分享到

机器之心

2219 文章 | 700万 总阅读

查看TA的文章>

深度 | 理解神经网络中的目标函数

2017-12-12 10:39

选自Kdnuggets

作者：Lars Hulstaer

参与：晏奇、李泽南

本文面向稍有经验的机器学习开发者，来自微软的 Lars 神经网络的几种目标函数。

介绍

本文的写作动机有以下三个方面：

首先，目前有很多文章都在介绍优化方法，比如如何对随机梯度下降进行优化，或是提出一个该方法的变种，很少有人会解释构建神经网络目标函数的方法。会去回答这样的问题：为什么将均方差（MSE）和交叉熵损失分别作为回归和分类任务的目标函数？为什么增加一个正则项是有意义的？所以，写作这篇博文的意义在于，通过对目标函数的考察，人们可以理解神经网络工作的原理，同时也就可以理解它们为何在其他领域却无法发挥作用。

$$CE = -\sum_{x} p(x) \log q(x)$$

在分类任务中，（监督学习中）正确的标注 p（ground truth）与网络输出 q 之间的交叉熵损失。

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \tilde{y}_i)^2$$

在回归任务中，（监督学习中）正确的标注 y 与网络输出 y_tilde 之间的均方差。

其次，神经网络作出错误概率预测是出了名的，并且，面对对抗性样本（adversarial example，即一种特殊的输入数据，它们由研究人员专门设计，用来让神经网络作出错误预测）它们也毫无办法。总之，神经网络经常过度自信，甚至当它们判断错误时也这样。这个问题在真实环境中可不容忽视，以自动驾驶为例，一辆自动驾驶汽车要保证在 145km/h 的

24小时热文

1 宁静坦白允许男人
在外面有十个女人

2 苹果重磅新品被移
死！国人爽

3 iOS 11坑惨iPhone
苹果神回应


FF完成超10亿美元
亭出任公司CEO


FF 称已完成 10 亿
跃亭出任 CEO

不是王宝强！熊乃瑾与小鲜肉恋情曝光

犹太人穆斯撒冷到底是


中国首次站在世界之巅：最先攻克划时代武器


俄罗斯美女用笔有点猛

24小时热文

1 一图看清美国正在
业

2 陪别人睡觉可以减
学生赚数万美金

3 iOS 11坑惨iPhone
苹果神回应


FF完成超10亿美元
亭出任公司CEO


FF 称已完成 10 亿
跃亭出任 CEO

搜狐号推荐

搜狐科技视界
搜狐科技官方原创账号。聚焦
件、大趋势和新变化，用我们

互联网圈内事官方
每日呈现最新热门有料内容，
事件、解读互联网百态人生，

IT之家
IT之家是业内领先的即时IT资
网站。IT之家快速精选泛科技

创业最前线
创业投资"第一自媒体"平台，
上的创业者、投资人关注，其

猎云网
猎云网是一家科技新媒体，聚
势、创业创新报道，关注新产

具优点，还要知道其缺点。

分享到

一直以来，我都想明白为何神经网络可以从概率的角度来加以解释，以及它们为什么适合作为广义的机器学习模型框架。人们喜欢把网络的输出作为概率来讨论。那么，神经网络的概率解释与其目标函数之间是否存在联系呢？

写作这篇文章的灵感来源于作者和其朋友 Brian Trippe 在剑桥大学计算与生物学习实验室工作期间对贝叶斯神经网络的研究，作者高度推荐读者朋友阅读其朋友 Brian 关于神经网络中变分推理的论文《Complex Uncertainty in Machine Learning: Bayesian Modeling for Conditional Density Estimation and Synaptic Plasticity》。

