机器学习问题构建

基本上，ML是训练一个模型的过程，以便使用一个数据集合来进行有用的预测。这个预测模型可以对之前没有见过的数据进行预测。

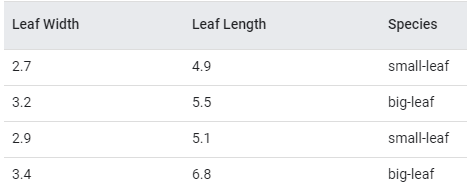
一般，ML包括2个范式，监督和非监督学习。

什么是监督学习？

监督学习使用带标签的数据来训练模型。

比如您是一个业余植物学家，想要区分两种看上去很像的植物，幸运的是一个植物学家已经将一些这两种植物的样本还有他们的标签提供给您。

以下是一些数据：



叶子宽度和长度是特征，都标记为X，物种是标签。真是的植物数据可能包括更多的特征，但是标签只有一个。特征是测量值或者描述，标签是答案。比如这个数据集的目的是帮助其他植物学家回答这个问题：这个植物是哪个物种？

这个数据集包括样本。现实世界中数据集包括非常多样本。

在监督学习中，您将样本特征和对应的标签输入训练算法过程。在训练过程中，算法逐渐确定特征和标签之间的关系。这个关系称为模型。

监督学习找到的数据和标签之间的模式可以表达成数学函数。输入一个特征，告诉系统期望输出的label是什么，这叫做监督学习。机器学习将学习标签数据的模式，然后再使用这些模式对训练中没有见过的数据进行预测。

现实世界中监督学习的一个例子是斯坦福大学使用图片来检测皮肤癌。

非监督学习

非监督学习的任务是找到数据中有意义的模式。数据是不带标签的。换句话说，模型没有任何数据分类的线索，必须自己来推断规则。

机器学习思维

在传统的软件工程领域，您可以完成从需求到可行设计的推理，但是在机器学习领域，找到一个可行模型的方式是实验。

许多机器学习系统通过不同于人类的信号解释来产生编码知识和智能的模型。神经网络可以使用embeding来解释一个单词。神经网络可以使用这些embeding符号来做准确的翻译或者语义分析，但是人类却无法理解这些符号。

模型会犯一些奇怪的错误，这些错误很难调试，这是由于从扭曲的训练数据到训练期间意外的数据解释等原因造成的。此外，当机器学习模型被合并到产品中时，交互可能会变得复杂，使得预测和测试所有可能的情况变得困难。这些挑战要求产品团队花费大量时间来了解他们的机器学习系统在做什么以及如何改进它们。

实验设计入门

习惯不确定性

在传统编程中，您设定参数并知道将得到的结果，使用ML，非代码工作将很复杂。

科学的方法

解决一个机器学习问题，将机器学习理解成一个实验过程，我们迭代测试最终收敛到一个可行的模型。就像一个实验，整个过程是令人兴奋、富有挑战性，最终是值得的。

|  |  |
| --- | --- |
| 步骤 | 例子 |
| 确定研究目标 | 我想预测某一天的交通拥堵 |
| 做一个假设 | 我认为天气预报是一个有用的信息 |
| 收集数据 | 收集历史交通数据和天气数据 |
| 测试假设 | 使用这些数据训练模型 |
| 分析接口过 | 这个模型比现有的系统表现更好么？ |
| 获得结论 | 我将采用(放弃)这个模型因为1、2、3。。。 |
| 重新定义假设并重复 | 一年中的时间段将是有用的信号 |

确定机器学习合适的问题

介绍了机器学习问题的特征。

使用传统的程序来解决自动邮件回复或者google 照片搜索，没有明显好的方法。但是，机器学习将通过检查数据之间的模式，并利用这些模式来解决这类问题。将机器学习当作工具箱中的工具并且只是在合适的时候拿来使用。

这样，首先问自己两个问题：

1. 我的产品面临的问题是什么？
2. 对于ML来说，这是一个好问题么？

在关注数据之前了解问题

如果您清楚地理解这个问题，您应该能够列出一些潜在的解决方案进行测试，以便生成最佳的模型。要明白，在得到一个好的模型之前，您可能需要尝试一些解决方案。

探索性数据分析可以帮助您理解您的数据，但在您对照以前未看到的数据检查这些模式之前，您还不能声称找到的模式是泛化的。不检查可能会导致你走错方向，或者强化刻板印象或偏见。

