特征可能包括：

A）电子邮件文本中的字词

B）发件人的地址

C）发送电子邮件的时段

D）电子邮件中包含“一种奇怪的把戏”这样的短语。

## 3、样本

样本是指数据的特定实例：x。（我们采用粗体 x 表示它是一个矢量。）我们将样本分为以下两类：

A）有标签样本

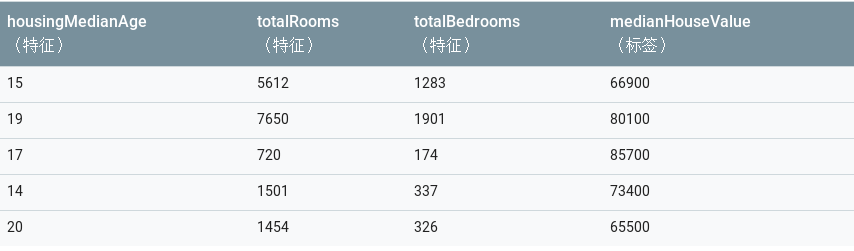
B）无标签样本

有标签样本同时包含特征和标签。即：

labeled examples: {features, label}: (x, y)

我们使用有标签样本来训练模型。在我们的垃圾邮件检测器示例中，有标签样本是用户明确标记为“垃圾邮件”或“非垃圾邮件”的各个电子邮件。

例如，下表显示了从包含加利福尼亚州房价信息的数据集中抽取的 5 个有标签样本：



无标签样本包含特征，但不包含标签。即：

unlabeled examples: {features, ?}: (x, ?)

在使用有标签样本训练了我们的模型之后，我们会使用该模型来预测无标签样本的标签。在垃圾邮件检测器示例中，无标签样本是用户尚未添加标签的新电子邮件。

## 4、模型

模型定义了特征与标签之间的关系。例如，垃圾邮件检测模型可能会将某些特征与“垃圾邮件”紧密联系起来。我们来重点介绍一下模型生命周期的两个阶段：

训练表示创建或学习模型。也就是说，您向模型展示有标签样本，让模型逐渐学习特征与标签之间的关系。

推断表示将训练后的模型应用于无标签样本。也就是说，您使用训练后的模型来做出有用的预测 (y')。例如，在推断期间，您可以针对新的无标签样本预测 medianHouseValue。

## 5、回归与分类

回归模型可预测连续值。例如，回归模型做出的预测可回答如下问题：

A）加利福尼亚州一栋房产的价值是多少？

B）用户点击此广告的概率是多少？

分类模型可预测离散值。例如，分类模型做出的预测可回答如下问题：

A）某个指定电子邮件是垃圾邮件还是非垃圾邮件？

B）这是一张狗、猫还是仓鼠图片？

## 备注

（1）特征必须是可量化的，否则不能做特征。