监督学习

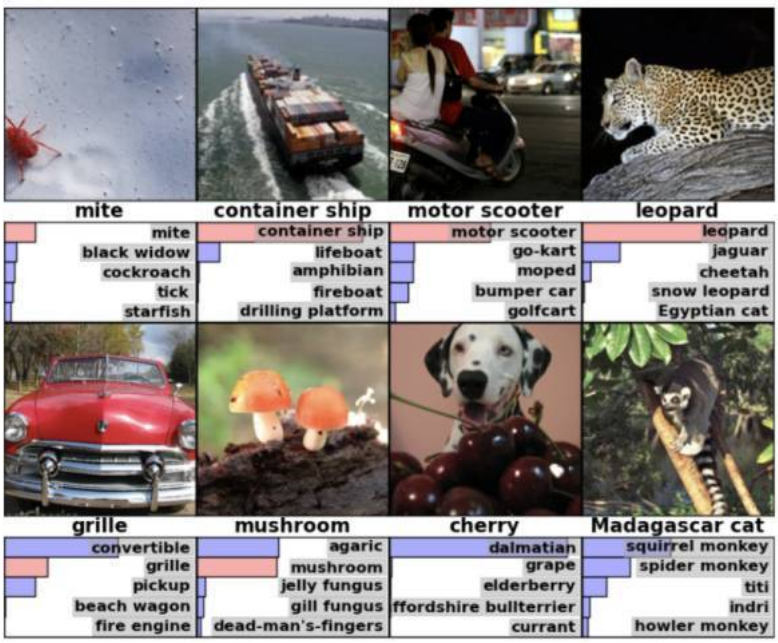
在监督学习问题中，我们一般会有一个数据集 D，x 是其中的样本，y 是样本标签，我们用（x, y）的方式来表示样本，我们要做的，是对 P(y | x, θ) 这个条件概率分布进行建模。

举个例子，在图像分类任务中，x 表示一个图像，y 表示与之对应的图像标签。P(y | x, θ) 表示：在图像 x 和一个由参数θ定义的模型下，出现标签 y 的概率。

按照这种方法建立的模型被称为判别式模型（discriminative model）。在判别式或条件模型中，定义条件概率分布函数 P(y|x, θ) 的参数θ是从训练集中推出的。

基于观察数据 x（输入数据或特征值），模型输出一个概率分布，之后会用这个分布来预测标签 y（类别或真值）。不同的机器学习模型要求预测不同的参数。对于线性模型（如：逻辑回归，由一系列值等于特征数量的权重来定义）与非线性模型（如：神经网络，由其每一层的一系列权重所定义）而言，这两类模型都可以近似等于条件概率分布。

对于典型的分类问题而言，（一系列可被学习的）参数θ用作定义一个 x 到范畴分布（它们基于不同的标签）的映射。一个判别式模型会将概率 N（N 等于类的数量）作为输出。每个 x 都属于一个单独的类，但是模型的不确定性是由在类上输出的一个分布来反映的。一般来说，概率最大的类会在做出决定的时候被选择。



