**第八章 可解释机器学习**

目录

1 基础知识 1

1.1 设计科学 1

1.2 机器学习 1

1.2.1 机器学习的定义 2

1.2.2 机器学习与相关领域的关系 2

2 机器学习方法 3

2.1 线性回归模型 3

2.2 决策树 5

2.3 XGboost 6

2.4 神经网络 8

2.5 支持向量机 9

3 可解释内涵 11

3.1 可解释的意义 11

3.1.1 案例引入 11

3.1.2 机器学习的优劣势 11

3.1.3 可解释机器学习的实际需求 12

3.1.4 可解释性的意义 13

3.2 可解释的分类 13

3.2.1 自解释与事后解释 13

3.2.2 特定于模型还是与模型无关 14

3.2.3 局部还是全局 14

3.3 可解释的范围 14

3.3.1 辨析：算法的透明度与可解释性 14

3.3.2 全局、整体的模型可解释性 14

3.3.3 局部可解释性 15

4 可解释机器学习方法 16

4.1 自解释模型 16

4.1.1 线性回归模型的解释 16

4.1.2 决策树模型的解释 21

4.2 事后解释方法 22

4.2.1 部分依赖图和个体条件期望图 22

4.2.2 置换特征重要性 25

4.2.3 LIME 27

4.2.4 SHAP 30

5 可解释机器学习的未来 37

5.1 重点将放在与模型无关的可解释性工具上 37

5.2 分析模型重于分析数据 37

5.3 机器人和程序会自我解释 37

5.4 可解释性可以促进机器智能研究 37

# 基础知识

## 设计科学

**行为科学**与**设计科学**对应管理领域两个范式，分别是**用户范式**与**系统范式**。

**行为科学**是从人的本性和需求入手，注重人的行为动机和行为的诱导。

**设计科学**的本质在于决策、问题求解和创造。整个设计过程都贯穿着信息的收集、整理、变换、传输、贮存、处理、反馈等基本活动要素。一方面，信息处理被用来解释设计思考过程；另一方面，信息处理技术又被广泛用作设计工具。

**数字经济**时代，计算社会科学成为社会科学领域研究热点。以**数据挖掘和机器学习**为代表的计算方法逐步融入**管理学**研究，形成了传统研究范式与计算科学相互融合的趋势。这一趋势在管理领域表现为**设计科学**与**行为科学**两个主要研究范式相互融合，形成了结合**计算方法创新**与**构念理论创新**多方面特色的研究成果。

## 机器学习

人工智能大师Herb Simon（见人物简介1.2.1）这样定义**学习**：

**学习**：系统在不断重复的工作中对本身能力的增强或改进，使得系统在下一次执行相同任务或类似任务（指的是具有相同分布的任务）时，比现在做的更好或效率更高。

**机器学习**：通过经验提高系统自身性能的过程(系统自我改进)。或者，计算机基于数据做出和改进预测或行为的一套方法。

### 机器学习的定义

**Herbert Alexander Simon**

**赫伯特 亚历山大 西蒙（1916-2001）**

研究领域涉及认知心理学、计算机科学、经济学、管理学、科学哲学等多个方面。他学识渊博，是现今很多学术领域的创始人之一，如人工智能、信息处理

、决策理论、注意力经济、组织行为学、复杂系统等。创造了术语“有限理性”、“满意度（satisficing）”等

**人物简介1.2.1**



1936年 芝加哥大学政治系并获得文学学士学位

1943年 加利福尼亚大学完成管理决策的博士论文，成就经典著作《管理行为》

1949年 卡内基美隆大学担任行政学和心理学教授，后来任计算机科学与心理学教授终生

1956年 西蒙、纽厄尔、麦卡锡、明斯基被公认为人工智能之父

1957年 西蒙与别人合作开发了IPL语言最早的AI程序设计语言。

1958年 荣获美国心理学会杰出贡献奖

1970年 提出决策支持系统DSS。

1978年 诺贝尔经济学奖。

1986年 美国国家科学家奖章。

1993年 美国心理协会终身成就奖。

1970年 发展和完善了语义网络。

1975年 荣获图灵奖。



图 1

如图 1，令W是这个给定系统的有限或无限所有对象的集合，由于观察能力的限制，我们只能获得这个世界的一个有限的子集Q⊂W，称为**样本集**。

**机器学习**就是根据这个有限样本集Q ，推算这个系统的模型，使得其对这个系统为真。

### 机器学习与相关领域的关系

**预测与学习**是**智能**的本质，**预测和学习**也是机器学习的主要功能性组成部分。

很难想象: 一个没有**学习**功能的系统是能被称为是具有**智能**的系统。**机器学习**是**人工智能**的主要核心研究领域之一，如图 2。

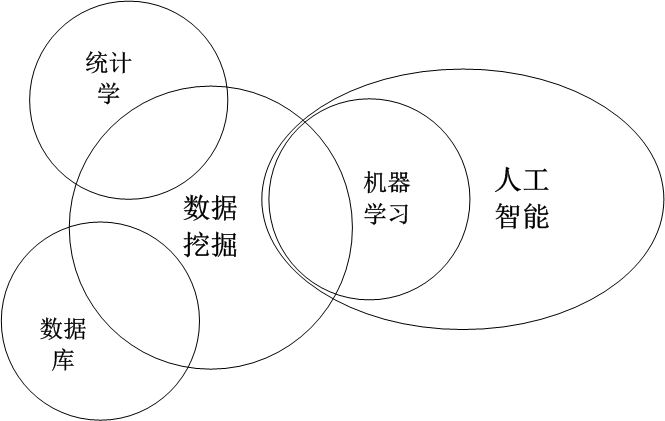


图 2

# 机器学习方法

## 线性回归模型

因为表哥达尔文《物种起源》一书，高尔顿对遗传问题产生了浓厚的兴趣，并积极收集数据证明自己的想法。**回归**(Regression)这一概念最早由英国生物统计学家高尔顿和他的学生皮尔逊在研究父母亲和子女的身高遗传特性时提出。

**线性回归**（Linear Regression）是一种预测性的建模技术，研究的是因变量（目标）和自变量之间的线性关系，**因变量是连续的**。

**目标函数** (Objective function)：

其中：

· 代表训练损失 (Training Loss)，表示模型多好的拟合训练数据。

·为正则化项 (Regularization)，衡量了模型复杂度。

训练数据损失函数 (Training Loss)：

·平方损失 (Square Loss)：

·逻辑损失 (Logistic Loss)：

正则化项 (Regularization)：描述了模型的复杂程度。

·L1 Norm (lasso)：

·L2 Norm：

**公式 2.1.1 线性回归的目标函数**

线性回归模型可以描述为：

假设第i个数据，对应的真实值为 ，模型的输出值为，所以我们的目标是使得 尽可能地接近，假设训练后时，对应的参数为和 ，其中代表一组最优，其中是特征的数量。

因此，线性回归是一个求解参数和的问题，具体有如下解法：

**公式 2.1.2 线性回归的解法**

在线性回归中，误差项为观测值与预测值之间的残差。

在Python中，使用sklearn.linear\_model模块的LinearRegression类来实现线性回归模型，示例代码如下：

from sklearn.linear\_model import LinearRegression

regressor = LinearRegression()

regressor = regressor.fit(X\_train, Y\_train)

Y\_pred = regressor.predict(X\_test)

**代码 2.1.1 线性回归的Python实现**

## 决策树

**决策树**(Decision Tree)是一种非参数的有监督学习，它能够从一系列有特征有标签的数据中总结出**决策规则**，并用树状图的结构来呈现这些规则，以解决分类和回归问题。

熵的定义：

信息增益的最大化：

**公式 2.2.1 决策树的信息增益计算**

在Python中，使用sklearn.linear\_model模块的LinearRegression类来实现线性回归模型，示例代码如下：

主代码部分：

from sklearn import tree

clf = tree.DecisionTreeClassifier(max\_depth=3)

clf = clf.fit(X\_train, y\_train)

from sklearn.metrics import accuracy\_score # acc准确率评估

print(accuracy\_score(y\_test, clf.predict(X\_test)))

可视化部分：

from sklearn.externals.six import StringIO

from IPython.display import Image

from sklearn.tree import export\_graphviz

import pydotplus

dot\_data = StringIO()

export\_graphviz(clf, out\_file=dot\_data, filled=True, rounded=True,

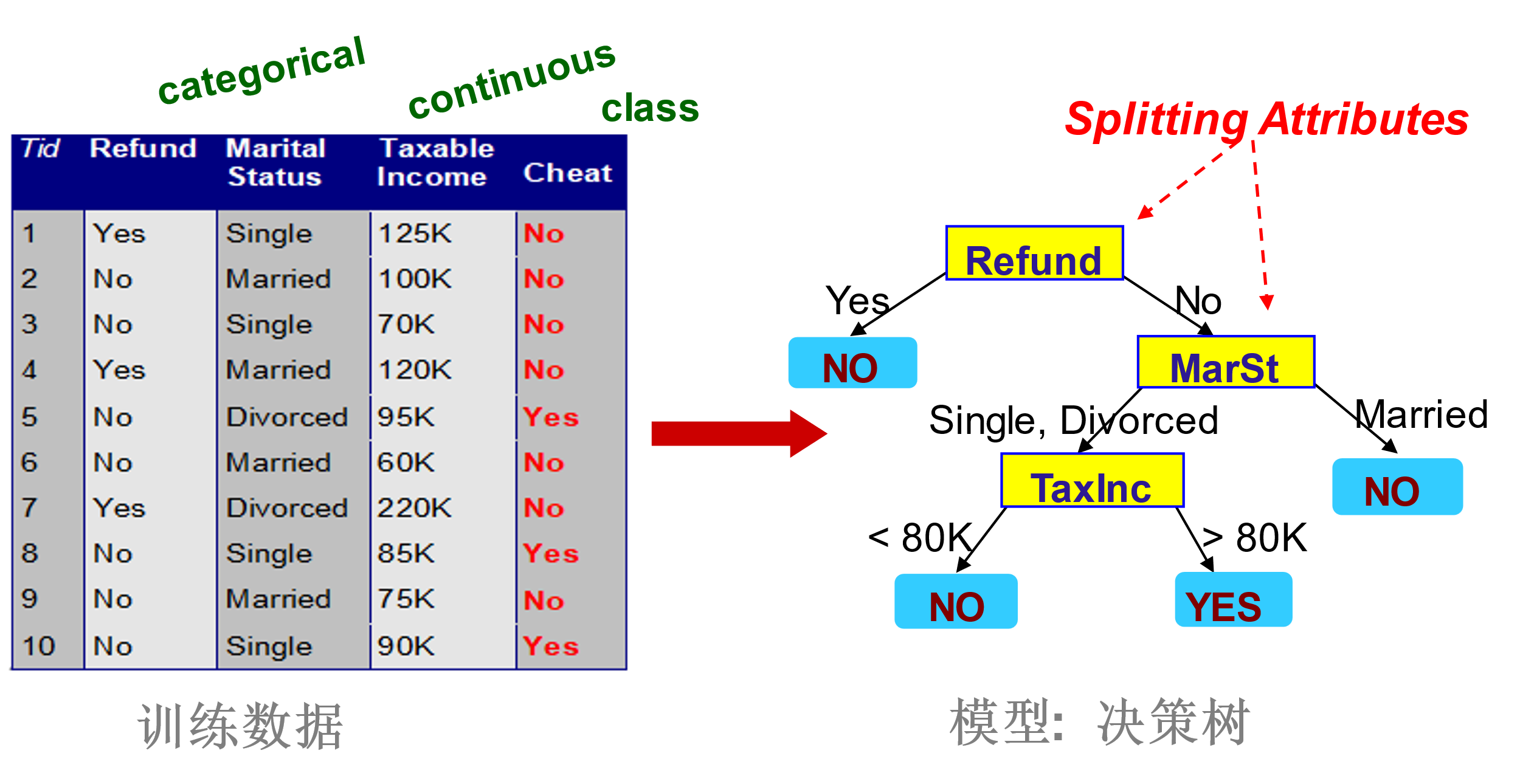
special\_characters=True, precision=2)

