无人驾驶汽车系统入门（八）——机器学习入门

原创 置顶 AdamShan 2017-12-29 12:31:38 13759 版权

收藏 28

分类专栏： 无人驾驶汽车专题 无人驾驶汽车系统入门 文章标签： 机器学习 计算机视觉 无人驾驶 无人车

无人驾驶汽车系统入门（八）——机器学习入门

> 在上一章中，我们介绍了基于传统的计算机视觉的技术实现的车道线检测，在这个过程中我们不难发现，使用传统的计算机视觉，往往需要人为地设计特征，这些特征对于不同的任务来说是不同的，车道线检测和行人检测要分别设计特征，另一方面，人为地设计特征往往会存在疏漏，对于无人驾驶汽车来说，忽视了某种情况的程序设计缺陷可能会造成严重的后果。
> 在本节我们开始了解机器学习——一类模式识别方式，在后面的章节中我们还是进一步接触深度学习。机器学习以及深度学习在无人驾驶领域有着广泛的应用，不仅能够提高感知的精度，在决策控制方面也有很大的研究和应用价值。

机器学习，是无人驾驶技术树中极其重要的一环，其中的深度学习更是近年来研究的热点。掌握机器学习的基本理论，是端到端无人驾驶研究，无人驾驶的行为克隆，强化学习控制，基于深度学习的视觉感知等研究的第一步。本节重点讲解机器学习的基本模式，不涉及算法和模型，为读者描述处理机器学习任务的过程。

基本概念

我们从一个实例来了解机器学习的基本概念。假设我们现在面临这样一个 **任务(Task)** ,任务的内容是识别手写体的数字。对于计算机而言，这些手写数字是一张张图片，如下所示：

对人来说，识别这些手写数字是非常简单的，但是对于计算机而言，这种任务很难通过固定的编程来完成，即使我们把我们已经知道的所有手写数字都存储到数据库中，一旦出现一个全新的手写数字（从未出现在数据库中），固定的程序就很难识别出这个数字来。所以，在这里，我们的 **任务** 指的就是这类很难通过固定编程解决的任务。要解决这类任务，我们的计算机需要有一定的“智能”，但是在我们的认知中，只有人类才具备这种“高级智能”（某些灵长类动物虽然具备一定的运用工具的能力，但我们认为那距离我们所说的智能还有很远的距离），所以如果我们想让计算机具备这种“智能”，由于这是人造的事物，我们称这种智能为 **人工智能（Artificial Intelligence, AI）** 。正式地讲，**人工智能，是指由人制造出来的机器所表现出来的智能。通常人工智能是指通过普通计算机程序的手段实现的类人智能技术。** 机器学习可以帮助我们解决这类 **任务** ，所以我们说，机器学习是一种人工智能技术。

那么机器学习是怎么解决这类任务的呢？

机器学习（Machine learning）是一类基于数据或者既往的经验,优化计算机程序的性能标准的方法。 这是机器学习的定义，看起来可能难以理解，我们对它进行分解：

1. 首先，对于手写数字识别这个任务来说，**数据或者既往的经验** 就是我们已经收集到的手写数字，我们要让我们的程序从这些数据中 **学习** 到一种 **能力/智能** ，这种能力就是：通过学习，这个程序能够像人一样识别手写数字。
2. **性能标准** 就是指衡量我们的程序的这种 **能力** 高低的指标了。在识别任务中，这个指标就是识别的精度。给定100个手写数字，有99个数字被我们的“智能”程序识别正确，那么精度就是 99% 。
3. **优化** 就是指我们基于既往的经验或者数据，让我们的“智能”程序变得越来越聪明，甚至比人类更加聪明。

机器学习，就是能够从经验中不断“学习进步”的算法，在很多情况下，我们将这些经验用数值描述，因此，**经验=数据** ，这些收集在一起的数据被成为 **数据集（Dataset）** ，在这些已有的数据集上学习的过程我们称之为 **训练（Train）** ，因此，这个数据集又被成为 **训练集** ，很显然，我们真正关心的并不是机器学习算法在训练集上的表现，我们希望我们的“智能”程序对从未见过的手写字也能够正确的识别，这种在新的样本（数据）上的性能我们称之为 **泛化能力（generalization ability）** ，对于一个任务而言，泛化能力越强，这个机器学习算法就越成功。

根据数据集的不同，机器学习可以分成如下三类：

- **监督学习（Supervised learning）**：数据集既包含样本（手写字图片），还包含其对应的标签（每张手写字图片对应的是那个数字）
- **无监督学习（Unsupervised learning）**：与监督学习相对，数据集仅包含样本，不包含样本对应的标签，机器学习算法需要自行确定样本的类别归属
- **强化学习（Reinforcement learning）**：又称为增强学习，是一种半监督学习，强调如何基于环境而行动，以取得最大化的预期利益。我们在后面的文章中会重点介绍。