24小时热文

1 宁静坦白允许男人
在外面有十个女人

2 苹果重磅新品被移
死！国人爽

3 iOS 11坑惨iPhone
苹果神回应


FF完成超10亿美元
亭出任公司CEO


FF 称已完成 10 亿
跃亭出任 CEO

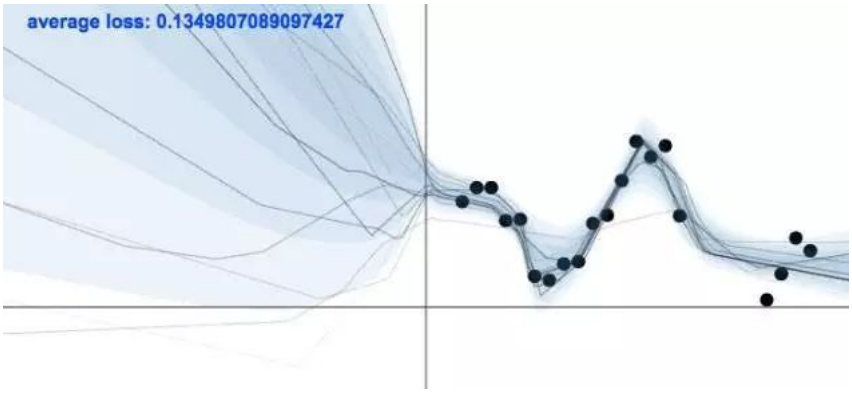
个类（以概率大小为标准筛选）。

分享到

我们注意到，判别式回归模型（discriminative regression model）经常只会输出一个预测值，而不是一个基于所有真值的分布。这与判别式分类模型（discriminative classification model）不同，后者会输出一个基于可能的类的分布。那么这是否意味着判别式模型因回归任务而瓦解了呢？模型的输出难道不应该告诉我们哪些回归值（regression value）会比其它值更有可能吗？

说判别式回归模型只有一个输出其实会让人误解，实际上，一个回归模型的输出与一个著名的概率分布有关：高斯分布。事实证明，判别式回归模型的输出代表了一个高斯分布的均值（一个高斯分布完全由一个均值与标准差决定）。有了这个信息，你就可以在输入*x*的情况下决定每个真值的相似度了。

通常，只有这个分布的均值才会建模，高斯分布的标准差要么没有建模，要么就是在所有 x 上保持一个常值（constant）。因此，在判别式回归模型中，θ规定了从 x 到高斯分布（y 从中取样得来）均值的一个映射。基本上每当要做出决定时，我们都会选择均值，因为模型能够通过提高标准差来表达哪个 x 是不确定的。



当没有训练数据的时候，一个模型是需要保持不确定的，相反，当有训练数据的时候，模型需要变得确定。上图展示了这样的一个模型，图片来自 Yarin Gal 的博文。

在回归问题里，其他的概率模型（比如高斯过程）在对不确定性进行建模的过程中效果好得多。因为当要同时对均值与标准差建模的时候，判别式回归模型会有过于自信的倾向。

高斯过程（Gaussian process）可以通过对标准差精确建模来量化不确定性。其仅有的一个缺点在于，高斯过程不能很好地扩大到大型数据集。在下图中你可以看到，GP 模型在具有大量数据的区域周围置信区间很小。在数据点很少的区域，置信区间又变得很大。