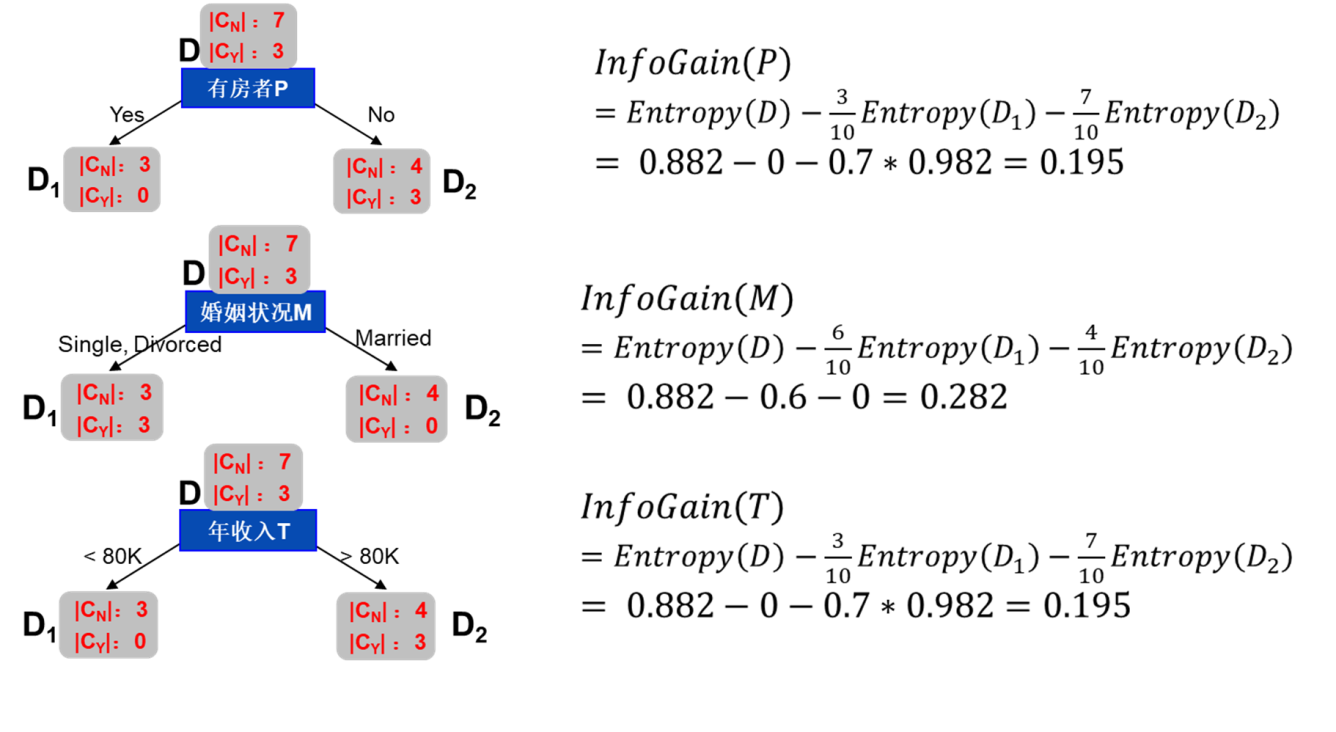
graph = pydotplus.graph\_from\_dot\_data(dot\_data.getvalue())

Image(graph.create\_png())

**代码 2.2.1 决策树的Python实现**

以下的案例2.2.1，是一个使用决策树模型解决实际问题的步骤演示：





**案例 2.2.1 决策树的应用**

## XGboost

**XGboost**的基本组成元素是：决策树；将这些决策树即”弱学习器”共同组成了XGboost。后一棵决策树的生成会考虑前一棵决策树的预测结果，即将前一棵决策树的偏差考虑在内。

**目标函数：**

**展开步骤：**

求导步骤：

优化目标：

**公式 2.3.1 XGBoost模型**

## 神经网络

一个**卷积神经网络模型**一般由若干个卷积层、池化层和全连接层组成。卷积层的作用是提取图像的特征；池化层的作用是对特征进行抽样，可以使用较少训练参数，同时还可以减轻网络模型的过拟合程度。卷积层和池化层一般交替出现在网络中，称一个卷积层加一个池化层为一个特征提取过程，但是并不是每个卷积层后都会跟池化层，大部分网络只有三层池化层。网络的最后一般为1~2层全连接层，全连接层负责把提取的特征图连接起来，最后通过分类器得到最终的分类结果。

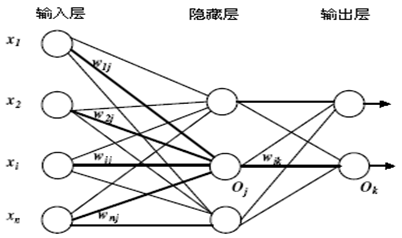


图 3

如图 3，训练样本X = {x1 ,x2 ,..., xi}馈入输入层，每层之间存在加权连接；其中 wij表示由前一层的单元i与其下一层的单元j的之间的连接的权。

在Python中，使用sklearn. tree模块来实现线性回归模型，示例代码如下：

import numpy as np

from keras.models import Sequential

from keras.layers import Dense, LSTM

# 定义 LSTM 模型

model = Sequential()

model.add(LSTM(64, input\_shape=(time\_steps, 1), return\_sequences=True))

model.add(LSTM(64))

model.add(Dense(1, activation='linear'))

# 编译模型

model.compile(loss='mse', optimizer='adam')

# 训练模型

model.fit(train\_input, train\_target, epochs=100, batch\_size=32, validation\_split=0.2)

# 进行预测

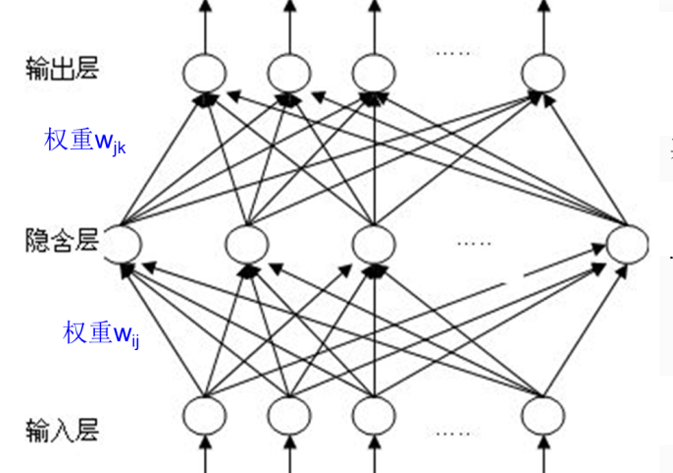
predictions = model.predict(test\_input)

**代码 2.4.1 神经网络的Python实现**

**输出层：**

**隐含层：**

**公式 2.4.1 神经网络**



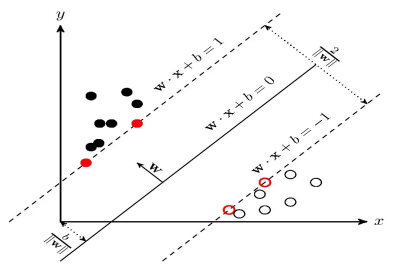
**输出结点的误差：**

**输出层权重求偏导：**

**隐含层权重求偏导：**

## 支持向量机

**SVM**（Support Vector Machine）的最核心的思想就是从输入空间（Ipnut 从输入向某一个中间的阶段做了一个映射，再进行分类。



分类函数:

sign函数定义：

优化问题的原始形式:

拉格朗日函数：

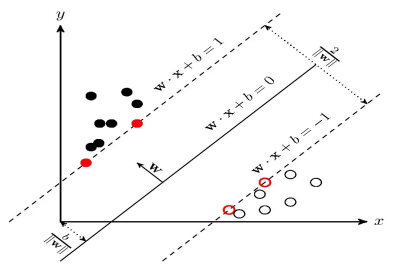
拉格朗日对偶问题:

其中 是拉格朗日乘子，对偶问题的解决涉及对 和 的最优化以及对 的最大化。  
导数和条件:

对偶问题的最优化问题:

核函数示例:  
- 线性核:   
- 多项式核:   
- 高斯核:   
- Sigmoid核:

**公式 2.5.1 支持向量机**



# 可解释内涵

## 可解释的意义

### 案例引入

深度学习（DL，Deep Learning）是机器学习（ML，Machine Learning）领域中一个新的研究方向，它被引入机器学习使其更接近于最初的目标——人工智能。深度学习源于人工神经网络的研究，含多个隐藏层的多层感知器就是一种深度学习结构。动机在于建立模拟人脑进行分析学习的神经网络，例如图像，声音和文本等。

