机器学习学习笔记

Project：机器学习

Author：胡小忠

Version：2019/5/6-1

前言

该文档产生自本人在学习机器学习方法及Google深度学习框架TensorFlow过程后。主要内容包括机器学习基本概念、神经网络等基本概念，也包括了一些TensorFlow使用的方法内容，也总结了在实际使用机器学习训练神经网络模型时遇到的一些问题和解决方法。

主要参考资料：

* 机器学习速成课程：<https://developers.google.cn/machine-learning/>
* 《实战Google深度学习框架》
* 《机器学习》
* 《Python神经网络编程》
* 《模式识别及MATLAB实现》

公式使用TeX格式渲染。

机器学习简介

机器学习可以帮助我：

* 缩短编程时间。
* 定义自己的产品。
* 解决自己作为编程人员却不知道如何用人工方法解决的问题（提供足够的样本和合适的模型）。

框架处理

（监督式）机器学习：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。

标签：要预测的对象，即简单线性回归中的y变量。

特征：输入变量，即简单线性回归中的x变量。复杂的机器学习可能使用数百万个特征，例如：。

样本：数据的特定实例：。可以为：无标签样本和有标签样本。使用有标签的样本训练模型，对无标签样本进行预测。

模型：特征与标签之间的关系。

回归模型：预测连续值。

分类模型：预测离散值。

深入了解机器学习

线性回归：。机器学习中为：。找到一种最适合一组点的直线或超平面的方法。

训练：让损失最小化。

损失：

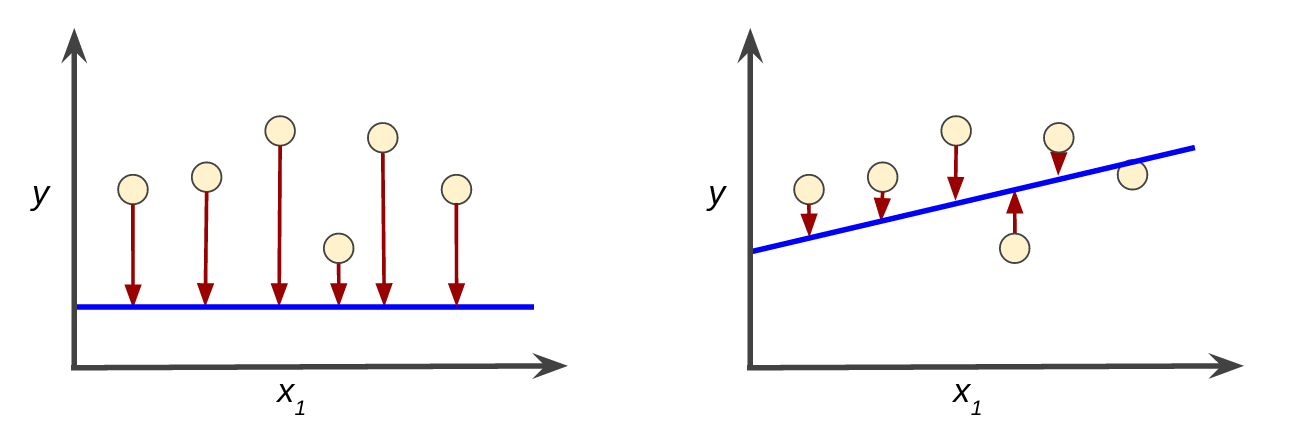


图 1 左侧损失大，右侧损失小

* 均方损失：，需要注意的是，该损失函数并不是唯一的损失函数，也不是适用于所有情况的损失函数。
* 平方损失：又称损失。即。

降低损失

迭代方法：

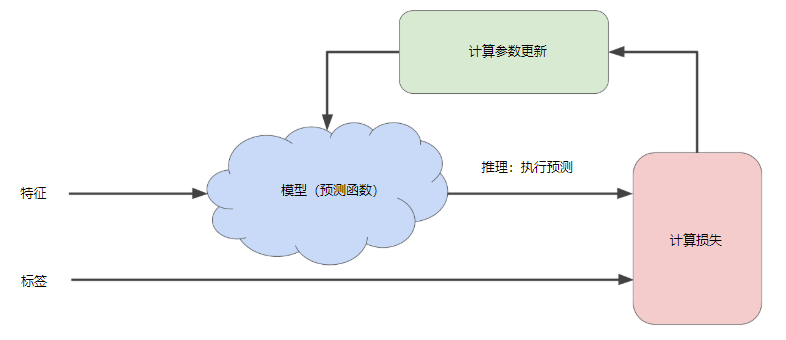


图 2 计算参数更新

梯度下降法：

梯度是一个矢量，因此具有两个重要特征：方向和大小。梯度始终指向损失函数中增长最为迅猛的方向。梯度下降算法会沿着负梯度的方向走，以便尽快降低损失。