24小时热文

1 宁静坦白允许男人
在外面有十个女人

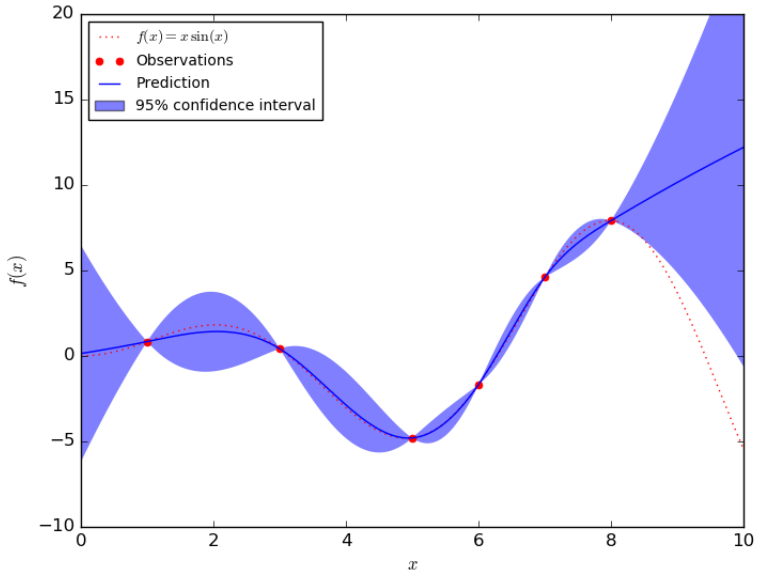
2 苹果重磅新品被移
死！国人爽

3 iOS 11坑惨iPhone
苹果神回应


FF完成超10亿美元
亭出任公司CEO


FF 称已完成 10 亿
跃亭出任 CEO

分享到



24小时热文

1 宁静坦白允许男人
在外面有十个女人

2 苹果重磅新品被移
死！国人爽

3 iOS 11坑惨iPhone
苹果神回应


FF完成超10亿美元
亭出任公司CEO


FF 称已完成 10 亿
跃亭出任 CEO

GP 模型在数据点上是确定的，但是在其他地方是不确定的（图片来自 Sklearn）。

通过在训练集上训练，判别式模型可以学习数据（代表了一个类或是真值）中的特征。如果一个模型能够将高概率赋给正确地的样本类，或是一个接近测试集中真值（true value）的均值（mean），那么我们说这个模型表现的不错。

链接神经网络

当用神经网络来进行分类或回归任务时，上述提到的参数分布（范畴分布与高斯分布）的建模就通过神经网络来完成。

这一点在当我们要决定神经网络参数θ的最大似然估计（MLE）的时候比较清楚。MLE 相当于找到训练数据集似然度（或等效对数似然度）最大时的参数θ。更具体的来说，下图的表述得到了最大化：

$$\theta^{MLE} = \text{argmax}_\theta \log p(Y|X,\theta)$$

$$= \text{argmax}_\theta \log \prod_{i=1}^{N} p(y_i|x_i,\theta)$$

$$= \text{argmax}_\theta \sum_{i=1}^{N} \log p(y_i|x_i,\theta)$$

新闻 体育 汽车 房产 旅游 教育 时尚 科技 财经 娱乐 更多

当 p(Y | X, θ) 由模型确定时，它表示了训练数据中真实标签的概率。如果 p(Y | X, θ) 接近于 1，这意味着模型能够确定训练集中正确的标签/均值。在给定由 N 个观察对组成的训练数据（X,Y）的条件下，训练数据的似然度可被改写成对数概率的总和。

在分类与回归的情况下，p(y|x, θ) 作为一个（x, y）的后验概率，可以被改写成范畴分布和高斯分布。在优化神经网络的情况下，目标则是去改变参数，具体方式是：对于一系列输入 X，概率分布 Y 的正确的参数可以在输出（回归值或类）中得到。一般这可以通过梯度下降

输出：

最大化一个范畴分布的对数值相当于最小化真实分布与其近似分布的交叉熵。

最大化高斯分布的对数值相当于最小化真实均值与其近似均值的均方差。

因此，前述图片中的表达就可以被改写，分别变成交叉熵损失和均方差，以及分类和回归的神经网络的目标函数。

相较于更加传统的概率模型，神经网络从输入数据到概率或是均数习得的非线性函数难以被解释。虽然这是神经网络的一个显著的缺点，但是其可以模拟大量复杂函数的能力也带来了极高的好处。根据这部分衍生讨论的内容，我们可以明显看到，神经网络的目标函数（在确定参数的 MLE 似然度过程中形成）可以以概率的方式来解释。

神经网络一个有趣的解释与它和那些一般的线性模型（线性回归、逻辑回归）的关系有关。相比于选择特征的线性组合（就像在 GLM 做的一样），神经网络会产生一个高度非线性的特征组合。

最大后验概率（MAP）

但是如果神经网络可以被解释成概率模型，那为什么它们给出的概率预测质量很差，而且还不能处理那些对抗性样本呢？为什么它们需要这么多数据？

在选择好的函数逼近器时，根据不同的搜索空间我倾向于选择不同的模型（逻辑回归，神经网络等等）。当面对一个极大的搜索空间，也即意味着你可以很灵活地模拟后验概率时，依然是有代价的。比如，神经网络被证明是一个通用的函数逼近器。也就是说只要有足够的参数，它们就可以模拟任何函数。然而，为了保证函数在整个数据空间上能够得到很好的校准，一定需要极大的数据集才行。