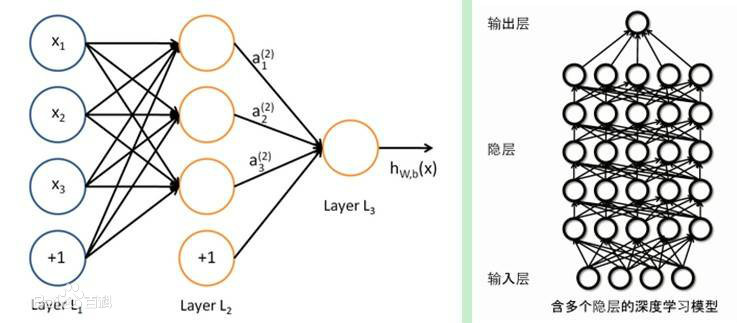


图 4

相⽐于传统的模式识别、机器学习需要⼿⼯提取特征，深度学习最⼤的优势是可以⾃动从数据中提取特征。通过多层处理，逐渐将初始的“低层”特征表示转化为“高层”特征表示，可将深度学习理解为进行“特征学习”（feature learning）或“表示学习”（representation learning）。

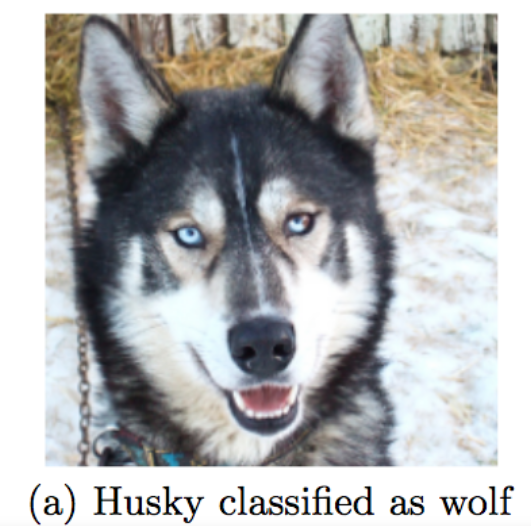
使⽤机器学习系统识别⼀只狼，我们需要⼿动提取、输⼊我们⼈类可以将动物识别为狼的特征，⽐如体表⽑茸茸、有两只凶狠的眼睛等；⽽使⽤深度学习系统，我们只需要将⼤量的狼的图⽚输⼊系统，它就会⾃动学习特征。

以识别狼的⼈⼯智能系统为例，在Macro等⼈研究的⼯作中发现，深度学习系统会将西伯利亚哈⼠奇识别为狼，为什么呢？

原因是在训练系统时，输⼊的狼的图⽚上的背景⼤多是皑皑⽩雪，使得系统在⾃动提取特征时将⽩雪背景作为了识别狼的标志。

图b为说明系统使⽤了图⽚的背景并忽略了动物的特征。模型原本应该关注动物的眼睛

**案例3.1.1 机器学习识别狼**



除此之外，还有许多例子，说明了对“黑盒”模型进行解释的重要性。

比如，在医疗领域，⽤了⼀个极深的神经⽹络，预测结果告诉⽤户很可能得了癌症。那么，为什么呢？为什么要相信你呢？你是不是应该给出⼀个合理的解释呢？你该如何给出⼀个合理的解释呢？

不光要提升预测的精度，还要减小发生“信任危机”的可能，只有明白模型这样预测的原因，才能信任该模型，因此，对“黑盒”的解释变得尤为重要。

### 机器学习的优劣势

机器学习、深度学习（预测）在医学、刑事司法系统和金融市场等关键领域的应用越来越广泛，基于GBDT的各种优化以及ensemble横扫所有数据竞赛，CNN则是霸榜图像领域。

在优势方面：机器学习算法（基于神经网络）通常遵循⾮线性，学习能⼒强，覆盖范围⼴，数据驱动，⽤测试集的准确率与召回率衡量模型的好坏，通常会具有良好的准确率、预测性能。

2016年5⽉，美国佛罗⾥达州⼀位男⼦驾驶开启⾃动驾驶模式后，特斯拉Model S撞上⼀辆正在马路中间⾏驶的半挂卡车，导致驾驶员当场死亡。什么原因造成的呢？

排查原因时发现，主要是因为图像识别系统没能把卡车的⽩⾊车厢与⼤背景中的蓝天⽩云区分开导致的，系统认为是卡车是蓝天⽩云时就不会⾃动刹车进⽽引发了事故。

**案例3.1.2 图像识别导致的事故**

在劣势方面：机器学习具有不可解释性，⿊盒模型（⽐如神经⽹络）计算过程中会⾃动从原始数据中提取特征并拆分、组合，构成其判别依据，却可能**⽆法理解其提取的特征**。

同时，虽然机器学习有着更⾼的准确率，但是模型最后输出结果时，它是根据哪些⽅⾯、哪些特征得到这个结果，**⽆法给出具体的规则**，**⽆法让⼈理解和信任**模型的预测结果。

另外，⼀个预测表现接近完美、却属于⿊盒的机器学习模型，可能**产⽣误导的决策**，还可能**招致系统性风险**，导致漏洞被攻击，因⽽变得不安全可靠。

### 可解释机器学习的实际需求

在不同领域，都有对机器学习模型进行解释的需求，比如：

* 在和客户的实际沟通时，单单抛出⼀个数字就想得到客户信任，那肯定是不够的，这就要求我们摆出规则，业务⼈员才知道如何营销。
* 当模型应⽤到银⾏业等⾦融领域时，透明度和可解释性是机器学习模型是否值得信任的重要考核标准。
* 对风控（安全）⼈员来讲，关注的是如何识别风险（可解释），如何进⾏针对性防御。

它们的共同需求，都是要建⽴⼀个解释器来解释模型。

在学术上，“解释机器学习模型”的任务定义似乎不够明确。学术文献中提出了为模型寻求可解释性的许多动机，并提供了无数的技术来提供可解释的模型。尽管存在这种模棱两可的情况，但许多作者宣称他们的模型在公理上是可解释的，然而对此却缺乏进一步的论证[[1]](#footnote-1)。

### 可解释性的意义

解释器的⼯作在于提供**对样本的解释**，来帮助⼈们对模型**产⽣信任**。

可解释性要求解释器的**特征与模型**须是可解释的。像决策树、线性模型都是很适合拿来**解释的模型**；⽽可解释的模型必须搭配**可解释的特征**，才是真正的可解释性，像是词嵌⼊[[2]](#footnote-2)（Word Embedding）的⽅式，即使是简单的线性回归也⽆法做到可解释性。

解释在于对**预测结果**的解释，哪些特征得到这个结果，结果产生的原因。

## 可解释的分类

### 自解释与事后解释

**自解释模型（Intrinsically Interpretable Model）**指将可解释性结合到具体的模型结构中使模型本⾝具备可解释能⼒。

* 常见的具备可解释性的机器学习模型有线性回归、决策树、逻辑回归、广义加性模型等。
* 背后的哲学意义通常是做出某些分布假设或限制模型的复杂性，使其具有可学习的结构和参数的模型，可以为其分配特定的解释。

**事后可解释性（Post-hoc Interpretable Method）**指通过开发可解释性技术解释已训练好的机器学习模型。

* 发⽣在模型训练之后，旨在利⽤解释⽅法或构建解释模型，解释学习模型的⼯作机制、决策⾏为和决策依据。
* Post-hoc 可解释性的重点在于设计⾼保真的解释⽅法或构建⾼精度的解释模型。

### 特定于模型还是与模型无关

**特定于模型（Model-specific）的解释方法**仅限于特定的模型类，例如线性模型中回归权重的解释。自解释模型的解释通常是特定于模型的解释。

**与模型无关（Model-agnostic）的解释方法**可以用于任何机器学习模型，在模型经过训练后应用（事后），通常通过分析特征输入和输出来实现，一般不能访问模型的内部信息，例如权重或结构信息。换句话说，⽆论多复杂的模型，像是SVM或神经⽹络，该解释器都可以⼯作。

### 局部还是全局

**全局可解释性**（Global Interpretability）旨在帮助⼈们理解复杂模型背后的整体逻辑以及内部的⼯作机制。

**局部可解释性**（Local Interpretability）旨在帮助⼈们理解机器学习模型针对每⼀个输⼊样本的决策过程和决策依据。

**局部保真度:**不可能期望简单的可解释模型在效果上等同于复杂模型（⽐如原始CNN分类器）。所以解释器不需要在全局上达到复杂模型的效果，但⾄少在局部上效果要很接近，⽽此处的局部代表我们想观察的那个样本的周围。

## 可解释的范围

### 辨析：算法的透明度与可解释性

**算法透明度**是指如何从数据中学习模型，以及它可以学习到什么样的关系。它是对算法如何工作的理解，但这既不是对最终学习的特定模型的理解，又不是对如何做出单个预测的理解。算法的透明度只需要对算法了解，而不需要对数据或学习模型有了解。

本研究的目标是预测模型的可解释性，而不是算法的透明度。

### 全局、整体的模型可解释性

要解释全局模型的输出，需要训练好的模型、算法和数据。这种级别的可解释性是基于对模型特征和每个学习部分（例如，权重、其他参数和结构）的整体认知来理解模型是如何做出决策的。

哪些特征很重要，以及它们之间有什么样的交互作用？全局模型的可解释性有助于基于特征理解目标结果的分布。

线性模型，**可解释部分是权重**，单个权重的解释与其他权重是相互关联的，即“其他输入特征的值保持不变”。

例如，一个预测房屋价格的线性模型，考虑到房屋面积大小和房间数量，其中房间数量的特征可能具有负权重。原因是房间数量与房屋面积高度相关。负权重的解释是如果两件房屋的面积相同，那么房间少的房屋比房间多的房屋更值钱。因此，权重仅在模型中其他特征的上下文中有意义。

**案例3.3.2 模块层面上的全局模型可解释性**

### 局部可解释性

**（1）单个预测的局部可解释性**

它着眼某一个实例，检查模型对这个实例输入的预测，并解释原因。在局部，预测可能只依赖于线性的或单调的某些特征，而不是对它们有复杂的依赖性。

例如，全局而言，房价预测可能与面积是非线性关系，但如果查看一个100平方米的房屋，对于该子集，模型预测可能与面积呈现线性关系。可以通过模拟当增加或减少10平方米的面积时，预测的价格是如何变化来发现。局部解释比全局解释更准确。

