Beijing Forest Studio
北京理工大学信息系统及安全对抗实验中心

深度模型可解释方法——树正则化

张寒青　硕士研究生

2018年03月25日

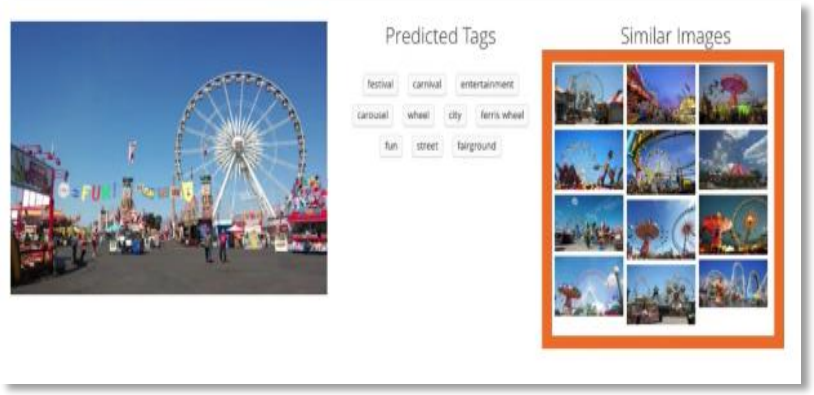
内容提要

- 背景简介
- 基本知识
- 算法原理
- 实验分析
- 应用总结

背景简介

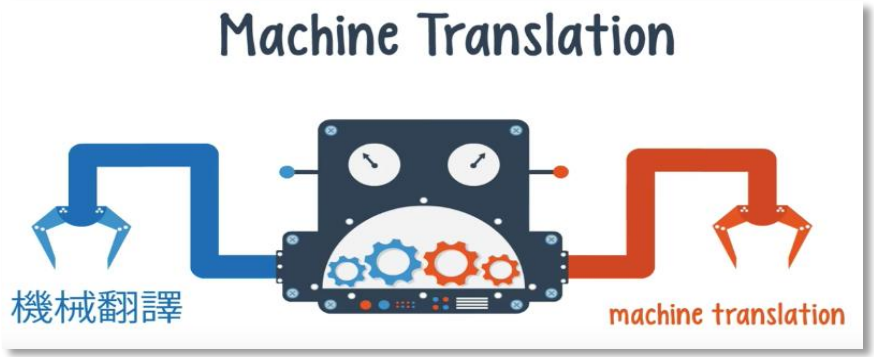
背景简介

- 深度学习的应用



计算机视觉

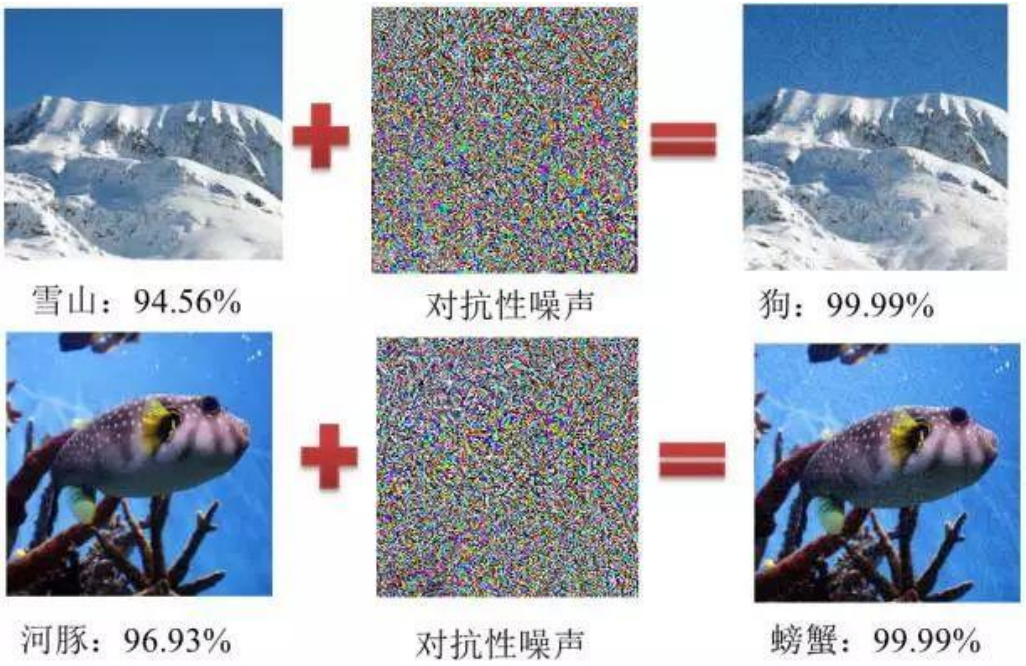
语音识别



自然语言处理

背景简介

- 黑盒问题
  - 无法理解深度模型这种黑盒算法
  - 实践走在理论之前



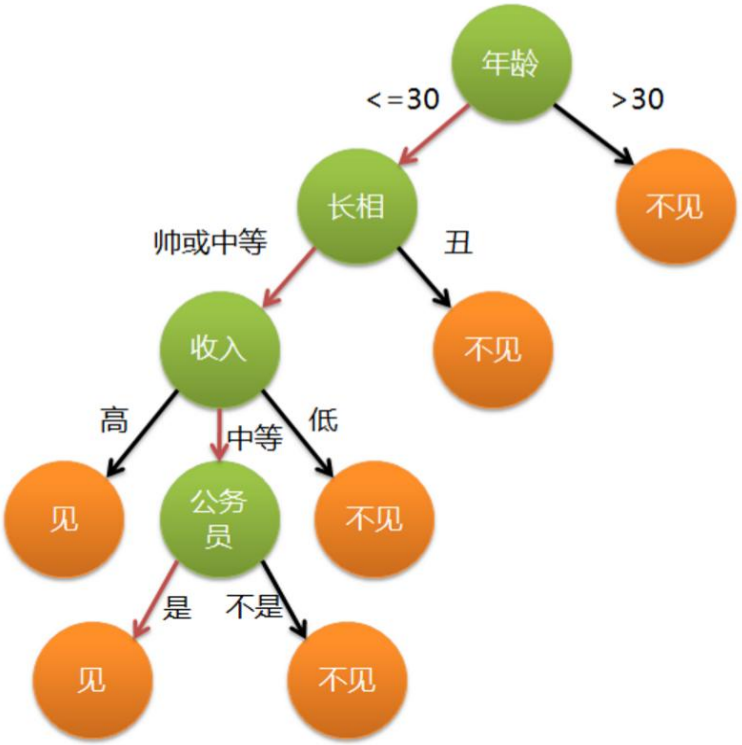
背景简介

- 深度模型可解释性&决策树
  - Frosst N, Hinton G. Distilling a Neural Network Into a Soft Decision Tree[J]. 2017.
  - Wu M, Hughes M C, Parbhoo S, et al. Beyond Sparsity: Tree Regularization of Deep Models for Interpretability[J]. AAAI 2018.

基本知识

- 可解释性
  - 我们应该把可解释性看作人类模仿性（human simulatability）如果人类可以在合适时间内根据输入数据和模型参数，经过每个计算步作出预测，则该模型具备模仿性（Lipton 2016）
  - 以医院生态系统为例：给定一个模仿性模型，医生可以轻松检查模型的每一步是否违背其专业知识，甚至推断数据中的公平性和系统偏差等。这可以帮助从业者利用正向反馈循环改进模型

- 决策树
  - 输入：[年龄，外貌，收入，公务员]
  - 输出：[见，不见]



- 平均路径长度APL（Average Path Length）
  - 公式：$APL = \frac{1}{N}\sum_{n=1}^{N} p(n)$
  - 实例