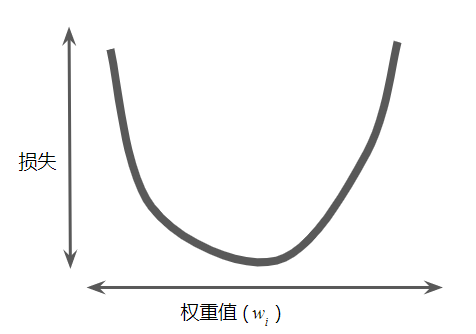


图 3 回归问题产生的损失与权重图为凸型

学习速率：

也就是梯度下降算法中每次下降的步长。过小，会收敛太慢。太大，会来回弹跳，不收敛。

* 一维空间中，最理想的学习速率是，也是对二阶导数的倒数。
* 二维或者多维空间中理想的学习速率是海森距离（由二阶偏导数构成的矩阵）的倒数。

优化学习速率：多次试用。

随机梯度下降法（SGD）：

每次迭代使用全批量不太现实，计算量太大，所示可以每次迭代只使用一个样本（批量大小为1）。或者使用折衷方案，小批量SGD，10-1000个随机选择的样本。**梯度下降法也同样适用于包含多个特征的特征集。**

使用TensorFlow基本步骤

|  |
| --- |
| 1. # 导入TensorFlow package. 2. import tensorflow as tf 3. # 创建一个线性分类器 4. classifier = tf.estimator.LinearClassifier() 5. # 训练样本 6. classifier.train(input\_fn=train\_input\_fn, steps=2000) 7. # 对指定样本进行预测，事实证明，并不好用 8. predictions = classifier.predict(input\_fn=predict\_input\_fn) |

代码 1 使用TensorFlow创建线性预测器并进行预测

具体参考：

书籍：TensorFlow实战Google深度学习框架（第二版）

文档：TensorFlow官网文档<https://tensorflow.google.cn/learn>

常用参数表：

|  |  |
| --- | --- |
| 名称 | 说明 |
| step | 训练迭代的总次数。一步计算一批样本产生的损失 |
| batch size | 单步的样本数量，随机选择。例如SGD的批次大小为1。 |
| periods | 控制报告的粒度，例如periods为7，steps为70，则每10步输出一次损失值。 |