**（2）一组预测的局部可解释性**

对一组实例进行特定的预测，即多个实例的模型预测的解释可以用全局模型解释方法或单个实例解释方法来实现。

* 全局方法可以通过获取实例组，将其视为一个完整的数据集，使用全局方法处理。
* 单个实例方法是对实例组中的每个实例分布进行局部解释方法，然后为整个组列出结果或对结果进行聚合。

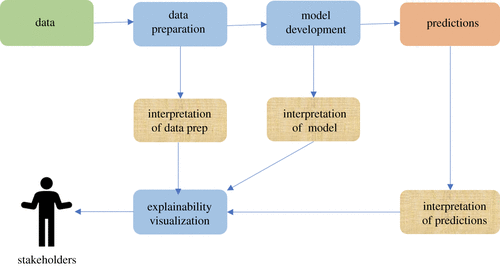


图 5[[3]](#footnote-3)

# 可解释机器学习方法

## 自解释模型

### 线性回归模型的解释

**（1）如何解释？**

* 数值（连续）特征的解释：当所有**其他特征值保持不变**时，特征增加一个单位，则预测结果增加
* 分类特征的解释：当所有其他特征值保持不变时，将特征从参照类别改变为该类别时，预测结果会增加。对于具有多分类的特征，使用独热编码（如图 6），即L个类别用L-1个特征。

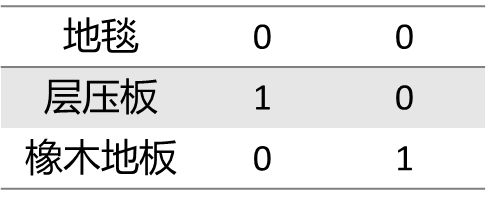


图 6

下面以自行车租赁数据集的线性回归模型解释为例，进行说明：

**实例说明4.1.1.1 自行车租赁数据集**

* season – 季节，四季
* holiday - 是否这天处于假期
* workingday – 是否工作日
* weather –天气情况
* temp-摄氏度温度
* humidity-相对湿度
* windspeed-风速
* days\_since\_2011：自数据集中第一天（20110101）起的天数
* count-总租赁量

该数据集由Capital-Bikeshare公开提供，可从机器学习数据库[[4]](#footnote-4),[[5]](#footnote-5)中下载。使用租借期之前的可用信息来预测租用的自行车总数。

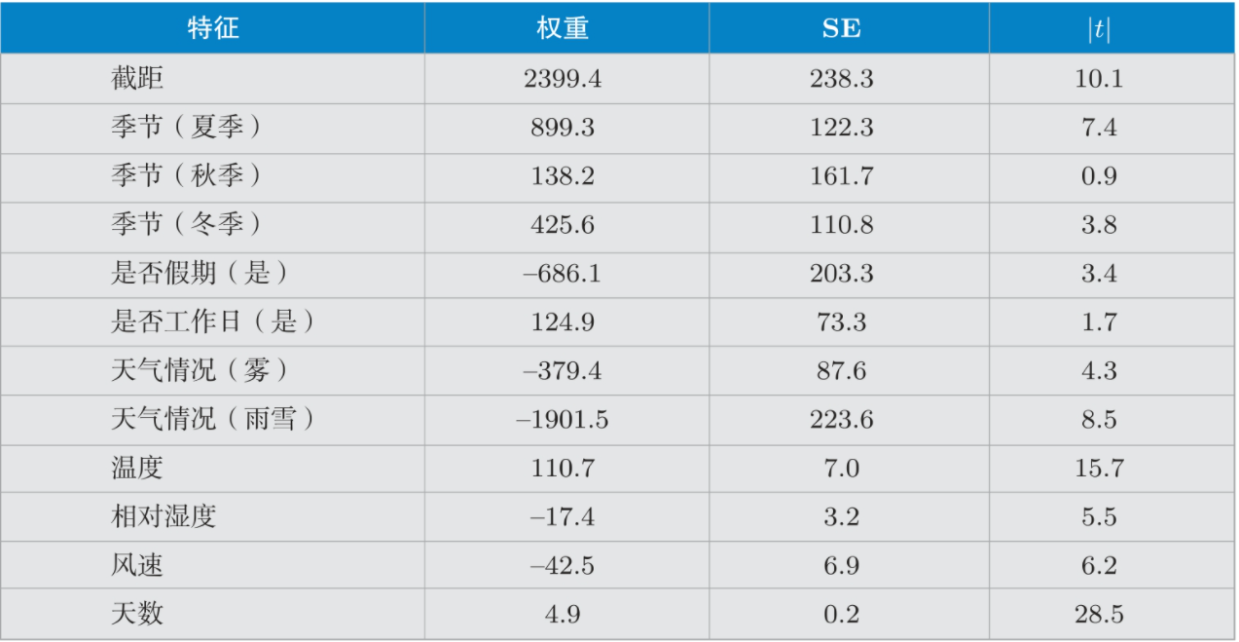


图 7

如图 7，在解释结果中：

* 对于数值（连续）特征的解释：如温度特征，当所有其他特征值保持不变时，将温度升高1度，自行车的预测数量增加110.7。
* 分类特征的解释： 如天气情况，当所有其他特征值保持不变时，当为雨雪天气时，自行车的预测数量与晴天相比减少1901.5。当所有其他特征不变时，当天气为雾时，自行车的预测数量与晴天相比少了379.4。

**（2）如何度量可解释性？**

* R-平方（R-squared, ）

其中，SSE是误差项（真实结果与预测结果）的平方和，显示拟合模型后还有多少方差。SST是数据方差的平方和，目标结果的总方差。

R-平方值越高，模型对数据的解释就越好。对于无法解释数据的模型，R-平方的值为0 ；对于能够解释数据的最优模型，R-平方的值为1。一般认为，R-平方大于0.75表明模型的可解释性较高；R-平方小于0.5表明模型的可解释性存在问题。

* 调整后的R-平方

R-平方随着模型中特征（自变量）数量的增加而增加。调整后的R-平方公式如下：

其中，p为特征（自变量）的数量，n为实例的数量。

解释调整后R-平方很低的模型是没有意义的，对权重的任何解释都没有意义。

一般认为，0.5为调整后R-平方的临界点；而且调整后的R-平方与R-平方之间的差距越大说明模型的可解释性越差。

* 特征重要性

某个特征的重要性可以用它的t-统计量（t-statistic）的绝对值来衡量。 t-统计量以标准差为尺度估计权重。

其中，SE为标准差（Standard Error）。

特征（自变量）的重要性随着权重的增加而增加。估计权重的方差越大，表明对正确值的把握越小，特征越不重要。一般认为，t值的绝对值大于1.96时，认为这个特征（自变量）是重要的。

**（3）如何对解释结果进行可视化？**

* 特征重要性图，如图 8：

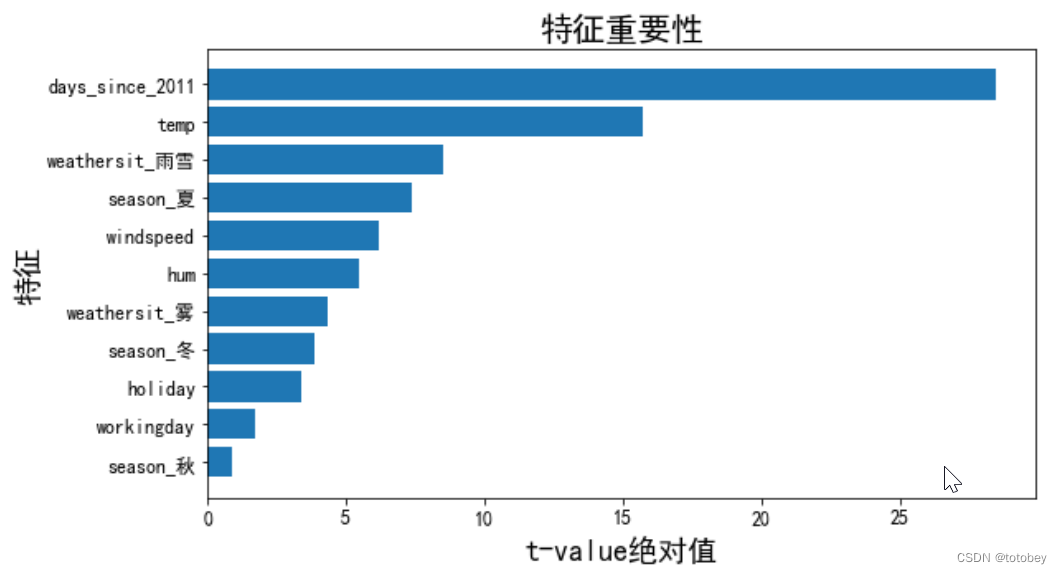


图 8

* 权重图

在该图中，展示了线性回归模型的结果，同时权重用点显示，95%置信区间[[6]](#footnote-6)显示为线。如图 9：

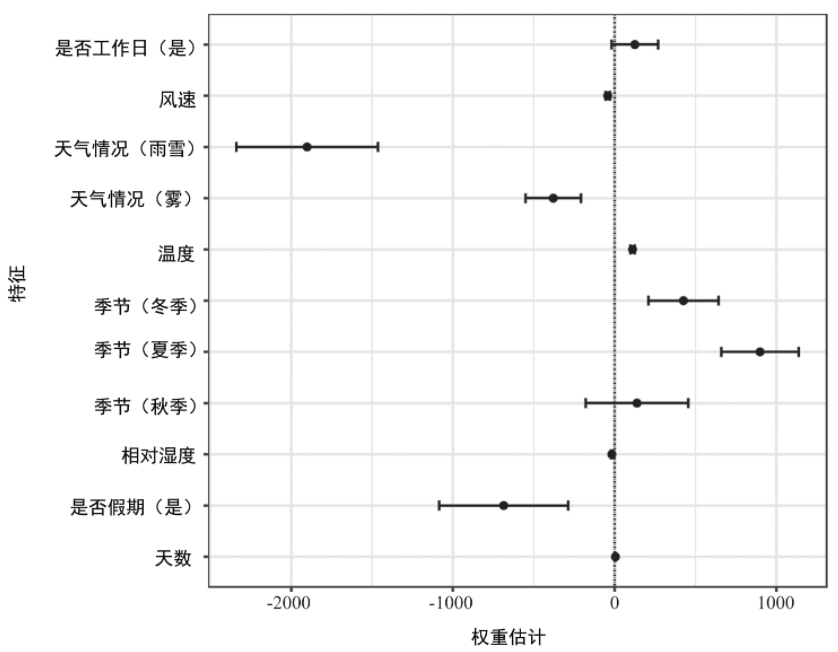


图 9

由图可以说明：一、雨雪天气对预测的自行车数量有很多大负效应。二、是否工作日（是）特征的权重接近于零，并且95%的区间中包含零。

权重图的问题是在不同尺度上的测量。天气情况权重反应了晴天与雨雪天气之间的差异。温度只反应了1度的增加情况。拟合线性模型之前，通过收缩特征可使得估计的权重更具有可比性。