    数据集：

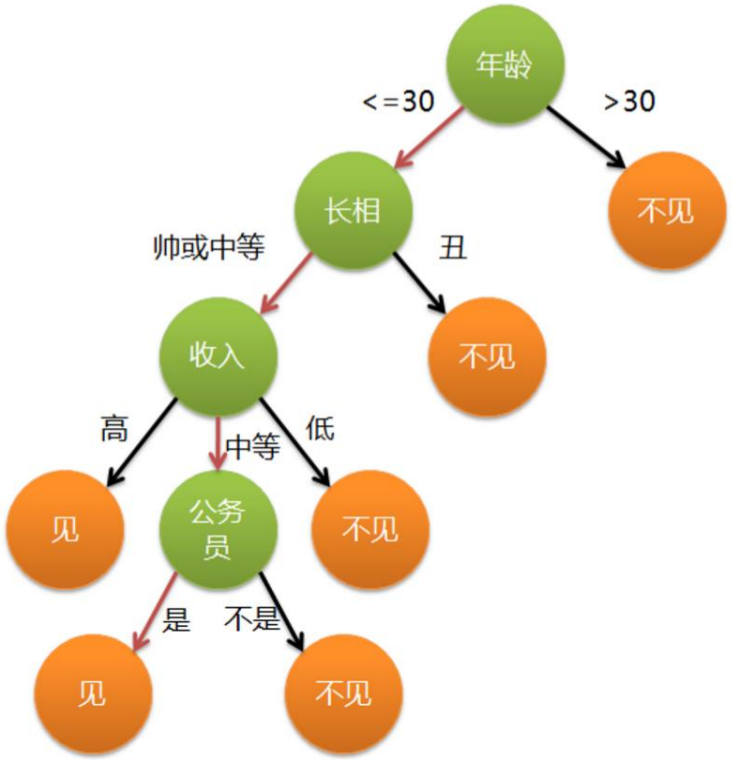
    a1=[20,帅，中，是公务员]

    a2=[35,帅，中，是公务员]

    a3=[25,丑，高，是公务员]

APL=1/3(p(a1)+p(a2)+p(a3))

