深度学习原理论文笔记

Leoeon

2016 年 10 月 20 日

Contents

1 神经网络拟合哈密顿量 **3**
1.1 神经网络有效性 3
1.1.1 理论目标 3
1.1.2 任意精度拟合 3
1.1.3 高效 4
1.2 分层 4

2 ε误差估计 **5**
2.1 结论 5
2.2 单变量函数上限 5
2.3 下限 5

3 机器学习展平流型 **7**

Chapter 1

神经网络拟合哈密顿量

Why does deep and cheap learning work so well? (2016.08.29)[1]

1.1 神经网络有效性

1.1.1 理论目标

多层网络可写为

$$f(\vec{y}) = \hat{\sigma}_n \hat{W}_n \cdots \hat{\sigma}_1 \hat{W}_1 \vec{y}$$

令：

$$H_x(\vec{y}) \stackrel{def}{=} -\ln p(\vec{y}|x)$$

$$\mu_x \stackrel{def}{=} -\ln p(x)$$

则：

$$p(x|\vec{y}) = \frac{1}{Z(\vec{y})} e^{-(H_x(\vec{y})+\mu_x)} = \hat{\sigma}(-(H_x(\vec{y})+\mu_x)) \left(Z(\vec{y}) = \int e^{-(H_x(\vec{y})+\mu_x)} dx \right)$$

故只要多层网络能估计$-(H_x(\vec{y})+\mu_x)$，则再加一层神经元为softmax即可

1.1.2 任意精度拟合

由$\sigma(x) = \sigma + \sigma' x + \sigma'' x^2 + O(u^3)$ 得：

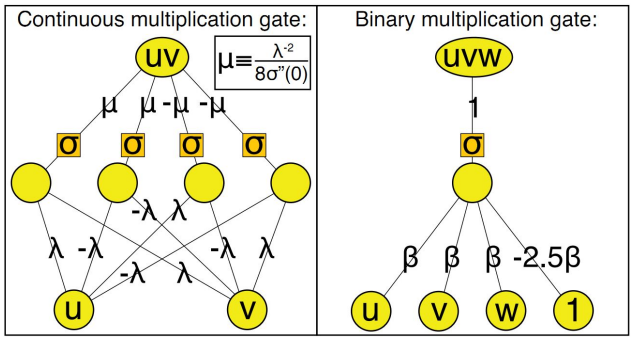
$$\frac{1}{8\sigma''\lambda^2}[\sigma(\lambda u + \lambda v) + \sigma(-\lambda u - \lambda v) - \sigma(\lambda u - \lambda v) - \sigma(-\lambda u + \lambda v)] = uv(1 + O(\lambda^2 u^2 + \lambda^2 v^2))$$

故理论上可用神经网络以任意精度表示乘积

微扰项

$$H_x(\vec{y}) = H + \sum_i H'_i y_i + \sum_{ij} H''_{ij} y_i y_j + \sum_{ijk} H'''_{ijk} y_i y_j y_k + \cdots$$

故理论上可用神经网络表示任意$H_x(\vec{y})$



1.1.3 高效

以下限制使得参数数量有限

*低阶项 $H_x(\vec{y})$ 只需展开到有限项，无需太高阶项

*局域性 只有短程作用，很多$H^{(k)}$ 为0

*对称性 很多$H^{(k)}$ 不独立，相互依赖

1.2 分层

类似重整化群理论，从低层次信息中抽象出（近似）统计充分量作为高层次信息，逐层提取直至所需。

Chapter 2

ε误差估计

Why Deep Neural Networks? (2016.10.13)[2]
神经网络$\tilde{f}(x)$拟合函数$f(x)$，满足$|f(x) - \tilde{f}(x)| \leq \varepsilon$

2.1 结论

以ε的误差拟合函数：

- 分段光滑函数，$\Theta(\log \frac{1}{\varepsilon})$层，$\mathcal{O}(polylog(\frac{1}{\varepsilon})$个神经元
- 分段光滑函数，$o(\log \frac{1}{\varepsilon})$层，$\Omega(poly(\frac{1}{\varepsilon})$个神经元
- 可微凸函数，$\Omega(\log \frac{1}{\varepsilon})$个神经元

2.2 单变量函数上限

2.3 下限

对强凸函数（$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \lambda ||x - y||_2^2$），$(x \in [0, 1])$，要求$N \geq L(\frac{\mu}{16\varepsilon})^{\frac{1}{2L}}$

Table 2.1: 单变量函数

拟合函数	层数	神经元个数
$f(x)=x^2\ (x\in[0,1])$	$O(\log\frac{1}{\varepsilon})$	$O(\log\frac{1}{\varepsilon})$
$f(x)=\sum_{i=0}^{p}a_ix^i\ (x\in[0,1],\sum_{i=0}^{p}\lvert a_i\rvert\le 1)$	$O(p+\log\frac{p}{\varepsilon})$	$O(p\log\frac{p}{\varepsilon})$
$f(x)\ (x\in[0,1])$ (对$\forall n\in[\lceil\log\frac{2}{\varepsilon}\rceil+1]$，$\lVert f^{(n)}\rVert_\infty\le n!$)	$O(\log\frac{1}{\varepsilon})$	$O((\log\frac{1}{\varepsilon})^2)$
$f=\prod_{i=1}^{k}h_i\ (x\in[0,1])$ (对$\forall n\in[\lceil 4k\log 4k+4k+2\log\frac{2}{\varepsilon}\rceil+1]$，$\lVert h_i^{(n)}\rVert_\infty\le n!$)	$O(k\log k+\log\frac{1}{\varepsilon})$	$O((k\log k)^2+(\log\frac{1}{\varepsilon})^2)$
$f(x)=h_1(h_2(\cdots(h_k(x))))\ (x\in[0,1])$ (对$\forall n\in[\lceil\log\frac{2}{\varepsilon}\rceil+1]$，$\lVert h_i^{(n)}\rVert_\infty\le n!$，$h_i:[0,1]\to[0,1]$)	$O(k\log k\log\frac{1}{\varepsilon}+\log k(\log\frac{1}{\varepsilon})^2)$	$O(k\log k\log\frac{1}{\varepsilon}+k^2(\log\frac{1}{\varepsilon})^2+(\log\frac{1}{\varepsilon})^4)$

Chapter 3

机器学习展平流型

Why Deep Learning Works: A Manifold Disentanglement Perspective (2014.11.07)[3]

原始数据为嵌在高维空间中的低维流型，且该流型很可能非常复杂。机器学习提取出特征，实质为将流型尽可能在低维空间中展平。

深度学习能自动提取特征，即为自动适应各类复杂流型，自动进行展平。优于传统机器学习中人工设计特征，局限于有限的流型。

流型展平程度判断依据：

- 流型上任意两点间测地线长度与两点间欧几里得空间长度比较，越接近则流型越平。
- 由流型上每个点切空间与周围点切空间求得该点曲率，曲率越小则流型越平。

Bibliography

[1] Henry W Lin and Max Tegmark. Why does deep and cheap learning work so well? *arXiv preprint arXiv:1608.08225*, 2016.

[2] Shiyu Liang and R Srikant. Why deep neural networks? *arXiv preprint arXiv:1610.04161*, 2016.

[3] Pratik Prabhanjan Brahma, Dapeng Wu, and Yiyuan She. Why deep learning works: A manifold disentanglement perspective. 2015.