* 效应图

权重取决于特征的单位，如身高特征，如果测量单位从米转换为厘米，权重会有所改变，但实际效应不会改变。而效应图帮助了解**权重与特征的组合**对数据预测的贡献程度，其计算公式如下：



图 10

如图 10，**箱线图**可视化效应，包含一般数据的特征效应范围（1/4分位数Q1到3/4分位数Q3），垂直线为中位线Q2，延伸线到Q1-1.5IQR与Q3+1.5IQR，IQR是四分位之间的距离，即Q3-Q1。其中，箱子的高度在一定程度上反映了数据的波动程度。

* 单个特征效应图

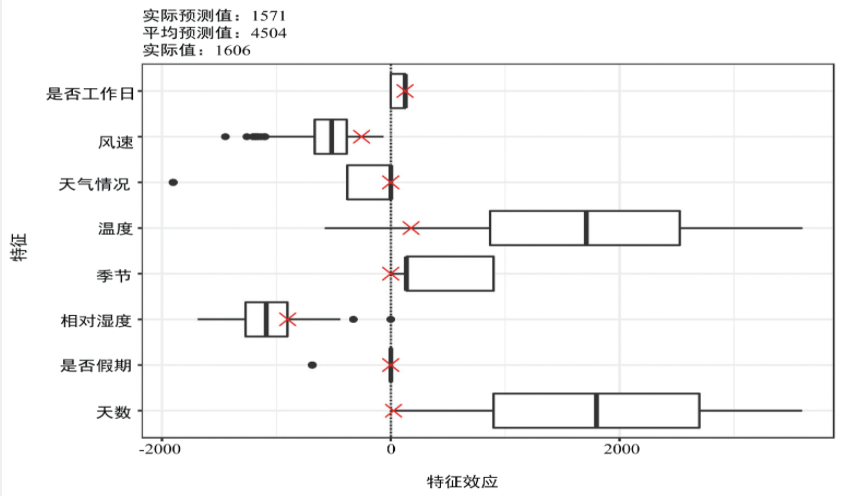


图 11



图 12

如图 11、图 12：

对训练数据预测求平均为4504，该单个实例的预测值很小1571。

温度特征效应较小，因为温度为1.604，较低。

天数特征效应也较小，仅过了5天。

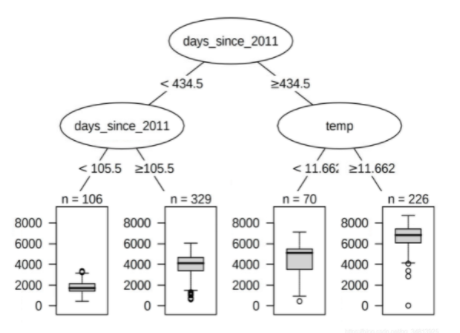
### 决策树模型的解释

[**Boston house prices dataset**](机器学习——线性回归模型的可解释性.pdf)

* CRIM per capita crime rate by town
  + - * + 地区人均犯罪率
* ZN proportion of residential land zoned for lots over 25,000 sq.ft.
  + - * + >25000英尺地块住宅用地比例
* INDUS proportion of non-retail business acres per town
  + - * + 非零售商业用地比例
* CHAS Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
  + - * + 查尔斯河虚拟变量（地区边界是河，值取1，否则为0）
* NOX nitric oxides concentration (parts per 10 million)
  + - * + 一氧化氮浓度（每一千万）
* RM average number of rooms per dwelling
  + - * + 每套住宅平均房间数
* AGE proportion of owner-occupied units built prior to 1940
  + - * + 1940年前建成的自用房的比例
* DIS weighted distances to five Boston employment centres
  + - * + 与五个波士顿就业中心区的加权距离
* RAD index of accessibility to radial highways
  + - * + 放射状公路的可达性指数
* TAX full-value property-tax rate per $10,000
  + - * + 财产税率
* PTRATIO pupil-teacher ratio by town
  + - * + 生师比
* B 1000(Bk - 0.63)2 where Bk is the proportion of blacks by town
  + - * + Bk为黑人比例
* LSTAT % lower status of the population
  + - * + 低地位人口比例
* MEDV Median value of owner-occupied homes in $1000's
  + - * + 自住房平均房价，以千美元计

**实例说明4.1.1.2 Boston房产数据集**

**（1）如何解释？**

从根节点开始，根据边的判断，转到下一个子集，直到走到叶节点，得到结果。所有的边都由AND连接。

解释：

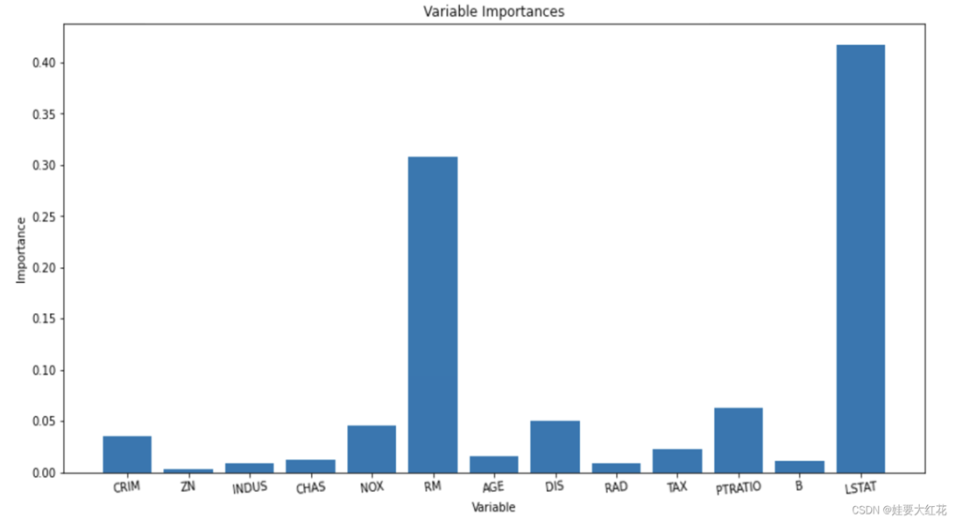
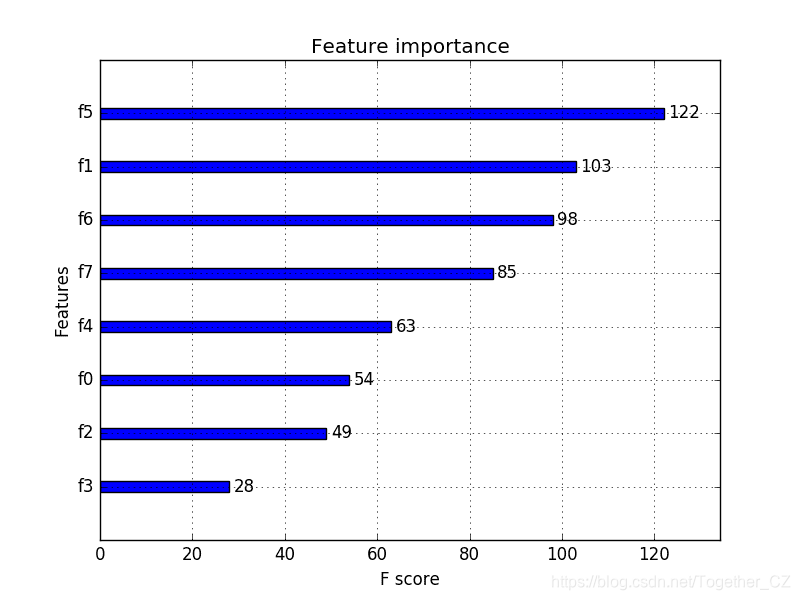
在第105天之前的天数中，自行车的预测数量大约是1800；在第106到第434天之间，大约3900。对于第434天之后的天数，如果温度低于11.662，则预测值为4600；如果温度高于11.662，则预测值是6600。

**案例4.1.2.1 决策树模型实例的解释**

**（2）解释的方法？**

**特征重要性**（Feature Importance） 也被称为变量重要性（Variable Importance），依据某个特征进⾏决策树分裂时，分裂前后的信息增益（如基尼系数）衡量。通过对模型的特征重要性进行排序，就可获得对模型的全局解释，如图 13：

图 13



特征重要性有如下的作用 ：

* 快速的让你知道哪些因素是比较重要的，但是不能得到这个因素对模型结果的正负向影响，同时传统方法对交互效应的考量会有些欠缺。
* 一个特征的意义在于降低预测目标的不确定性，能够更多的降低这种不确定性的特征就更重要。即特征重要性计算依据某个特征进行决策树分裂时，分裂前后的信息增益（基尼系数）。

## 事后解释方法

在事后解释方法中，部分依赖图、个体条件期望图、置换特征重要性、LIME、SHAP成为了事后揭开机器学习模型⿊箱的有⼒⼯具：

* + - * 部分依赖图和个体条件期望图：考察某项特征的不同取值对模型输出值的影响。
      * 置换特征重要性：衡量了对特征值进行打乱后,模型的预测性能的变化。
      * LIME：对于每条样本，寻找⼀个更容易解释的代理模型解释原模型。
      * SHAP：源于博弈论，计算特征对模型输出的边际贡献。