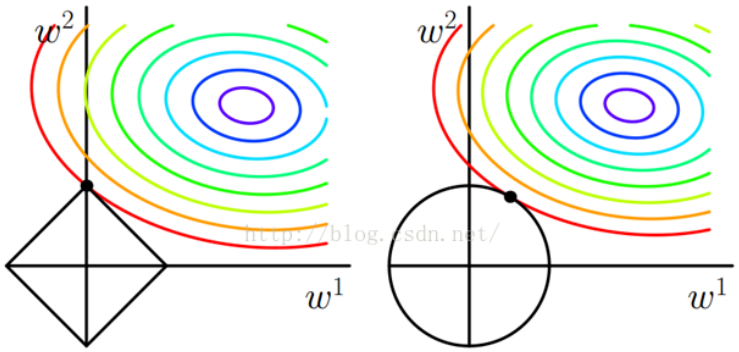
=1/3(4+1+2)=2.3



- 正则化：
  - 机器学习中，许多策略被显式地设计来减少泛化误差，这些策略统称为正则化。
  - 对学习算法的修改-旨在减少泛化误差而不是训练误差。

L1正则化： $J = J_0 + \partial \sum_{n=1}^{N} |w_n|$

L2正则化： $J = J_0 + \partial \left(\sum_{n=1}^{N} w_n^{\ 2} \right)$



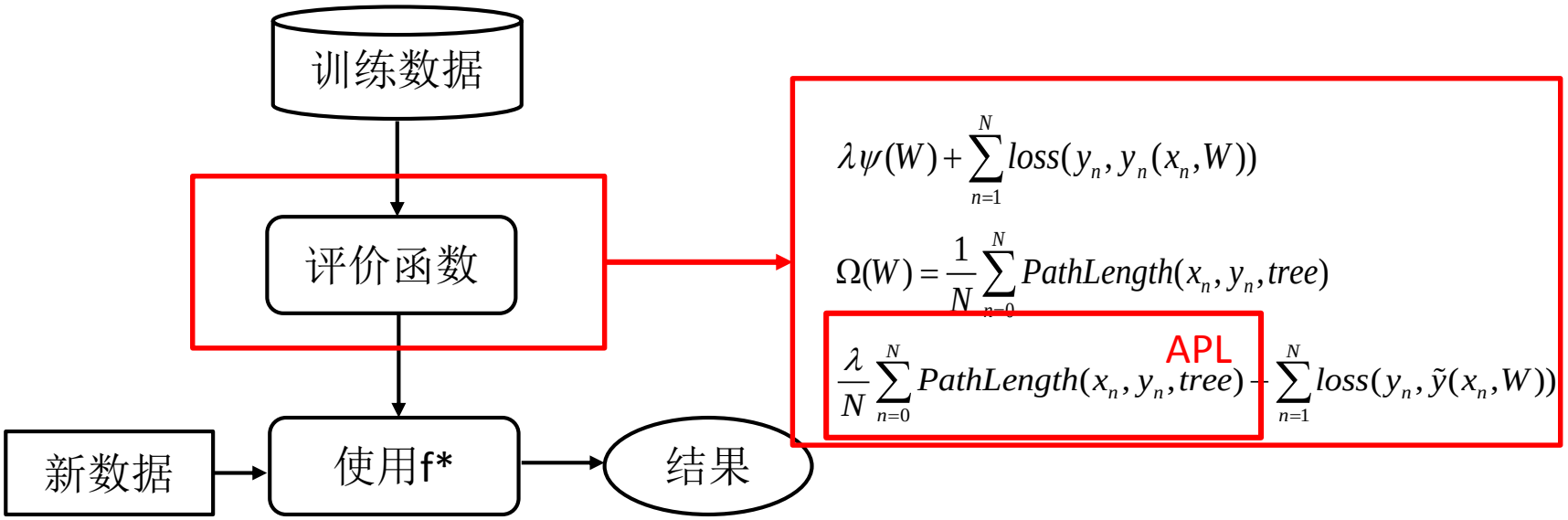
(a) ℓ_1-ball meets quadratic function. ℓ_1-ball has corners. It's very likely that the meet-point is at one of the corners.

(b) ℓ_2-ball meets quadratic function. ℓ_2-ball has no corner. It is very unlikely that the meet-point is on any of axes.

算法原理

算法原理

- 总体思路
  - 在训练深度神经网络中，通过**树正则化**方式训练决策树。让决策树模拟深度神经网络。



算法原理

- APL值计算

1. 将训练数据D输入深度神经网络，得到预测结果A。

2. 由步骤1中可得训练数据集C，根据C训练得到模拟决策树。

3. 将 $D = \{x_n\}_{n=1}^{N}$ 作为步骤二得到的决策树（DT）输入，得到APL值

算法原理

- 为什么惩罚项是APL
  - 和人类模仿性性最相关的参数

    人类的模仿性需要逐步完成预测所需的每个计算。平均路径长度准确地计算了进行平均预测所需的布尔计算的数量

Algorithm 1 Average-Path-Length Cost Function

Require:

$\hat{y}(\cdot, W)$: binary prediction function, with parameters W

$D = \{x_n\}_{n=1}^{N}$: reference dataset with N examples

1: **function** $\Omega(W)$

2: tree $\leftarrow$ TRAINTREE$(\{x_n, \hat{y}(x_n, W)\})$

3: **return** $\frac{1}{N}\sum_n$ PATHLENGTH(tree, x_n)

算法原理

- 问题：APL计算不可微

SGD算法：$\theta^i = \theta^{i-1} - \eta \nabla C(\theta^{i-1})$

目标函数：$C(\lambda, W) = \frac{\lambda}{N} \sum_{n=0}^{N} PathLength(x_n, y_n, tree) + \sum_{n=1}^{N} loss(y_n, \tilde{y}(x_n, W))$