依靠团队日志

专门为您的任务收集的数据将是最有用的。实际上，你可能做不到这一点，你会依赖任何你能得到的足够接近的数据。这很好，只要你知道成本，并且你最终可以得到产品日志，你可以用它们来建立一些更有针对性的任务。

“很多”是多少？这取决于问题，但更多的数据通常会改进模型，从而提高模型的预测能力。一个好的经验法则是基本线性模型至少有数千个样本，神经网络有数十万个样本。如果数据较少，请首先考虑使用非ml解决方案。

您不应该试图让ml做很多工作来发现哪些特性与您的问题相关。如果你把所有东西都扔到模型上，看看什么看起来有用，那么你的模型很可能会变得过于复杂、昂贵，并且充满了不重要的特性。在较小的数据集中，在您的数据样本中，您有更高的机会偶然地将某个功能与您的标签关联起来。如果你在没有假设的情况下尝试了很多特性，你会错误地认为这些是模型的相关信号。只有当你试图用你的模型进行预测并意识到它没有普遍性时，你才会明白这一点。

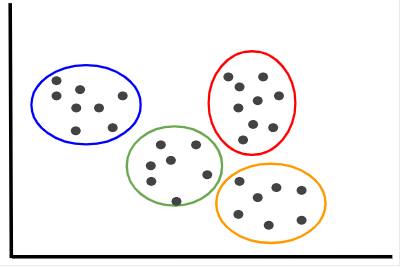
机器学习的难题

聚类

非监督学习中聚类代表什么意思？比如，如果你的模型显示用户属于蓝色集群，你将确定蓝色集群表示什么。

有时，您可以基于集群采取操作。例如，Google Photos使用集群将同一个人的图片组合在一起。其他时候，根据集群确定要采取的操作是很有挑战性的。您可以尝试为集群分配一个含义，但这可能会很棘手，因为模型可能不会按您认为直观的条件进行分组。

另一种方法是在集群之前标记一些项，然后尝试在整个集群中传播这些标签。例如，如果标签为X的所有项最终都在一个集群中，则可以将标签X分散到其他示例中。

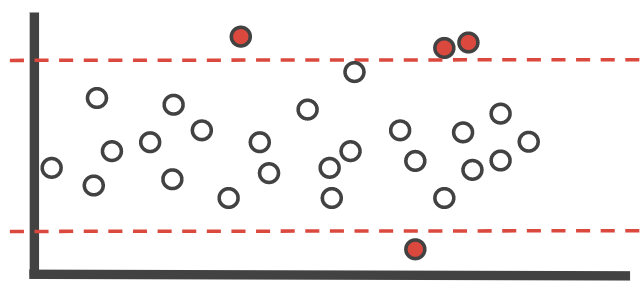


异常检测

有时，人们希望使用ML来识别异常。诀窍是，如何确定什么构成了标记数据的异常？一种方法是定义一个启发式方法并用它来标记异常。然而，一旦定义了这种启发式，您就可以在生产系统中使用启发式，因为ML模型无法胜过用于训练它的启发式。

有时，你可以设计一个高精度的低召回启发式算法，然后使用半监督方法训练一个模型，使其从一组“种子”预测增长到对一组更大的未标记数据进行分类。

如果你的启发式方法足够复杂，那么考虑使用ML来替换该系统可能是值得的。但是，继续前进时要小心，因为您无法像优化启发式方法那样轻松地优化模型。



因果关系

ML可以识别两个或多个事物之间的相互关系或联系。确定因果关系（一个事件或导致另一个事件或因素）要困难得多。换言之，很容易看出发生了什么，但很难理解为什么会发生。

例子：

消费者购买某本书是因为他们在前一周看到了正面的评论，还是即使没有评论他们也会购买？

你不能仅仅从观察数据来确定原因。如上例所示，您不能仅仅通过查看过去的事件来确定观看评论是否导致了购买。您需要运行一个实验，将没有看到评论的用户与看到评论的类似用户进行比较。一般来说，你需要介入这个世界——进行一个实验——来确定原因，你不能从纯粹的观察数据中看到。

