

跟我一起学ML20 (二) : Regression

原创 多杰平措 多杰平措OPL 2020-04-09 22:25:53 手机阅读

这篇文章开始我们正式进入精灵宝可梦大师Machine Learning 2020课程正式内容的学习，今天我们将要学习的专题是Regression。

前言

进入学习之前，首先公布一下本次课程的GitHub仓库，这个仓库包含了课程的slides、作业源码和课程相关资料，方便大家学习。仓库地址：

<https://github.com/Aierhaimian/ml20>

另外，本课程虽然是公开课，但是课程视频都是公开在YouTube网站上，对于国内的同学们来说非常不便，那么这里我们要特别感谢B站大佬的迷弟的粉丝同学，将视频从YouTube搬运到了B站，为大家带来了方便。视频观看地址：

https://www.bilibili.com/video/BV1JE411g7XF/?spm_id_from=333.788.videocard.0

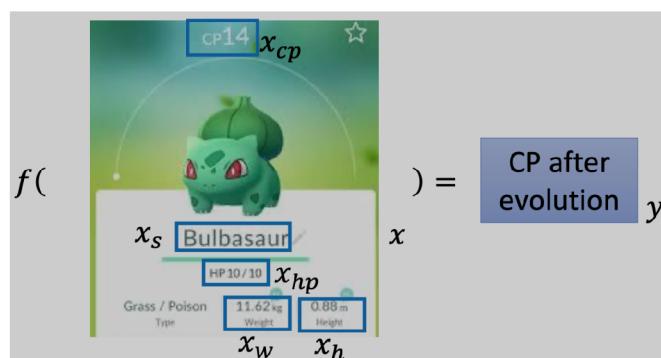
接下来，我们正式开始今天的内容。

问题提出：如何预测宝可梦的CP值？

相信爱玩游戏的同学们都很熟悉精灵宝可梦，小时候我们都有看过精灵宝可梦的动画片。假如我们抓到了一只妙蛙种子，可以看到它的各种属性，那么我们就想知道它进化后到底厉不厉害，能不能打得过其他的精灵？这时候我们会期望根据已有宝可梦进化前的属性，来预测某只宝可梦进化后的CP值的大小。

假设我们已经有一批宝可梦，我们知道其进化前的各种属性以及其进化后的CP值（labeled）。因为我们通过已知数据区训练模型，然后通过模型预测未知的数据，因此这是有监督学习（Supervised Learning）。

那么我们是不是可以写一个模型，根据已知宝可梦数据训练这个模型，模型训练完成后，输入需要预测的某只宝可梦进化前的各种属性，通过模型运算后输出该宝可梦进化后的预测CP值。因为这个预测值是一个标量（scalar），因此我们的任务就是回归（Regression）类型。



上图参数说明：

X : 表示一只宝可梦，用下标表示该宝可梦的某种属性
 X_q : 表示该宝可梦进化前的CP值。
 X_s : 表示该宝可梦属于哪一种物种，如皮卡丘、杰尼龟...
 X_{hp} : 表示该宝可梦的HP值
 X_w : 表示该宝可梦的重量
 X_h : 表示该宝可梦的高度
 $f(\cdot)$: 表示我们寻找的model
 y : model输出，即宝可梦进化后的CP值(scalar)

问题解决：Regression

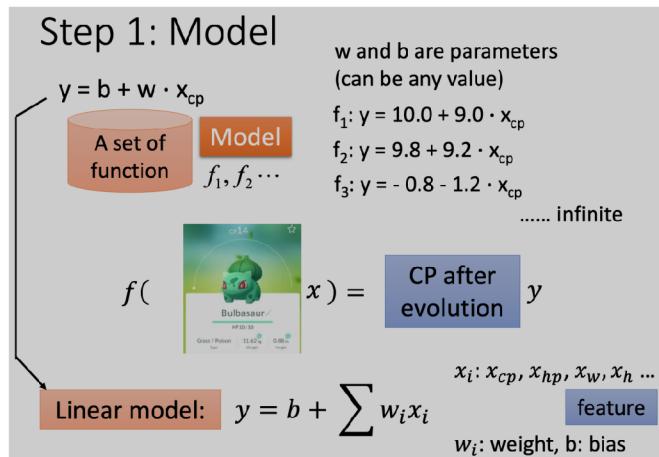
宝可梦CP值预测这个问题需要用Machine Learning的方法去解决，通常来说解决一个Machine Learning问题需要三步骤：