通常，一个标准的神经网络都会使用 MLE 来进行优化，知道这一点很重要。使用 MLE 进行优化可能会让模型发生过拟合，所以模型需要大量数据来让过拟合问题减弱。机器学习的目标不是去寻找一个对训练数据解释度最好的模型。我们更需要的是找到一个可以在训练集外的数据上也有很好泛化能力的模型。

在这里，最大后验概率（MAP）方法是一个有效的可选方案，当概率模型遭遇过拟合问题时我们经常会使用它。所以 MAP 相当于神经网络的语境下的什么呢？对于目标函数它会有什么影响呢？

与 MLE 类似，MAP 也可以在神经网络的语境下被改写成一个目标函数。就本质而言，使用了 MAP 你就是在最大化一系列参数θ（给定数据下，在θ上假设一个先验概率分布）的概率：

$$\theta^{MAP} = \text{argmax}_\theta \log p(\theta|X,Y)$$
$$\approx \text{argmax}_\theta \log p(Y|X,\theta) \cdot p(\theta)$$

使用 MLE 时，我们只会考虑方程的第一个元素（模型在何种程度上解释了训练数据）。使用了 MAP，为了降低过拟合，模型满足先验概率也很重要（θ在何种程度上满足先验概率）。

分享到

24小时热文

1 宁静坦白允许男人
在外面有十个女人

2 苹果重磅新品被移
死！国人爽

3 iOS 11坑惨iPhone
苹果神回应


FF完成超10亿美元
亭出任公司CEO


FF 称已完成 10 亿
跃亭出任 CEO

多小权重），然而在θ上使用一个拉普拉斯先验概率与把 L1 正则化应用到目标函数上是一致的（确保很多权重的值为 0）。

分享到

$$\lambda \sum_{i=1}^{N} |w_i| \qquad \lambda \sum_{i=1}^{N} (w_i)^2$$

左边是 L1 正则化，右边是 L2 正则化。

一种完全贝叶斯方法

在 MLE 和 MAP 两种情况中，都只使用了一个模型（它只有一组参数）。对于复杂的数据尤其如此，比如图像，数据空间中特定的区域没有被覆盖这个问题不太可能出现。模型在这些地方的输出由模型的随机初始化与训练过程决定，模型对处于数据空间覆盖区域之外的点会给出很低的概率估计。

尽管 MAP 保证了模型在这些地方的过拟合程度不会太高，但是它还是会让模型变得过于自信。在完全贝叶斯方法中，我们通过在多个模型上取平均值来解决这个问题，这样可以得到更好的不确定性预测。我们的目标是模拟参数的一个分布，而不是仅仅一组参数。如果所有的模型（不同参数设置）在覆盖区域之外都给出了不同的预测，那么这意味着这个区域有很大的不确定性。通过对这些模型取平均，最终我们会得到一个在那些区域不确定的模型，这正是我们想要的。

原文链接：https://www.kdnuggets.com/2017/11/understanding-objective-functions-neural-networks.html

本文为机器之心编译，转载请联系本公众号获得授权。 返回搜狐，查看更多

声明：本文由入驻搜狐号的作者撰写，除搜狐官方账号外，观点仅代表作者本人，不代表搜狐立场。

阅读 (1108) 不感兴趣 投诉

本文相关推荐

神经网络激活函数 深度卷积神经网络 二次函数根的判别式
非线性目标函数的最值 目标函数+约束条件 函数的表示方法的教学目标
bp神经网络模型 python+map函数 excel中标准差的函数
matlab求均值函数 函数模型及其应用教案 几类不同增长的函数模型

我来说两句 0人参与，0条评论

来说两句吧……

登录并发表

搜狐“我来说两句”用户公约

24小时热文

1 宁静坦白允许男人
在外面有十个女人

2 苹果重磅新品被移
死！国人爽

3 iOS 11坑惨iPhone
苹果神回应


FF完成超10亿美元
亭出任公司CEO


FF 称已完成 10 亿
跃亭出任 CEO