当前大热的神经网络，深度学习等等都是监督学习，随着大数据时代的到来以及GPU带来的计算能力的提升，监督学习已经在诸如图像识别，目标检测和跟踪，机器翻译，语音识别，自然语言处理的大量领域取得了突破性的进展。然而，**当前在无监督学习领域并没有取得像监督学习那样的突破性进展。**由于在无人驾驶领域主要应用的机器学习技术仍然是监督学习，本文将重点讲监督学习的相关内容。当然，在后续的文章中，我还将介绍强化学习在无人驾驶领域的研究。

> 在本文中，为了便于读者理解，我们使用手写数字识别来描述处理的任务，实际上，机器学习算法能够处理的任务还有很多，例如：分类，回归，转录，机器翻译，结构化输出，异常检测，合成与采样， 缺失值填补等等。这些任务看似不同，却有着一个共性，那就是很难通过人为设计的确定性程序来解决。

监督学习

经验风险最小化

监督学习，本质上就是在给定一个集合 (X, Y) 的基础上去学得一个函数：

$$y = f(x)$$

在MNIST问题中， X 就表示我们收集到的所有的手写数字图片的集合，Y 表示这些图片对应的真实的数字，函数 f 则表示输入一张手写字图片，输出这张图片表示的数值这样的一个映射关系。

很显然，这样的映射关系中的 x 有着一个极其巨大的取值域（甚至有无限种可能取值）， 所以我们可以把我们已有的样本集合 (X, Y) 理解为从某个更大甚至是无限的母体中，根据某种未知的概率分布 p ，以独立同分布随机变量方式来取样。现在，我们假定存在一个 **损失函数(Loss function)** L ，这个损失函数可以表述为：

$$L(f(x), y)$$

这个损失函数描述的是我们学得的函数 $f(x)$ 的输出和 x 样本对应的真实值 y 之间的距离，很显然，这个损失越小，表示我们学得的函数 f 更贴近于真实映射 g 。以损失函数为基础，我们定义 **风险** ：

函数 f 的风险，就是损失函数的期望值。 由于我们以手写字分类为例，所以这里各个样本的概率分布 p 是离散的，我们可以用如下公式定义风险：

$$R(f) = \sum_i L(f(x_i), g(x_i))p(x_i)$$

如果是连续的，则可以使用定积分和概率密度函数来表示。这里的 x_i 是指整个样本空间的所有可能取值，所以，现在的目标就变成了： 在很多很多可能的函数中，去寻找一个 f，使得风险 $R(f)$ 最小 。然而，真实的风险是建立在对整个样本空间进行考量的，我们并不能获得整个样本空间，我们有的只是一个从我们要解决的任务的样本空间中使用独立同分布的方法随机采样得到的子集 (X, Y)，那么，在这个子集上，我们可以求出这个真实分布的近似值，比如说经验风险 ：

$$\bar{R}(f) = \frac{1}{n}\sum_{i=1}^{n} L(f(x_i), y_i)$$

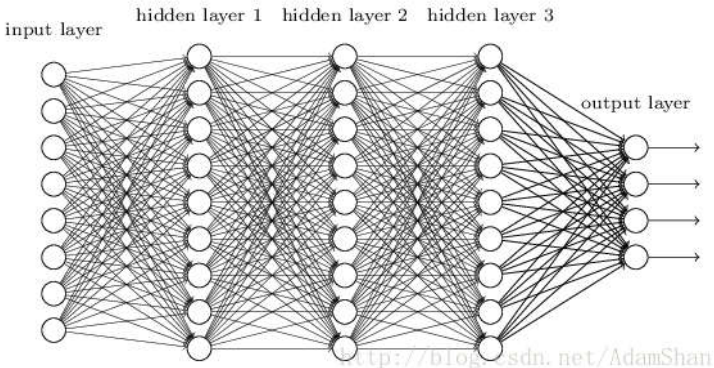
其中 (x_i, y_i) 是我们已有的数据集中的样本，所以，我们选择能够最小化经验风险的函数 f 这样的一个策略就被称之为 经验风险最小化原则 。

很显然，当训练数据集足够大的时候，经验风险最小化这一策略能够保证很好的学习效果——这也就是我们当代深度神经网络取得很多方面的成功的一个重要原因。专业的说，我们把我们已有的数据集的大小称之为 样本容量 。不论是什么应用领域，规范的大数据集合，就意味着我们的机器学习任务已经成功了一半。