1. 定义一个模型即function set。
2. 找到一个goodness of function损失函数评估function好坏。
3. 找到一个最优的function。

接下来，我们就照着三个步骤去解决宝可梦CP值预测的问题。

第一步：设计模型（function set）

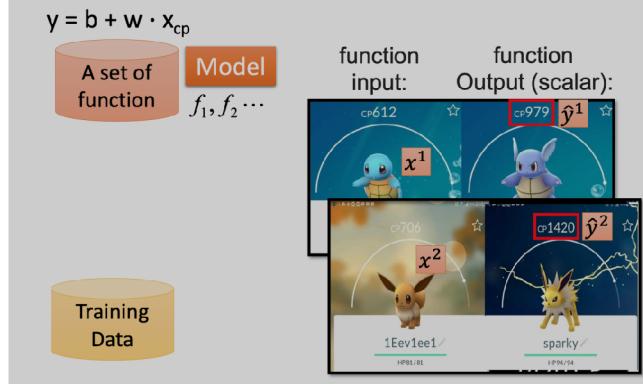
设计模型考验的是对问题深刻的理解、深厚的机器学习功力和经验，否则只能每个模型都试一遍，找到一个最好的模型，这也不失为一种方法。课程中我们选择了线性模型（Linear Model）来解决宝可梦CP值预测这个问题。



第二步：Goodness of Function

我们选择了一个模型，但正如上文所说的，模型千千万，我们选择的这个模型就一定是最好的吗？有没有什么方法可以评判我们选择的模型好坏呢？

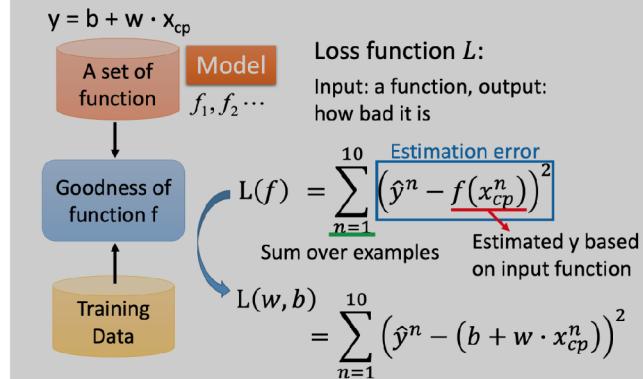
Step 2: Goodness of Function



为了衡量function set中的某个function的好坏，我们需要一个评估函数，即损失函数（loss function），简称L。可以将loss function理解为是一个function的function。

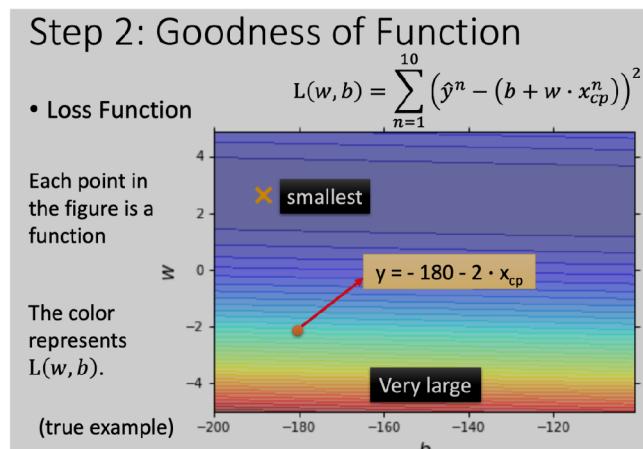
之前提到的model是由我们自主选择的，这里的loss function也是，最常用的方法就是采用类似于方差和的形式来衡量参数的好坏，即预测值与真值差的平方和；这里真正的数值减估测数值的平方，叫做估测误差（Estimation error），将10个估测误差合起来就是loss function。