泛化

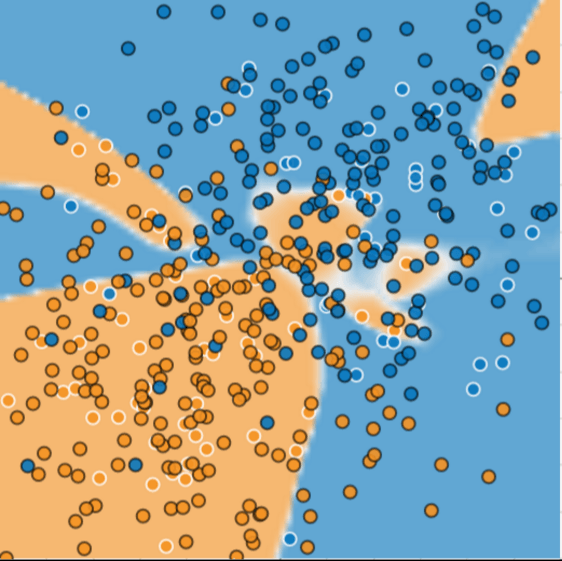


图 4 过拟合的模型

如上图所示模型，在训练过程中损失很低，但是在预测新数据方面却表现不一定很好。过拟合是由于模型的复杂程度超过了所需程度造成的。

机器学习哲学之一：

机器学习模型越简单，良好的实证结果就越有可能不仅仅基于样本的特性。

——奥卡莫剃刀定律

训练集和测试集

训练集：用于训练模型的子集。

测试集：用于测试训练后模型的子集，至少满足以下两个条件：

* 规模足够大，可以产生具有统计意义的结果
* 能够代表整个数据集，也就是挑选的测试集的特征应该与训练集的特征相同。**不可以对测试数据进行训练！**如果评估指标取得了意外的好结果，可能是训练时不小心对测试集进行了训练。

验证

基于给定的测试集执行评估的次数越多，不知不觉地过拟合该测试集地风险就越高。通过将数据集划分为训练集和测试集，可以判定给定模型是否能够很好的泛化到新的数据。

将样本集分成训练集、验证集、测试集三部分，目的：降低过拟合的发生率。

过程：使用训练集进行训练，使用验证集进行调整模型，使用测试集确认最终的模型效果。

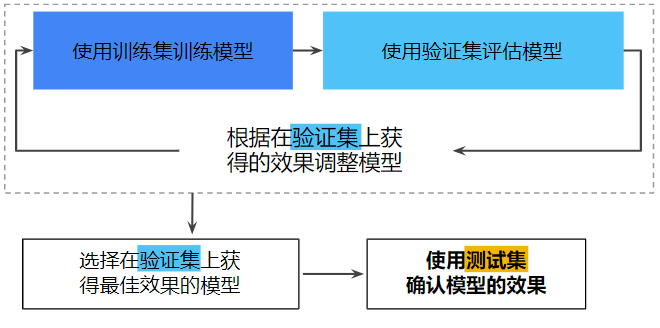


图 5 更好的工作流程

改良原因：暴露给测试集的信息更少。

表示法

必须提供给机器学习模型有用的信号来了解数据的关键特性，因此必须创建数据表示。也就是说，为了训练模型，必须选择最能代表数据的特征集。

特征工程：传统编程关注于代码，机器学习项目中，关注点为特征表示。开发者需要通过添加和改善特征来调整模型。

原始数据映射到特征：如下图所示，左侧表示原始数据，右侧表示特征矢量。

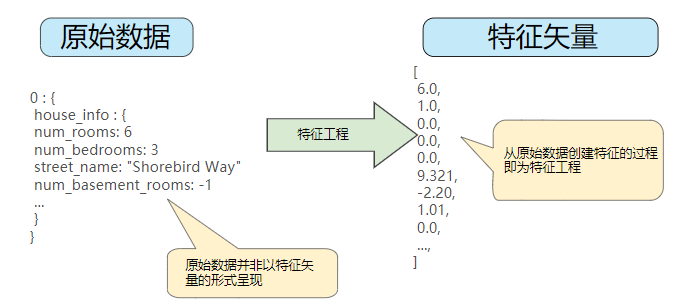


图 6 特征工程将原始数据映射到机器学习特征

特征工程指的是将原始数据转换为特征矢量，该过程需要大量时间。许多机器学习模型都需要将特征表示为实数向量，因为特征值必须与模型权重相乘。

特征组合

正则化：简单性正则

逻辑回归

分类

正则化：稀疏性正则

神经网络简介

多类别神经网络

嵌入

静态训练与动态训练

静态推理与动态推理

数据依赖关系

应用准则