算法原理

- 代理模型
  - 找到一个可微的代理模型，代理原来APL的计算方法。

$$\tilde{\Omega}(W) \approx \frac{1}{N}\sum_{n=0}^{N} PathLength(x_n, y_n, tree)$$

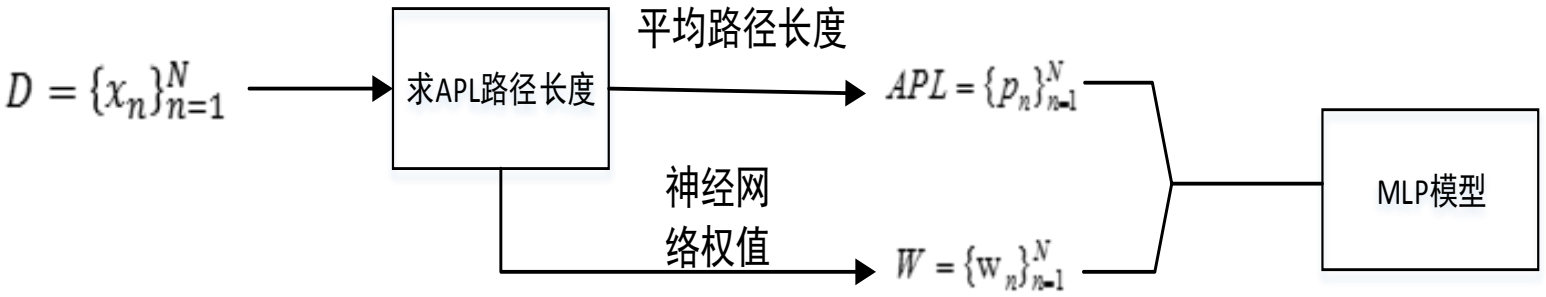
$$\min_{\xi} \sum_{j=1}^{J} (\Omega(W_j) - \tilde{\Omega}(W_j))^2 + \varepsilon \| \xi \|_2^2$$