Step 2: Goodness of Function



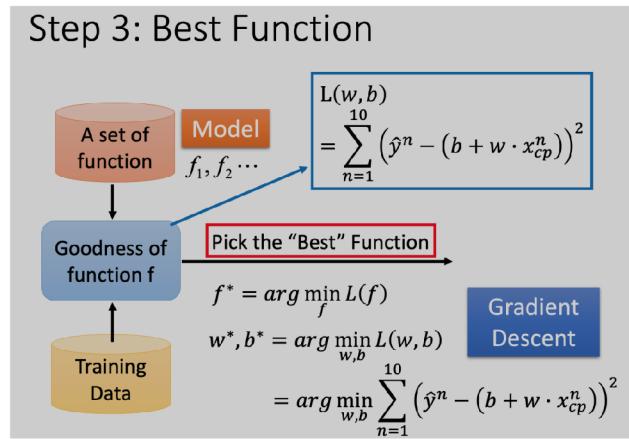
如果 $L(f)$ 越大，说明该function表现得越不好； $L(f)$ 越小，说明该function表现得越好。

下图中是loss function的可视化，该图中的每一个点都代表一组 (w, b) ，也就是对应着一个function；而该点的颜色对应着的loss function的结果 $L(w, b)$ ，它表示该点对应function的表现有多糟糕，颜色越偏红色代表Loss的数值越大，这个function的表现越不好，越偏蓝色代表Loss的数值越小，这个function的表现越好。



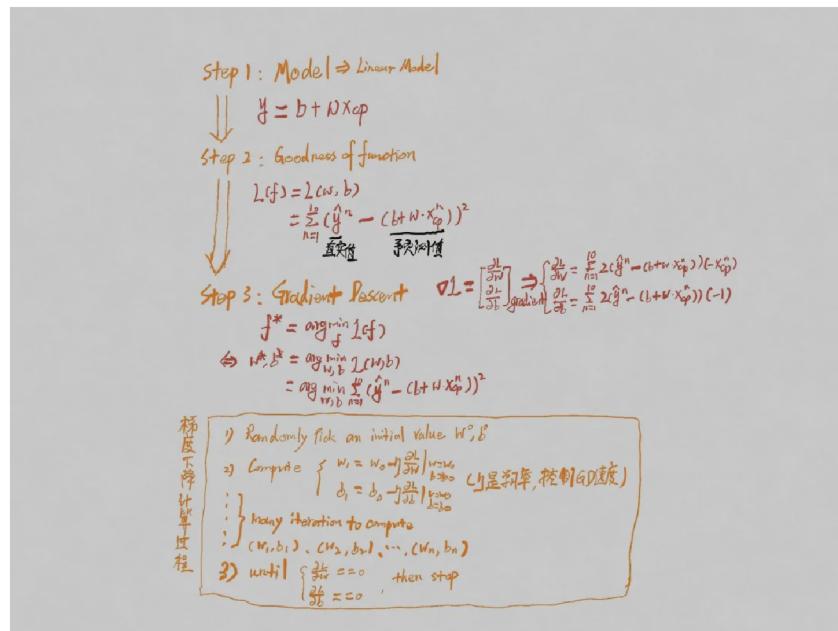
第三步：Best Function

我们已经确定了loss function，他可以衡量我们的model里面每一个function的好坏，接下来我们要做的事情就是，从这个function set里面，挑选一个最好的一个function，如图所示。