没有现有数据

如前所述，如果您没有数据来训练模型，那么机器学习无法帮助您。没有数据，使用一个简单的、启发式的、基于规则的系统。许多没有训练数据的新产品都是从启发式规则系统开始的，只有在用户与之交互后才能获得训练数据。一旦你有了训练数据，试着找出其中的模式。如果没有模式或者只有微不足道的模式，那么机器学习可能不会提供价值。如果有很多模式，并且做出准确的预测很重要，那么使用机器学习可能是正确的方法。

**构建一个问题**

决定使用ML

在你深入研究你的数据之前，为成功做好准备是很重要的。下一节概述了在尝试为机器学习设计问题之前需要考虑的几个问题。

从简单开始

简单地说，您希望您的ML模型做什么？

举例：

我们希望ML模型预测上传的视频未来是否流行

在这一点上，陈述可以是定性的，但要确保它抓住了你真正的目标，而不是间接的目标。

您理想的输出结果是什么

将ML模型添加到系统中应该会产生理想的结果。这个结果可能与您评估模型及其质量的方式大不相同。

举例：

1. 编码

我们的理想结果是在转码不太流行的视频时使用更少的资源。

转码是将用户上传的内容转换为YouTube向用户显示的更高效格式的过程。YouTube上超过1/3的视频观看次数少于10次！如果我们能够识别这些视频，并且只准备低分辨率的版本，我们可以节省大量资源。

1. 视频推荐

我们的理想结果是推荐人们觉得有用、有趣、值得花时间的视频。

不要局限于产品已经优化过的指标。相反，专注于产品或服务的更大目标。

成功与失败指标

量化

你怎么知道你的系统是成功的还是失败的？

你的成功和失败指标的措辞应该独立于你的评估指标，如精确性、召回率或AUC。相反，指定预期的结果。在开始之前设定你的成功标准，以防止沉没成本激励你推出一个平庸的模型。

举例：

1. 编码

一个成功指标是CPU资源利用率。成功意味着将转码的CPU成本降低35%。失败意味着降低的成本小于培训和服务模型的CPU成本。

1. 视频推荐

一个成功指标是模型正确预测的流行视频的数量。成功意味着预测到95%的在上传后28天内观看时间最长的视频。失败意味着正确预测的热门视频数量并不比当前的启发式方法好多少。

指标是否可度量？

一个可衡量的指标为成功的现实世界评估提供了足够的信息。例如，监测果园健康状况的系统可能希望减少病死树的比例。但是如果你不能测量有多少树生病，这不是一个有用的指标。

思考以下几个问题：

如何度量这些指标？

什么时候可以度量这些指标

需要多久能判断新的ML系统是成功还是失败？

理想情况下，你想提早失败。注意你的数据中信号太少，或者数据不具有预测性，以确定你的假设是否可能是错误的。快速的失败将使你能够在这个过程的早期修正你的假设，并防止失去时间。

你认为你可以根据用户的位置来决定哪些视频会被观看。这似乎是有可能的，但当你尝试它，信号太弱或有太多的噪音，它不起作用。你会坚持这个假设多久？

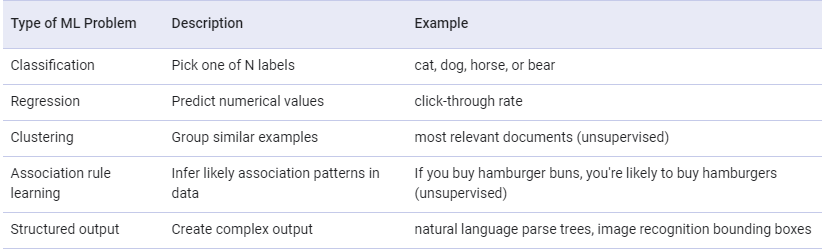
考虑长期的工程和维护成本。

其他失败的情形

注意与你的成功标准无关的失败。例如，总是推荐“clickbait”视频的视频建议系统在提供高质量的观看体验方面并不成功。

您希望ML模型产生什么输出？

参考下表，你想要获得哪一种输出：



好的输出的特性

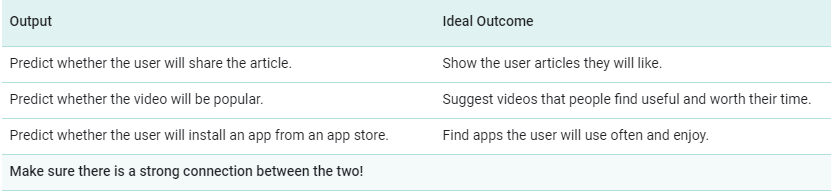
输出必须是可量化的，并且具有机器可以产生的明确定义。

如果不询问用户，就无法判断用户是否喜欢视频。如果无法询问用户，则需要使用代理标签。也就是说，你需要用一个替代标签来代替真实的东西。用户共享视频的程度是一个很好的代理标签。的确，分享不是享受的完美近似，因为用户可能因为娱乐而分享视频（例