算法原理

- 代理模型
  - 思路：找到权重W和APL之间的映射关系

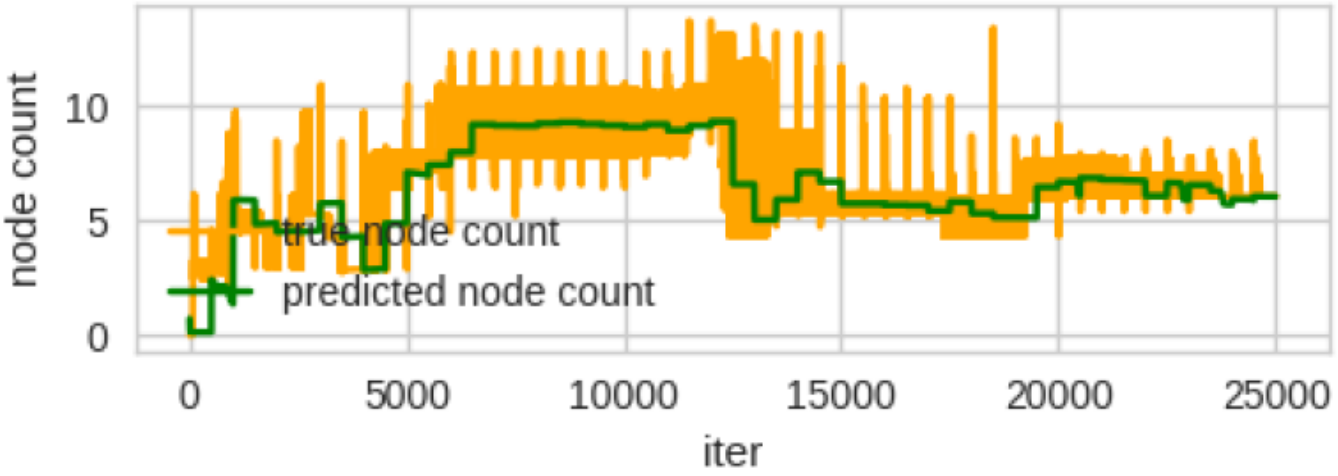
$$\tilde{\Omega}(W) = \text{APL}$$

  - 方法：训练MLP,建立映射关系
  - 具体过程：



算法原理

- 代理模型实验结果



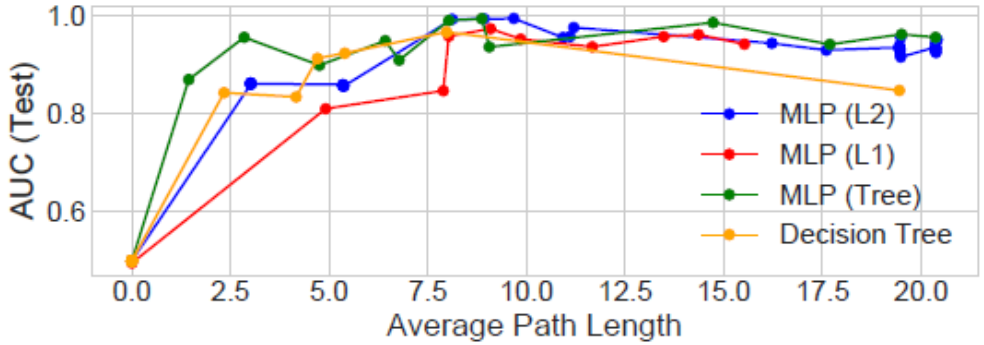
(a) Path length estimates $\hat{\Omega}$ for 2D Parabola task

使用带有25个隐藏节点的单层 MLP效果图

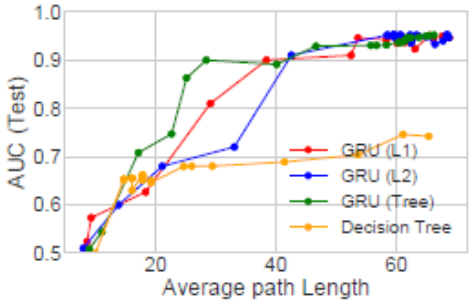
实验分析

实验分析

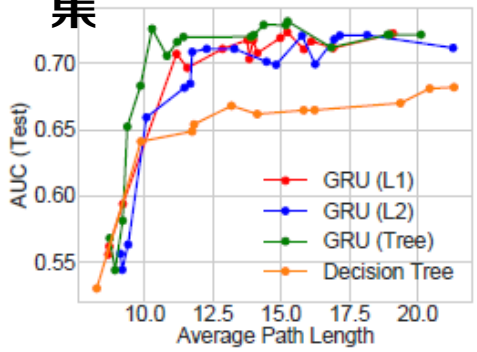
- 准确率
  - 在经典数据集上，深度模型在树正则化后和其他正则化后效果的比较