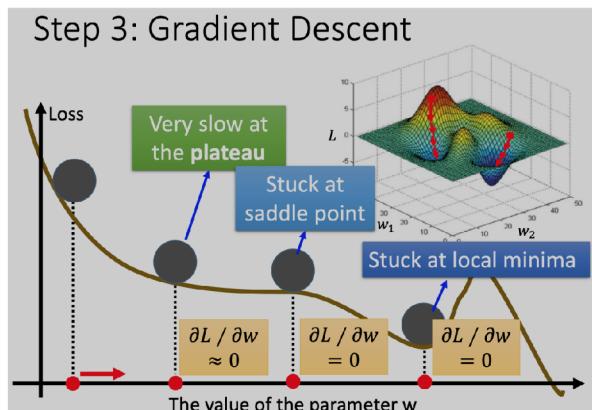


挑选最好的function，实质上是Loss最小，那如何才能找到使Loss最小的w, b呢？就是采用梯度下降（gradient descent）法。

梯度下降法的厉害之处在于，只要 $L(f)$ 是可微分的，梯度下降法都可以拿来处理这个 f ，找到表现比较好的参数。关于本节内容中梯度下降法的具体操作过程详见精灵宝可梦大师的课程视频和slides（后面有一节专门讲梯度下降的内容，到时候再具体展开）。下图总结了梯度下降法的过程，看过相关内容后，这幅图一定会使你有一种醍醐灌顶的感觉。

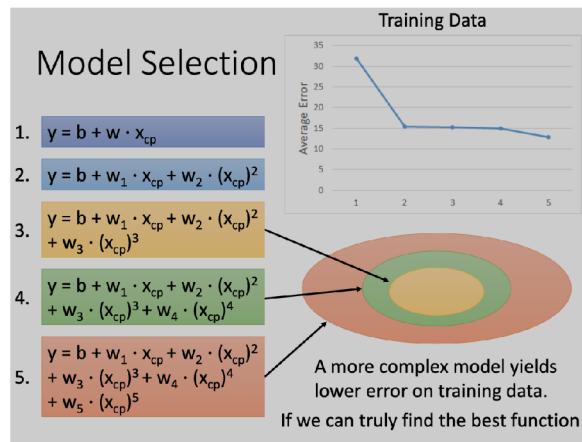


这里提出一个问题，如下图所示采用梯度下降法就一定能找到最合适那个function吗？希望大家能够学习完本节课程内容后可以自己回答这个问题。

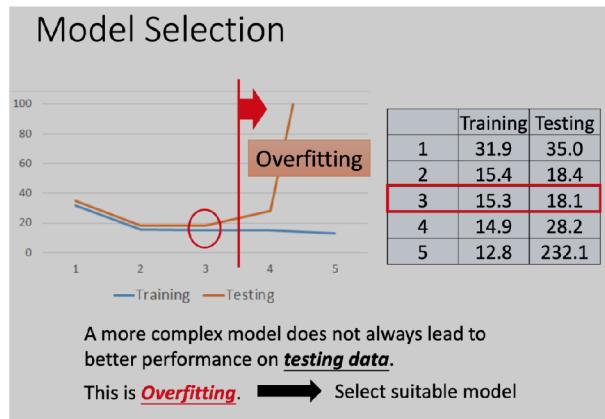


回到我们求宝可梦CP值预测的问题上来，通过梯度下降法，我们得到了最佳的 $b=-188.4$, $w=2.7$ ，训练集中每只宝可梦进化后的实际CP值与预测CP值之间的误差平均值为31.9。此时，如果我们用一只刚抓来的宝可梦带入到训练好的模型去预测它进化后的CP值，可以得到其误差值为35，这个值是大于在训练集上的平均预测误差的，那我们还没有可能做的更好呢？