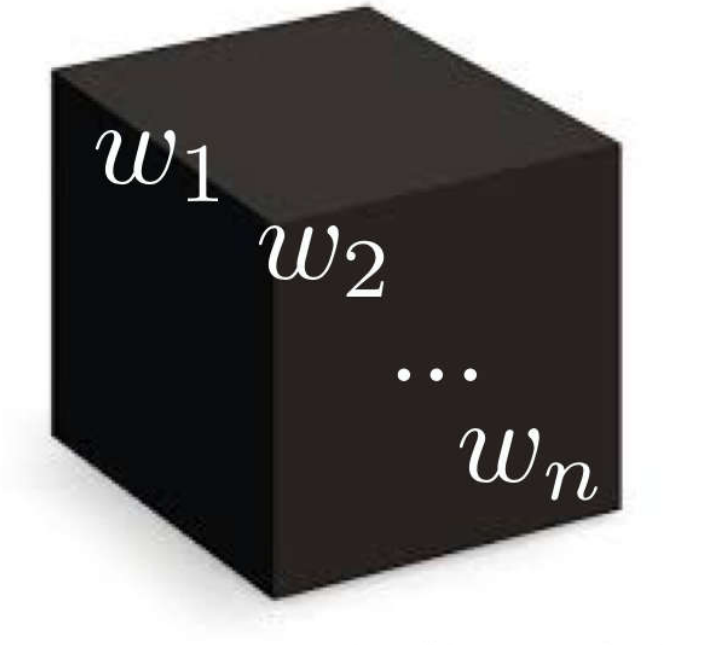
模型，过拟合，欠拟合

那么学习这个 f 需要一个 载体 , 这个载体的作用就是，用它我们可以表述各种各样的函数 f 这样我们就可以通过调整这个载体去选择一个最优的 f ,这个最优的 f 能够使经验风险最小化，这个载体我们专业地说，就是机器学习中的 模型（model） , 单纯地说模型的抽象概念可能让人难以理解，我们选取一种模型的实例来看。

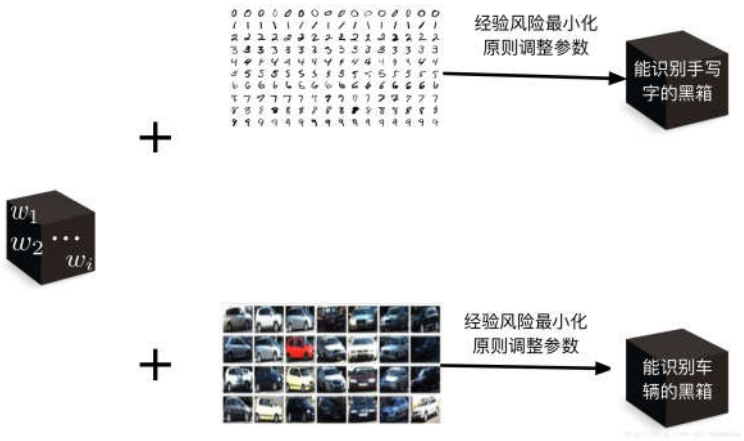
我们以 人工神经网络（artificial neural network，ANN） 为例来讨论。首先，我们知道我们现在需要的是一个模型，这个模型具有能够描述各种各样的函数的能力，下图是一个神经网络：



它看起来很复杂，让人费解，那么我们把它简化，如下图：



我们把这个模型理解成一个黑箱，这个黑箱里有很多参数：
$(w_1, w_2, w_3, \ldots, w_n)$**，我们用** W **来描述这个黑箱中的参数，这些参数叫 模型参数，即使模型内部的结构不变，仅仅修改这些参数，模型也能表现出不同的本领，具体来说：对于手写字识别任务，我们在手写字数据集上通过 一定的算法 调整神经网络的参数，使得神经网络拟合出一个函数** f **，这个** f **是经验风险最小化的函数，那么我们训练出来的这个“黑箱”就可以用于手写字识别了;另一方面，对于车辆识别来说，假设我们有车辆数据集，相同的思路，我们可以训练出一个黑箱来最做车辆识别。如下图所示：**