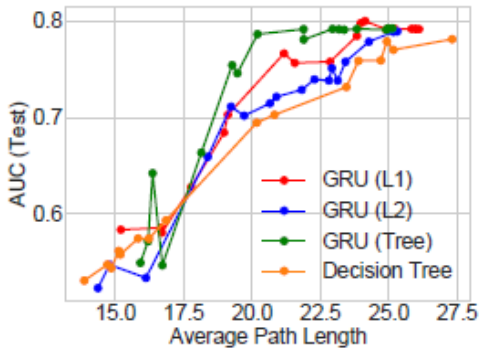
HMM数据集



(a) TIMIT Stop Phonemes



(b) HIV: CD4$^+$ $\leq$ 200 cells/ml



(c) HIV Therapy Adherence

实验分析

- 运行效率
  - 各种数据集上，不同模型效率对比
  - 结论：代理模型的加入，会让训练时间变长

Dataset	Model	Epoch Time (Sec.)
Signal-and-noise HMM	HMM	16.66 ± 2.53
Signal-and-noise HMM	GRU	30.48 ± 1.92
Signal-and-noise HMM	GRU-HMM	50.40 ± 5.56
Signal-and-noise HMM	GRU-TREE	43.83 ± 3.84
Signal-and-noise HMM	GRU-HMM-TREE	73.24 ± 7.86
SEPSIS	HMM	589.80 ± 24.11
SEPSIS	GRU	822.27 ± 11.17
SEPSIS	GRU-HMM	$1\,666.98 \pm 147.00$
SEPSIS	GRU-TREE	$2\,015.15 \pm 388.12$
SEPSIS	GRU-HMM-TREE	$2\,443.66 \pm 351.22$
TIMIT	HMM	$1\,668.96 \pm 126.96$
TIMIT	GRU	$2\,116.83 \pm 438.83$
TIMIT	GRU-HMM	3207.16 ± 651.85
TIMIT	GRU-TREE	$3\,977.01 \pm 812.11$
TIMIT	GRU-HMM-TREE	$4\,601.44 \pm 805.88$

实验分析

- 模拟决策树置信度
  - 模拟决策树和原深度模型在各种数据上集置信度表现
  - 结论：模拟决策树是可信的

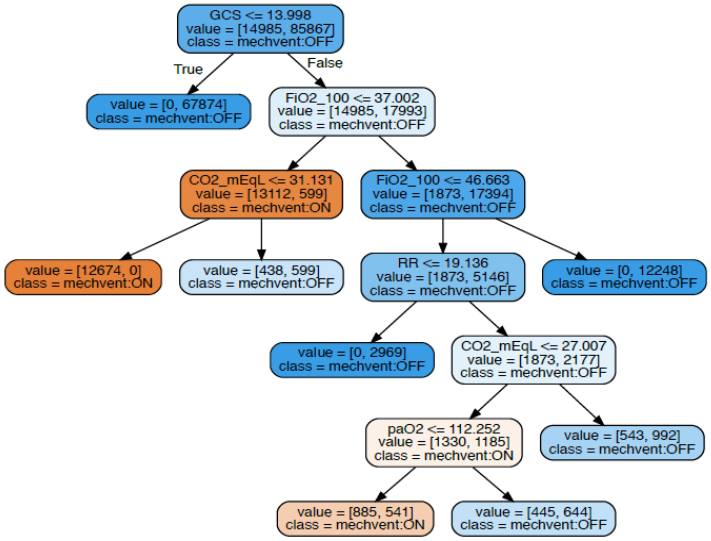
Dataset	Fidelity
signal-and-noise HMM	0.88
SEPSIS (In-Hospital Mortality)	0.81
SEPSIS (90-Day Mortality)	0.88
SEPSIS (Mech. Vent.)	0.90
SEPSIS (Median Vaso.)	0.92
SEPSIS (Max Vaso.)	0.93
HIV ($CD4^+$ below 200)	0.84
HIV (Therapy Success)	0.88
HIV (Mortality)	0.93
HIV (Poor Adherence)	0.90
HIV (AIDS Onset)	0.93
TIMIT	0.85

实验分析

- 可解释性
  - 数据集：Sepsis-超过 1.1 万败血症 ICU 病人的时序数据。
  - 结论

临床医生注意到树节点上的特征（FiO2、RR、CO2 和 paO2）

以及中断点上的值是医学上有效的，可解释性合理。



(d) Mechanical Ventilation

应用总结

应用总结

- 新的正则方法
- 用于特征选择
- 可解释性研究提供思路

参考文献

参考文献

[1] Wu M, Hughes M C, Parbhoo S, et al. Beyond Sparsity: Tree Regularization of Deep Models for Interpretability[J]. AAAI 2018.

[2] https://blog.csdn.net/zouxy09/article/details/24971995机器学习中的范数规则化之L0、L1与L2范数

道德经

大成若缺，其用不弊。
大盈若冲，其用不穷。
大直若屈。大巧若拙。
大辩若讷。静胜躁，寒胜热。清静为天下正。

谢谢！