如，取笑视频）。然而，共享是可量化的，可跟踪的，并提供了一个像样的预测信号。

输出应该与您的理想结果相关联

您的模型将优化输出。因此，确保输出确实是您所关心的。代理标签经常是必要的，因为我们不能总是直接衡量我们的理想结果。然而，我们的标签和真实结果之间的联系越紧密，我们就越有信心优化正确的事情。



可以获得训练数据的示例输出么？

如何从哪里获得？

有监督的机器学习依赖于标记的数据。如果很难获得用于培训的示例输出，则可能需要重新查看对过去练习的响应，以重新制定问题和目标，以便可以对数据进行模型培训。您的输出示例可能需要进行工程设计，如上面的示例所示，这会将观看时间转换为百分比。

不好的目标

如果设置得当，ML系统非常擅长实现给定的目标。相反，当给定错误的目标时，ML系统可能会产生意外的结果。因此，仔细考虑系统的目标将如何帮助您解决问题。

回想一下我们的模型，它预测有人将在YouTube上观看下一个视频。下面都是一些不好的目标：

Maximize Click Rate

Maximize Watch Time

Maximize Watch Time

Increase Diversity & Maximize Session Watch Time

启发式规则

如果没有ML，你怎么解决你的问题呢？

假设您明天需要交付一个产品，而您只有足够的时间来硬编码业务逻辑。您可以尝试以下启发式（非ML解决方案）：

前面的启发式可能不是世界上最伟大的启发式，但它确实提供了一个基线。永远不要推出一个无法超越启发式的奇特的ML模型。启发式的练习通常会帮助您在ML模型中识别出好的信号。

非ML解决方法有时比ML解决方法更容易维护

**将问题表述为ML问题**

本节是构建ML问题的建议方法的指南：

1. 指出问题。
2. 从简单开始。
3. 确定数据源。
4. 为模型设计数据。
5. 确定数据的来源。
6. 确定容易获得的输入。
7. 学习能力。
8. 考虑潜在的偏见

指出问题

分类和回归有几个亚型。使用相应的流程图来标识您正在使用的子类型。此流程图帮助您组合正确的语言，以便与其他ML实践者讨论您的问题。根据您的业务问题使用分类或回归流程图。

