序列数据的有监督标注
——循环神经网络

[著] Alex Graves
[译]xxxxx

2015 年 10 月 6 日

摘要

循环神经网络是一种强有力的学习方法，它们可以灵活地使用上下文信息，并且对数据的局部变形具有较强的鲁棒性。在序列的标注这个任务上，输入的序列被转换成流式的标签，循环神经网络的特性使得它们十分适用于这个任务。LSTM（Long short-term memory）是一种非常有前景的循环结构，它可以在输入和输出间构建持久的延迟，因此可以保留很长的上下文信息。这篇文章的关注于以下几点内容，首先我们讨论目前在有监督序列标注这个任务上取得最好效果的模型 —循环神经网络，随后会讨论一个特殊的结构，也就是 LSTM。文章的的主要贡献在于 (1) 提出一种新的输出层形式，使得循环网络可以在输入和输出映射关系未知的情况下直接进行训练 (2) 将 LSTM 扩展到高维数据，例如图像或视频序列。我们将会展示语音识别、在线的和离线的手写数字识别、关键词标注、图像分割及分类的实验效果，用以说明循环神经网络对比其他时序算法（比如 HMM）的优点。

目录

第一章　导言

在机器学习中，序列标注是所有将序列数据转换为序列标签任务的总称。一些有名的任务包括语音识别、手写数字识别、蛋白质二级结构的预测以及词性标注。

$$\frac{\partial}{\partial\theta_j}J(\theta) = \frac{\partial}{\partial\theta_j} \tag{1.1}$$

1.1　文章主要贡献

1.2　文章概览

第二章　有监督的序列标注

本章主要介绍关于有监督序列标注的基本背景知识以及回顾相关文献。2.1小节简单的回顾有监督学习，2.2小节介绍有监督模式分类的经典、非序列框架，2.3小节提出有监督序列标注的定义，并介绍基于对标签序列的不同假设下的各种序列标注任务。

2.1　有监督学习

在机器学习中，使用 $<$ 输入 $\rightarrow$ 输出 $>$ 作为训练集的任务称为有监督学习。这与增强学习不同，在增强学习中，仅仅只能在训练过程中量化激励值。而在非监督学习中，将不会存在响应数据，算法只能通过不断的试探，寻求数据的结构。本书中，我们不会讨论增强学习和非监督学习。

有监督学习包含了一个训练集 S，集合的元素是 $<$ 输入 $\rightarrow$ 输出 $>$ 对，记为 (x, z)，其中，x 是输入空间 $\mathcal{X}$ 中的一个元素，z 是目标空间 $\mathcal{Z}$ 中的一个元素，伴随训练集一同出现的是测试集 S'，我们有时候也会将训练集称为训练样本。一般的，我们假定 S 和 S' 都是从相同的输入 -输出分布 $D_{\mathcal{X}\times\mathcal{Z}}$ 中独立地抽取出来。在一些案例中，我们会从训练集中选取一部分样本组合成验证集，用于训练过程中验证算法的性能，特别的，为了防止过拟合，往往会把验证集用于确定训练停止的时间。整个过程的目标是，使用测试集训练模型，最小化定义在测试集上的某个误差度量 E（这个误差度量往往会结合具体任务给出定义），例如，在回归问题中，经常使用的误差度量是平方和，或者输出值与实际值的欧几里得距离的平方。对于参数化方法（比如神经网络），最小化误差的最常用方法是逐步调整参数，使其最优化定义在训练集上的一个损失函数，尽可能地逼近 E。将训练集上学习到的模型应用到测试集的过程称之为泛化，我们会在本章稍后一点的地方讨论它。

对于不同的监督学习任务，监督者需要提供的基本特征以及监督力度有着很大的不同，例如，训练一个模型让它预测出图片中飞机的每一个像素点的取值要比让它学会识别图中是否存在一辆飞机更困难，前者需要提供更多的信息。为了区分这两种极端情况，人们有时会将这两种任务所需的数据称为强标签数据以及弱标签数据。

2.2　模式分类

模式分类，也称为模式识别，是机器学习中的一个广泛的研究领域，一些特定的分类器，比如多层感知器、支持向量机等，现在已在科研圈子广为人知。

尽管模式分类处理的是非序列数据，但底层的实践和理论的框架涵盖了序列数据的情况，因此在讨论序列标注之前，花一点时间简短的回顾它还是有意义的。

有监督模式分类的输入空间 $\mathcal{X}$ 一般记为 $\mathbf{R}^M$，也就是说，集合中的每个样本是一个固定长度 M 的实数向量。目标空间 Z 是一个包含了 K 个类别的离散空间。一个模式分类器 $h : \mathcal{X} \to Z$ 就是一个将向量映射到标签的函数。如果说所有的误分类情况都是坏情况，那么 h 最常用的误差度量是测试集 S' 上的误分率 $E^{class}(h, S')$

$$E^{class}(h, S') = \frac{1}{|S'|} \sum_{(x,z)\in S'} \begin{cases} 0 & \text{若} h(x) = z \\ 1 & \text{其他} \end{cases} \tag{2.1}$$

2.2.1　概率分类

分类器直接输出的是标签，正如支持向量机那样，我们也称这种分类器为判别式函数。另一种方法是概率分类器，它在给定输入模式 x 时返回属于 K 个类别中的 k 的概率 $p(C_k|x)$，然后选取最可能的类别作为分类器的输出 $h(x)$：

$$h(x) = \arg\max xp(C_k|x) \tag{2.2}$$

2.2.2　训练概率分类器

2.2.3　生成式模型和判别式模型

2.3　序列标注

2.3.1　序列标注任务的一种解决方法

2.3.2　序列分类

2.3.3　段分类

2.3.4　时序分类

第三章　神经网络

3.1　多层感知器

3.1.1　前馈

3.1.2　输出层

3.1.3　目标函数

3.1.4　反馈

数值梯度

3.2　循环神经网络

3.2.1　前馈

3.2.2　反馈

3.2.3　双向 RNN

BRNNs 和因果分析任务

3.2.4　连续的 Jacobian 矩阵

3.3　网络的训练

3.3.1　梯度下降算法

3.3.2　泛化

提前停止学习

噪声下学习

3.3.3　输入重表达

3.3.4　权值初始化

第四章　长短期记忆网络 (LSTM)

第五章　逐帧音素的分类

5.1　实验阶段

5.2　网络构型

5.2.1　计算复杂度

5.2.2　上下文范围

5.2.3　输出层

5.3　网络训练

5.3.1　重训练

5.4　实验结果

5.4.1　与之前工作对比

5.4.2　增加上下文的作用

5.4.3　权差