在前文中我们知道，考量一个机器学习模型的关键在于其泛化能力，一个考量泛化能力的重要指标就是模型的训练误差和测试误差的情况：

- 训练误差：模型在训练集上的误差
- 测试误差：模型在从未“见过的”测试集上的误差

这两个误差，分别对应了机器学习任务中需要解决的两个问题： 欠拟合 和 过拟合 。当训练误差过高时，模型学到的函数并没有满足经验风险最小化 ,对手写字识别来说，模型即使在我们的训练集中识别的精度也很差，我们称这种情况为欠拟合。当训练误差低但是测试误差高，即训练误差和测试误差的差距过大时，我们称之为过拟合，此时模型学到了训练集上的一些“多余的规律”，表现为在训练数据集上识别精度很高，在测试数据集（未被用于训练，或者说未被用于调整模型参数的数据集合）上识别精度不高。

模型的 容量（capacity） 决定了模型是否倾向于过拟合还是欠拟合。模型的容量指的是模型拟合各种函数的能力，很显然，越复杂的模型就能够表述越复杂的函数（或者说规律，或者说模式）。那么对于一个特定的任务（比如说手写字识别），如何去选择合适的模型容量来拟合相应的函数呢？这里就引入了 奥卡姆剃刀原则 ：

> 奥卡姆剃刀原则:在同样能够解释已知观测现象的假设中,我们应该挑选”最简单”的那一个。

这可以理解为一个简约设计原则，在处理一个任务是，我们应当使用尽可能简单的模型结构。

“一定的算法” –>梯度下降算法

前面我们说到我们可以通过 一定的算法 调整神经网络的参数，这里我们就来介绍一下这个 定向（朝着经验风险最小化的方向）调整模型参数的算法——梯度下降算法 。

要最小化经验风险 $\bar{R}(f)$，等同于最小化损失函数，在机器学习中，损失函数可以写成每个样本的损失函数的总和：

$$L(\theta)=\frac{1}{n}\sum_{i=1}^{n}L(x_i,\,y_y,\,\theta)$$

其中 θ 表示模型中的所有参数，现在我们要最小化 $L(\theta)$，我们首先想到的是求解导数，我们把这个 L 对 θ 的导数记作 $L'(\theta)$ 或者 $\frac{dL}{d\theta}$ ，导数 $L'(\theta)$ 就代表了函数 $L(\theta)$ 在 θ 处的斜率，我们可以把函数的输入输出关联性用斜率来描述：

$$L(\theta+\alpha)\approx L(\theta)+\alpha L'(\theta)$$

其中， α 是一个变化量，利用这个公式，我们就可以利用导数来逐渐使 L 变小，具体来说，我们只要让 α 的符号和导数的符号相反，即：

$$sign(\alpha) = -sign(L'(\theta))$$

这样，$L(\theta+\alpha)$ 就会比原来的 $L(\theta)$ 更小：

$$L(\theta+\alpha) = L(\theta) - |\alpha L'(\theta)|$$

这种通过向导数的反方向移动一小步来最小化目标函数（在我们机器学习中，也就是损失函数）的方法，我们称之为 梯度下降（gradient descent）。对于神经网络这种复杂的模型来说，模型包含了很多参数，所以这里的 θ 就表示一个参数集合，或者说参数向量，所以我们要求的导数就变成了包含所有参数的偏导数的向量 $\nabla_\theta L(\theta)$。这里的 α 就可以理解为我们进行梯度下降的过程中的步长了，我们将学习的步长称为 学习率（learning rate），它描述了梯度下降的速度。

小结

在本节中，我们没有介绍任何一种具体的机器学习算法和模型，但是我们快速的了解了机器学习任务中的重要成分和结构，以下我们来进行一个小的总结：

1. 首先，机器学习是用来完成特定的 **任务** 的:比如说手写字识别，行人检测，房价预测等等。这个任务必须要有一定的 **性能度量**，比如说识别精度，预测误差等等。
2. 然后，为了处理这个 **任务**，我们需要设计 **模型**，这个模型能够从 **数据** 中基于一定的 **策略** (比如说经验风险最小化原则) 和一定的 **算法** (比如说梯度下降算法) 去 **学习** 一个 **函数**。
3. 最后，这个 **函数** 要能够处理这个任务中的各种各样的情况（包括没有出现在训练集中的情况），这个模型要有很好的 **泛化能力**，这样，我们的机器学习任务就成功了。

之后的文章会介绍各种各样的任务，模型，算法，但是总的来说还是遵照这样的基本模式，机器学习尤其是深度学习在无人驾驶的研究中起着非常重要的作用，我们将逐步深入学习无人驾驶中的机器学习算法。

AdamShan 专家 图像处理 深度学习 TensorFlow

奔驰高级自动驾驶扫地僧，谷歌认证机器学习专家，兰州大学无人驾驶团队创始人，主攻深度学习，无人驾驶汽车方向，著有《无人驾驶原理与实践》一书。