我们的问题最好概括为：

二进制分类

一维回归

多类单标签分类

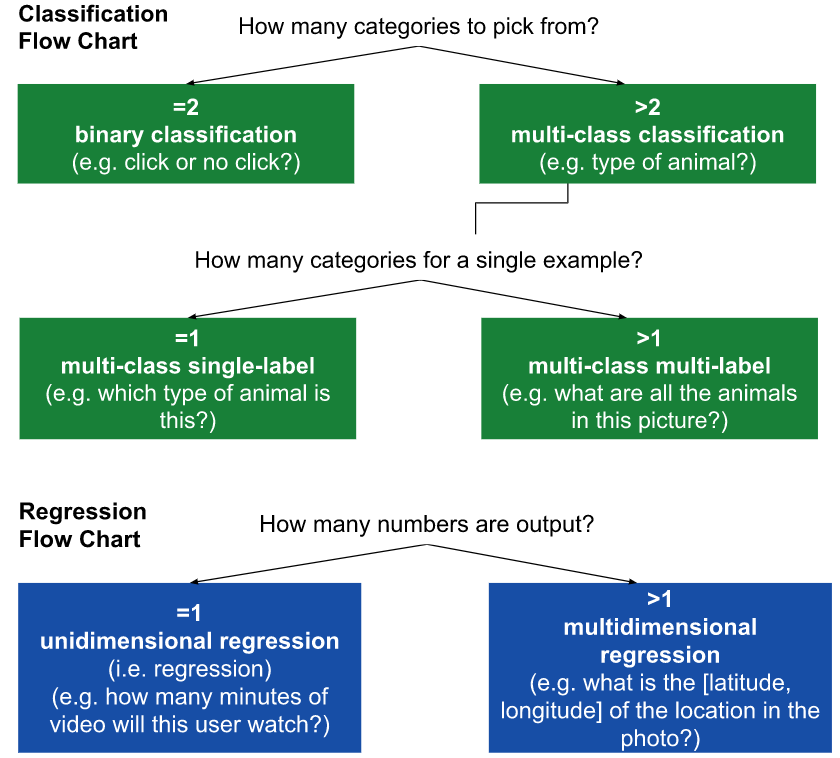
多类别多标签分类

多维回归

群集（无监督）

其他（翻译、解析、边界框id等）

然后，在提出问题之后，解释模型将预测什么



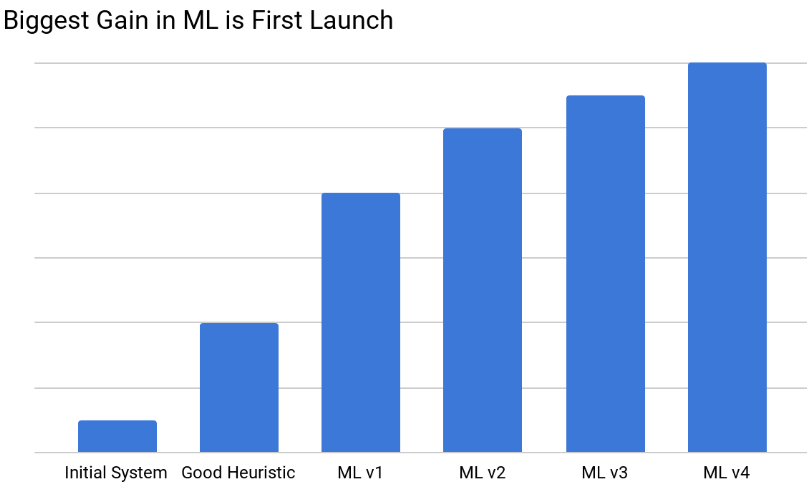
从简单开始

首先，简化建模任务。将给定问题声明为二进制分类或一维回归问题（或两者）。这两个问题都是经过很好的处理和监督的方法，它们有大量的工具和专家支持来帮助您开始。

然后，对于这个任务，使用尽可能简单的模型。一个简单的模型更容易实现和理解。一旦拥有完整的ML管道，就可以更轻松地迭代简单模型。

简单的模型提供了一个很好的基线，即使你最终没有发布它们。事实上，一个简单的模型可能比你想象的要好。从简单开始可以帮助您确定复杂模型是否合理。更复杂的模型训练更困难、更慢，更难理解，所以要保持简单，除非复杂性为模型质量提供了足够大的改进，以证明这些权衡。

从ML获得的最大收益往往是首次发布，因为这是您可以首先利用数据的时候。进一步的调整仍然会带来成功，但是，一般来说，最大的收获是在开始时，所以最好选择经过良好测试的方法，以使过程更容易。



确定数据源

有多少标记数据

标记的来源是什么

这个标记是否和最终做的决定紧密相关

为模型设计数据

确定您的ML系统应该用来进行预测的数据（输入->输出）

确定容易获得的输入

选择1-3个容易获得并且你认为会产生合理的初步结果的输入。

哪些输入对于实现前面提到的启发式有帮助？

考虑开发数据管道以准备输入的工程成本，以及在模型中包含每个输入的预期收益。

关注可以从一个简单的管道系统中获得的输入。从尽可能少的基础设施开始。

确定数据来源

评估开发一个数据管道来为一行构造每一列需要做多少工作。示例输出何时可用于培训目的？如果很难获得示例输出，则可能需要重新访问输出，并检查是否可以为模型使用不同的输出。

确保您的所有输入在预测时都是可用的，格式完全符合您写下的格式。如果在预测时很难获得某些特征值，请从模型中忽略这些特征。

学习能力

ML模型能够学习吗？列出可能导致学习困难的问题。例如：

数据集没有包含足够的正标签

训练数据没有包含足够的示例

噪声数据太多

系统记住了练数据，但难以推广到新的案例

考虑潜在的偏见

许多数据集在某种程度上有偏差。这些偏见可能会对训练和预测产生不利影响。例如：

有偏差的数据源不能跨多个上下文进行转换。

训练集可能不能代表模型的最终用户，因此可能会给他们带来负面体验。