那我们可以重新去设计模型，让它变的更加复杂，通过观看视频和slides的内容我们也实际看到了精灵宝可梦大师设计的5个模型，会让平均误差越来越小。

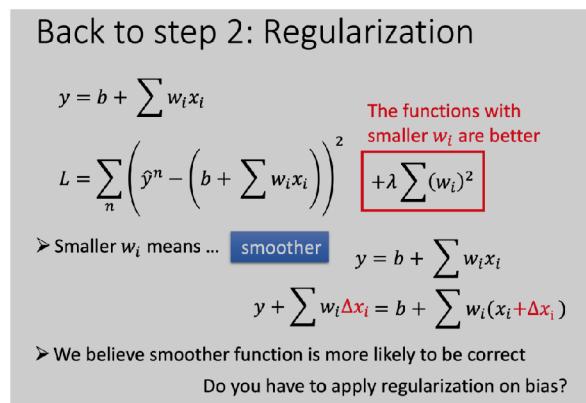


在training data上，model越复杂，error就会越低；但是在testing data上，model复杂到一定程度之后，error非但不会减小，反而会暴增，在该例中，从含有四次项的model开始往后的model，testing data上的error出现了大幅增长的现象，通常被称为过拟合（overfitting）。



因此，模型也不是越复杂越好，还是得选择最适合的。另外，精灵宝可梦大师还尝试了在模型中加入宝可梦的其他属性，可以看到模型在training data上的平均误差可以做到非常小，但是在testing data上的平均误差却差到令人发指。

怎么解决这个问题呢？我们可以通过regularization解决overfitting，regularization可以使曲线变得更加smooth，training data上的error变大，但是 testing data上的error变小。



从图中我们可以发现我们是非常喜欢w越小甚至接近于0的function。因为参数值接近0的function，是比较平滑的；所谓的平滑的意思是，当今天的输入有变化的时候，output对输入的变化是比较不敏感的。因此测试的时候，一些noises噪声

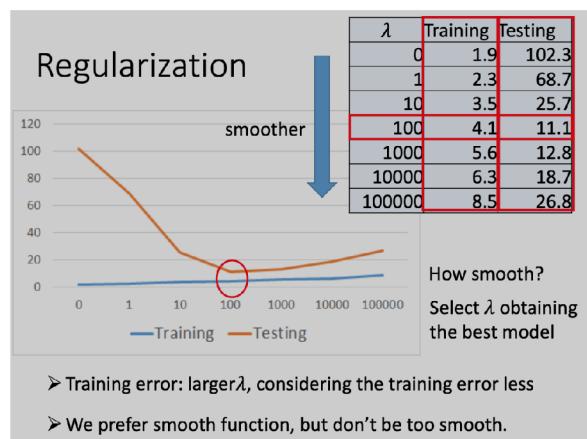
对这个平滑的function的影响就会比较小，而给我们一个比较好的结果。

λ 值越大代表考虑smooth的那个regularization那一项的影响力越大，我们找到的function就越平滑。

观察下图可知，当我们的 λ 越大的时候，在training data上得到的error其实是越大的，但是这件事情是非常合理的，因为当 λ 越大的时候，我们就越倾向于考虑w的值而越少考虑error的大小；但是有趣的是，虽然在training data上得到的error越大，但是在testing data上得到的error可能会是比较小的。

下图中，当 λ 从0到100变大的时候，training error不断变大，testing error反而不断变小；但是当 λ 太大的时候(>100)，在testing data上的error就会越来越大。

我们喜欢比较平滑的function，因为它对noise不那么sensitive；但是我们又不喜欢太平滑的function，因为它就失去了对data拟合的能力；而function的平滑程度，就需要通过调整 λ 来决定，就像下图中，当 $\lambda=100$ 时，在testing data上的error最小，因此我们选择 $\lambda=100$ 。



通过这波操作，我们就完成了一个宝可梦进化后CP值的预测模型，以后我们抓到一只宝可梦，就可以将它的属性值输入到这个模型中，然后就可以知道这只宝可梦进化后到底厉不厉害啦~

以上就是本次Regression章节所有的内容啦，欢迎大家踊跃留言~

注：以上部分图片来自李宏毅老师机器学习2020课程slides。

END

欢迎关注公众号 **多杰平措OPL**，为您带来更多精彩内容~