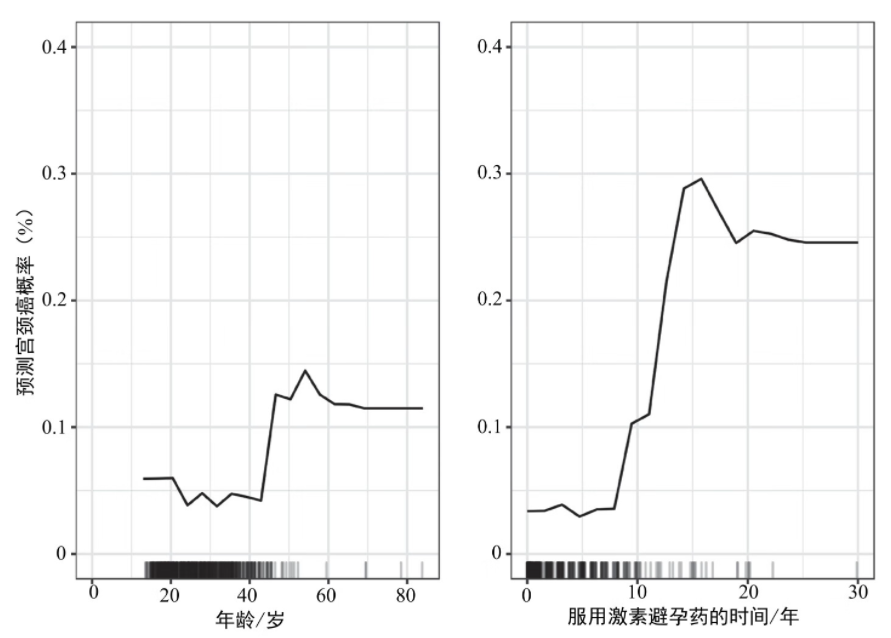
### 部分依赖图和个体条件期望图

**部分依赖图**（Partial Dependence Plot，PDP图或PD图）模型：

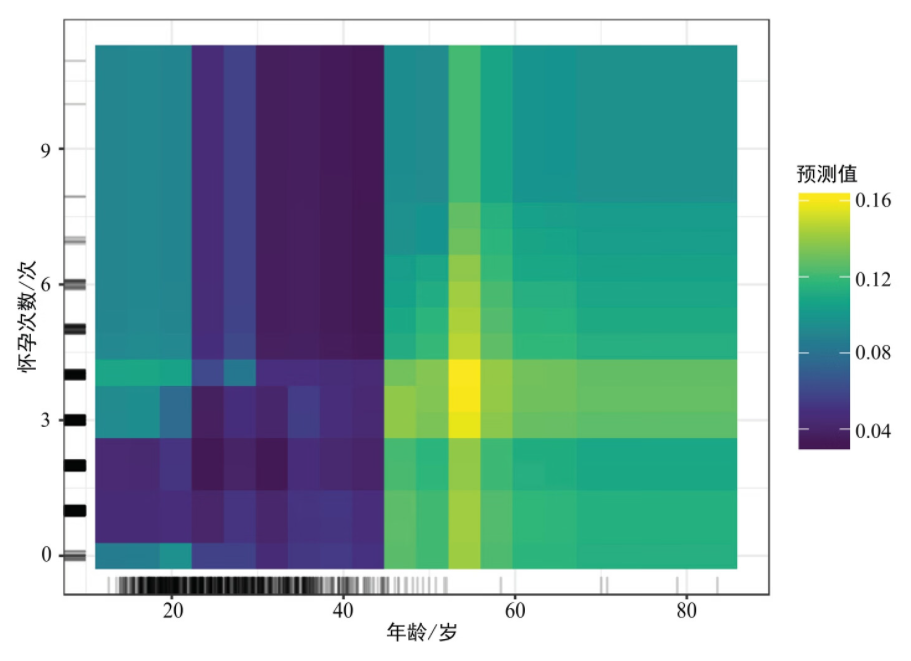
是关心的1到2个特征，而则是其他剩余输入模型的特征，是模型。通过不断改变的取值，来看模型预测结果的变化。

具体操作就是：先固定住其他特征，把你感兴趣的这个特征全部批量的改成某个值，不断**从某个低值改到某个高值**，看所有样本预测结果会做什么变化。

**案例4.2.1.2 部分依赖图解释实例2**



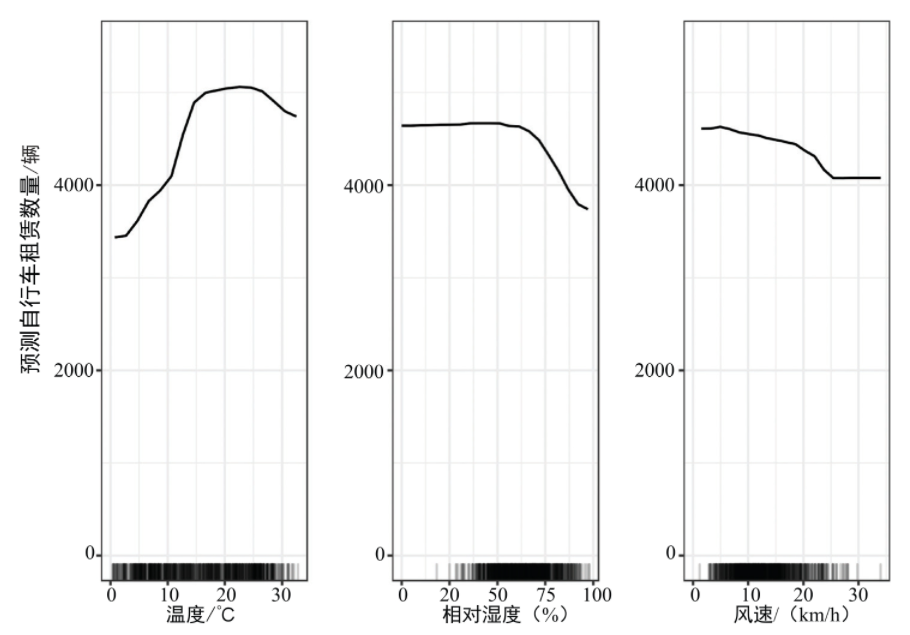
**宫颈癌分类**



* **两个特征的部分依赖关系：**

**案例4.2.1.1 部分依赖图解释实例1**

* 温度越高，自行车租赁数量越多，这种趋势上升到20度，在30度处变平略有下降。
* 当相对湿度超过60%时，潜在的骑自行车的人会越来越不愿意租赁。
* 风越大，骑自行车的人就越少。



**(1) 连续（数值）特征对预测自行车租赁数量的影响**



* 所有季节对模型预测产生相似的影响，模型预测的自行车租赁数量仅在春季会减少。

**(2) 分类特征季节特征对预测自行车租赁数量的影响**

部分依赖图具有如下特点：

* 部分依赖图是一种“全局解释模型的方法”。
* 实际分析PDP时的最大特征数为2个。
* PDP的假设前提是需要计算特征与其他特征之间不相关。如果不能满足这个假设的话，解释的结果可能会存在无法理解的地方。
* PDP最终的返回结果是平均值，多种类的影响可能会被隐藏。

**个体条件期望图**（Individual Conditional Expectation, ICE）计算方法与 PDP 类似:

PDP是一种全局方法，关注的是整体平均值；ICE刻画的是每个个体的预测值与单一变量之间的关系，关注的是特定实例。

* 原理和实现方法：对某一个体，保持其他变量不变，置换选定的特征变量的取值，放入黑箱模型输出预测结果，最后绘制出针对这个个体的单一特征变量与预测值之间的关系图。
* ICE 图可视化预测对每个实例的特征的依赖性，导致每个实例一条线；而PDP是一条线整体， 是 ICE 图线的平均值。

**以宫颈癌数据集为例**，看每个实例的预测如何与特征“年龄”相关联：

ICE 图显示，对于大多数女性而言，年龄效应遵循 50 岁时增加的平均模式；例外：对于少数在年轻时具有高预测概率的女性。

**案例4.2.1.3 个体条件期望图解释实例1**

**累积局部效应图ALE**（Accumulated Local Effects）解决了特征之间存在相关性的问题 。如图 14：

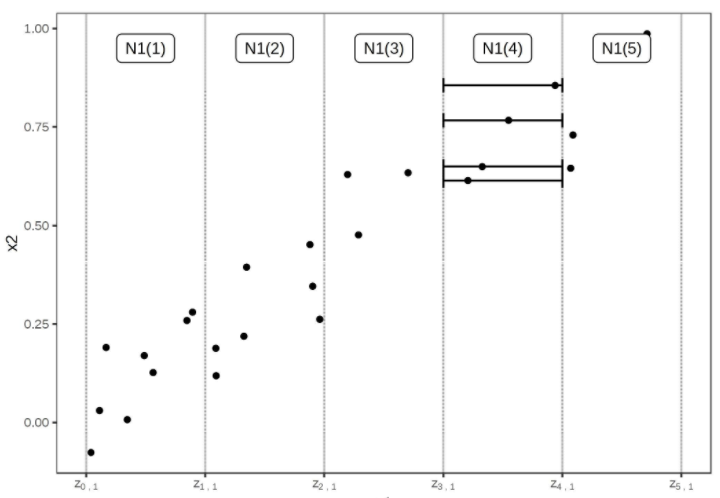


图 14

* 特征X1​ 和 X2​ 存在一定的线性关系，意味着假设特征之间互不相关的算法PDP无法应用。
* 当这个特征与其他特征有一定相关性时，部分依赖图会输出超过实际情况的预测结果。
* ALE将特征划分为间隔(垂直线)，这样在间隔中的点很大程度是符合实际的点。
* 对于间隔中的数据实例(点)，用间隔的上限和下限(⽔平线) 替换特征时，计算预测值的差异。这些差异被累积并中心化，形成ALE曲线。

### 置换特征重要性

置换特征重要性（Permutation feature importance）是首先由Breiman[[7]](#footnote-7) (2001)提出用在随机森林模型上， 后Fisher[[8]](#footnote-8)等(2019)把它一般化，用在任何一个模型上。

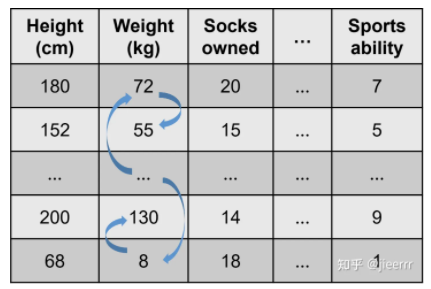
**算法4.2.2.1 置换特征重要性**

输入: 训练模型 , 特征矩阵 , 目标向量 , 误差度量 。  
估计原始模型误差 (例如均方误差)。  
for 特征 to do  
通过置换数据 中的特征 生成特征矩阵 。这破坏了特征 与真实结果 之间的关联。  
基于置换数据的预测, 估计误差 。  
计算置换特征重要性, 使用比率: ; 或者可以使用差异:   
end for  
按降序对特征进行排序。

置换特征重要性：通过置换某个特征后，计算模型预测误差的变化程度来衡量该特征的重要性。举例如下：

**案例4.2.2.1 置换特征重要性**

用左图的特征来预测一个人的运动能力，改变体重的数据，比如随机交换里面的数据，发现模型的预测误差明显增加了，说明体重特征很重要。相反，如果模型的预测误差几乎没有变化，说明体重特征不重要。



应该计算训练数据的重要性还是测试数据的重要性？

**是想知道模型在多大程度上依赖于每个特征来进行预测（训练数据），还是想知道特征对模型在未见数据上的性能有多大贡献（测试数据）。目前，还没有研究解决训练数据与测试数据的问题。**

置换特征重要性具有如下的特点：

* 特征重要性提供了对模型行为的高度压缩的全局洞察力。使用错误率代替错误差的一个积极方面是，在不同问题之间，特征重要性度量是可比较的。
* 重要性测量会自动考虑到与其他特征的所有交互作用。特征重要性既考虑了主要特征效应，也考虑了模型性能的交互效应。
* 置换特征的重要性不需要重新训练模型。

如果感兴趣的结果昂贵、耗时或难以测量（例如，因为它来自复杂的计算机模拟），则可以使用廉价且快速的结果替代模型。

基础模型是机器学习模型，**代理模型**必须是可解释的。代理模型的目的是尽可能准确地近似基础模型的预测，同时使其具有可解释性。

训练代理模型是一种与模型无关的方法，如果底层机器学习模型被另一个替换，仍然可以使用代理方法。黑盒模型类型和代理模型类型的选择是分离的。

执行以下步骤以获得代理模型：

**步骤4.2.2.1 代理模型的获取步骤**

1. 选择数据集 X。
2. 对于选定的数据集 X，获取黑盒模型的预测。
3. 选择可解释的模型类型（线性模型、决策树……）。
4. 在数据集 X 及其预测上训练可解释模型。
5. 衡量代理模型复制黑盒模型预测的程度。
6. 解释代理模型。

一种衡量代理复制黑盒模型能力的方法是R-平方:

其中：  
 是的是代理模型的第 个实例的预测， 是黑盒模型的预测和 是黑盒模型预测的平均值。SSE 代表误差平方和，SST 代表总平方和。 -平方度量可以解释为代理模型捕获的方差百分比。如果 - 平方接近 1 (= 低 SSE)，则可解释模型非常接近黑盒模型的行为。如果 平方接近 0 (= 高 SSE)，则可解释模型无法解释黑盒模型。

### LIME

LIME（Local Interpretable Model-agnostic Explanations）算法是在Marco Tulio Ribeiro2016年发表的论文《“Why Should I Trust You?” Explaining the Predictions of Any Classiﬁer》中介绍的。

Local: 基于想要解释的预测值及其附近的样本，构建局部的线性模型或其他代理模型。Interpretable: LIME做出的解释易被人类理解。Model-Agnostic: LIME解释的算法与模型无关，无论是用Random Forest、SVM还是XGBoost等各种复杂的模型，得到的预测结果都能使用LIME方法来解释。Explanations: LIME是一种解释方法。

LIME专注于训练局部代理模型以解释单个预测，即构造局部样本特征和预测结果之间的关系，利用局部可解释的模型对黑盒模型的预测结果进行解释。

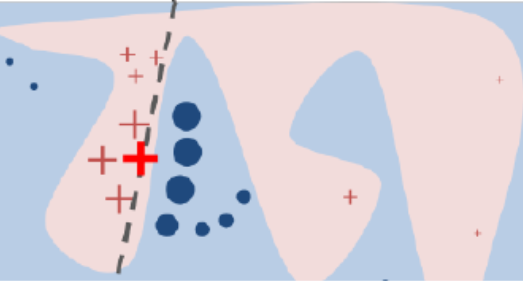
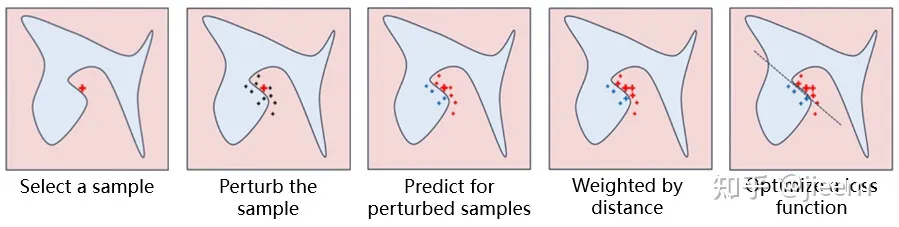


图 15

如图 15，红色/蓝色区域边界为决策函数，表示一个复杂模型，很难从全局用一个可解释的模型（例如线性模型）去逼近拟合它。这说明了，关注点从全局放到局部时，局部是可以用线性模型去拟合的。

**步骤4.2.3.1 LIME算法**

* 1. 选取关注的样本点，如图加粗的红色十字表示需要解释的样本点x；
  2. 在该样本点周围进行扰动采样z；
  3. 将采样出的样本z用复杂预测模型并得到结果（一些样本红色，一些样本蓝色点）；
  4. 定义一个**相似度**计算方式，按照他们到x的距离赋予样本权重；
  5. 选取的K个特征来解释，通过采样样本学习**局部可解释模型**(如线性或决策树)，

虚线表示。

**（1）目标函数中引入相似度**

对这个样本进行扰动，对扰动前后的**样本相似度的距离**进行了定义，使用**指数核函数**定义，这取决于样本的类型（文本用余弦相似性，图像用L2范数）

**公式4.2.3.1 指数核函数**

σ 控制核函数的作用范围，其值越大，核函数的局部影响范围就越大。

文本数据使用余弦相似性度量，先将文档数据整理成文档—词矩阵格式，如图 16：

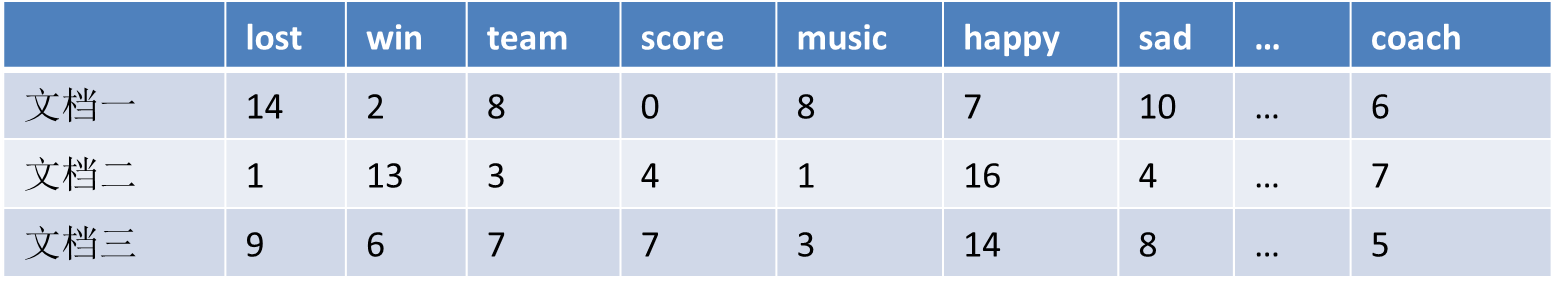


图 16

两个文档之间的余弦相似度的计算公式为：

similarity

**概念阐释4.2.3.1 L2范数**

范数是具有“长度”概念的函数。在向量空间内，为所有的向量的赋予非零的增长度或者大小。不同的范数，所求的向量的长度或者大小是不同的。

举例：2维空间中，向量(3,4)的长度是5，那么5就是这个向量的一个范数的值，更确切的说，是欧式范数或者L2范数的值。

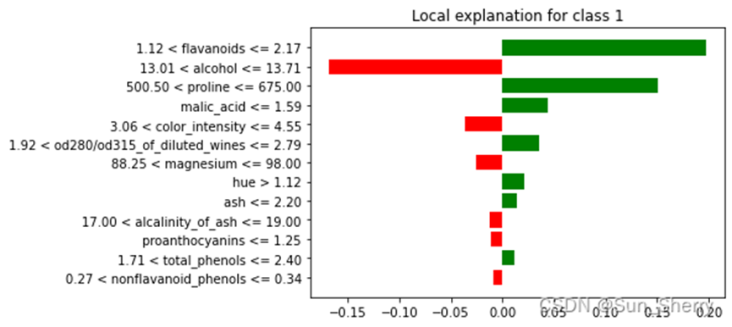
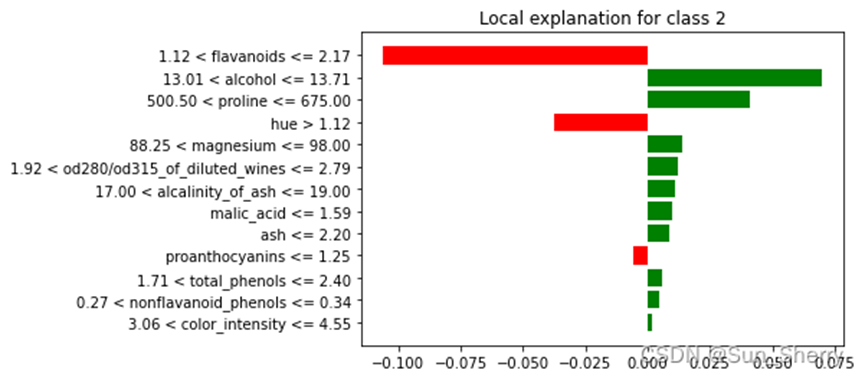
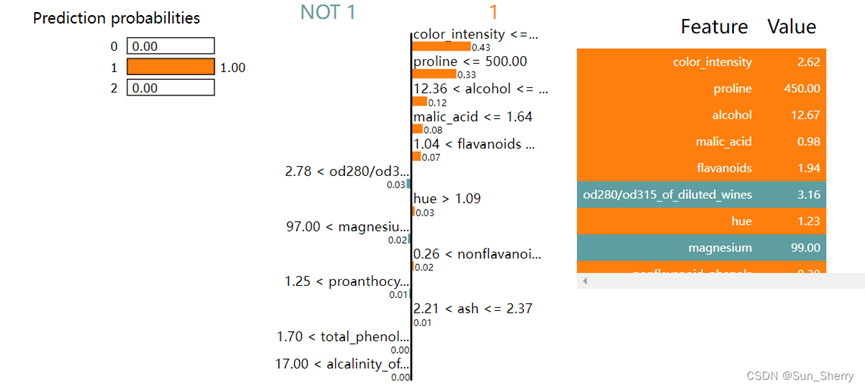
对于 -范数，如果

那么向量 的 -范数就是

L1范数 :

L2范数 :

**（2）目标函数**

其中*f*(z)就是扰动样本在d维空间（原始特征）上的预测值，*g*(z’)则是在d’维空间（可解释特征）上的预测值，然后以相似度作为权重。上述的目标函数可以通过线性回归的方式进行优化。

**案例4.2.3.1 LIME在不同类型数据上的应用**

（1）表格数据的LIME

（2）文本数据的LIME

（3）图像数据的LIME

**（3）LIME方法的特性**

* 可解释性。要求解释器的模型与特征都必须是可解释的，像决策树、线性模型都是很适合拿来解释的模型；而可解释的模型必须搭配可解释的特征，才是真正的可解释性，让不了解机器学习的人也能通过解释器理解模型。
* 局部保真度。不可能期望简单的可解释模型在效果上等同于复杂预测模型。解释器不需要在全局上达到复杂模型的效果，但至少在局部上效果要很接近，而此处的局部代表想观察的那个样本的周围。
* 与模型无关。与复杂预测模型无关，无论多复杂的预测模型，像是SVM或神经网络，该解释器都可以工作。

“汤姆给你 3 个金币，因为你们是朋友，你应该接受它; 如果你要公正的话，那么公正的分法是，你应当得到1个金夏普利币，而你的朋友汤姆应当得到7个金币。”

* 诺贝尔奖获得者

美国运筹学家  
罗伊德夏普利 ( Lloyd S. Shapley )

**案例引入 4.2.4.1 shapley值**



采用逻辑建模方法归纳出了三条合理分配原则，即在 人合作博奕 中，参与人 从 人大联盟博栾所获得的收益 应当满足的基本性质 (用公理形式表示 )，进而证明满足这些基本性质的合作博弈解是惟一存在的，从而妥善地解决了某类合作博峦的合理分配问题。



### SHAP

**（1）Shapley值的定义**

谈起沙普利，许多中国学者会对他有一种天然的亲切感，这主要源于他曾经在中国的土地上与中国军民并肩抗击过日本侵略军。1943年，作为哈佛大学数学系的一名本科生，他应征入伍成为一名空军中士，并很快奔赴中国成都战区。当时，沙普利就展现出卓越的数学天才，曾因为破解气象密码获得铜星奖章 (Bronze Star)。

战争结束后，沙普利回到哈佛大学继续念书，在1948年取得数学学士学位，随后进入普林斯顿大学数学系，一路念到博士毕业，他的博士导师也是纳什的导师：塔克（A.W.Tucker）教授。此后，他长期在美国著名的“战略思想库”兰德公司工作，1981年后，则一直担任美国加州大学洛杉矶分校数学和经济系教授。

2002年8月14日到17日，沙普利因为参加青岛大学承办的“2002国际数学家大会对策论及其应用’卫星会议”，再次来到中国。青岛大学的高红伟教授作为会议组织者，至今还留着一份为沙普利办理入境签证时青岛市政府出具的邀请函原件，“沙普利被誉为博峦论的无㴔之王，精通博弈理论，但却不太喜欢现代的信息技术，不喜欢使用电子邮件与别人进行沟通”。昔日的英武少年已成为一个科学“老顽童”，他不拘小节，动不动玩“消失”，高红伟派十几个学生把他找到时，才发现79岁的教授没有回宾馆，竟然在大厅的沙发上睡了，而且一睡就是3个小时。

青岛之行，沙普利再次讲述起他与中国将近等的那段渊源时，依然激动。

**案例引入 4.2.4.2 沙普利的故事**

满足下列三条分配原则的 ，称为Shapley值：

* 对称性原则：每个参与人获得的分配与他在集合 中的排列位置无
* 有效性原则：  
  ①若参与人 i对他所参加的任一合作都无贡献，则给他的分配应为 0 。数学表达式为: 任意 ，若 ，则 。  
  ②完全分配:
* 可加性原则：对 上任意两个特征函数 U 与   
  可加性原则表明 : 个人同时进行两项互不影响的合作，则两项合作的分配也应互不影响，每人的分配额是两项合作单独进行时应分配数的和。

任一 人合作博弈 ，Shapley值是唯一存在的，且 。 为集合 的元素个数：

Shapley值可以作为局中人 在联盟 的贡献 的一个加权因子。因此局中人树所有他可能参加的联盟所作贡献的加权平均 (期望值) 就是Shapley值。

实际计算中通过估计方式获得shaplev value的:

**案例4.2.4.1**

设有 三人经商。若各人单干，则每人仅能获利 10 万元 ; 若 合作, 可获利70万元； A、C 合作可获利50万元； B、C 合作可获利40万元 ; 三人合作可获利100万元。

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |

又因为   
因此三人合作可获利最大  
为 确定一个分配方案

参与人 的分配计算

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  | 10 | 70 | 50 | 100 |
|  | 0 | 10 | 10 | 40 |
|  | 10 | 60 | 40 | 60 |
|  | 1 | 2 | 2 | 3 |
|  | 2 | 1 | 1 | 2 |
|  |  |  |  |  |
|  |  |  |  |  |

同理可计算出参与人 的分配值

（1）列出所有子集，计算某一子集S上 的contribution时:

* 固定 中的feature的value，固定 的feature value，其他feature通过在数据集中多次随机抽样填充形成多个拼接样本 。用模型对这些合成样本进行预测获得预测结果。计算预测结果的平均值作为 （即除 和 外其他feature都是平均值）；
* 固定这S中的feature的value，连同 和其他feature一同做随机抽样替换，并计算预测值平均值（即目标feature 也计算平均值，就是w/o该feature的情况 ）；
* 两值相减，得到目标feature在子集 上的贡献值；

（2）计算所有子集上目标feature的贡献值，并做加权和，得到feature 的贡献值;

因此：

* Shapley value是针对于单个instance 的，不同instance的特征对应的Shapley value都不一样。
* 该实际估计方法本质上是用抽样频率估计来逼近理论概率值，逼近边际贡献。
* 计算所有子集 ，时间复杂度高。（指数级 ，即使不遍历子集 的各种排列也仍需要大量采样，后面SHAP给出新的估算方法)

SHAP与Shapley value的区别：

* SHAP(**SH**apley **A**dditive ex**P**lanation)：利用shapley value构建了一个加性的解释模型。对于某一个预测值pred，不同feature在其中起到的作用，可以理解为在“预测值均值”的基础上每个feature不断“拉扯”最后得到的预测值pred
* SHAP估计方法：kernelSHAP 高效的估算shapley value的方式
* 提供了局部与全局模型解释方法（Shapley value 是 instance-level的）

SHAP通过加减从预测值均值到预测值，用于解释模型。

其中， 为预测值均值， 为Shapley value， 是样本的指示向量， (i) 表示第 条样本。该公式解释的是对某一个instance 上特征贡献。  
**如何计算其中的每个 , 即Shapley值?**

Kernel SHAP Linear LIME + Shapley values。Kernel SHAP是**model-agnostic**。其中，用 Kernel SHAP 计算每个指示向量z对应的kernel权重

其中， 是总feature个数， 是 中 1 的个数/选定的feature个数/子集 中的特征个数。

其中，原始数据空间 , 计算对应的预测值 .

除了Kernel SHAP之外，tree SHAP可以模型无关，也可以根据树模型本身的结构，建立专门针对于树模型的计算Shapley Value的算法。

SHAP有如下的结果可视化方法：

**（1）样本特征影响图：waterfall**

import xgboost

import shap

shap.initjs()

# train an XGBoost model

X, y=shap.datasets.boston( )

model = xgboost.XGBRegressor (). fit (X, y)

# explain the model's predictions using SHAP

# (same syntax works for LightGBM, CatBoost, scikit-learn, transformers, Spark, etc.)

explainer = shap. Explainer(model)

shap\_values = explainer (X)

# visualize the first prediction's explanation

shap.plots.waterfall(shap\_values [0] )

**代码4.2.4.1 样本特征影响图waterfall的Python实现**

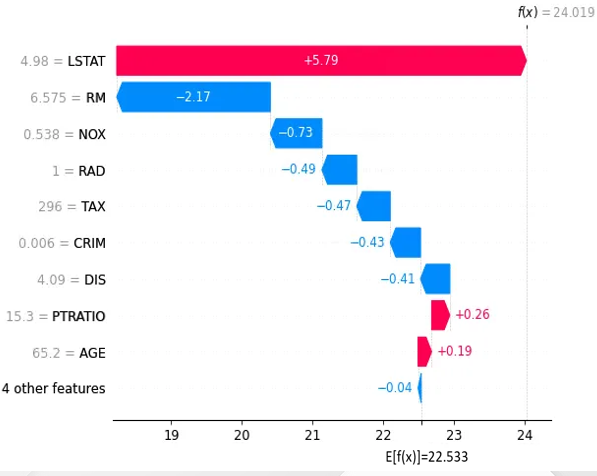


图 17

如图 17：

* Y轴代表不同特征值
* X代表SHAP值
* 红色代表该特征对第0个样本是正向影响，蓝色为负向
* E[f(x)]代表所有样本f(x)的期望base\_values，model.predict(X)的预测值的平均值
* f(x)值的大小为第0个样本的预测值，f(x) = E[f(x)]+特征值 \* 特征SHAP值

**（2）样本特征影响图：force plot**

# visualize the first prediction's explanation with a force plot

shap.plots.force(shap\_values[0])

shap\_values[0]

.values =

array([-4.2850167e-01, -6.6636719e-02, 7.7860229e-02, -1.5295845e-03,

-7.2922713e-01, -2.1700280e+00, 1.9213372e-01, -4.1425943e-01,

-4.9156108e-01, -4.7296646e-01, 2.5669456e-01, -5.3970132e-02,

5.7883248e+00], dtype=float32)

.base\_values =

22.532942

.data =

array([6.320e-03, 1.800e+01, 2.310e+00, 0.000e+00, 5.380e-01, 6.575e+00,

6.520e+01, 4.090e+00, 1.000e+00, 2.960e+02, 1.530e+01, 3.969e+02,

4.980e+00])

其中：

* `shap\_values[0]` - 第0个样本
* `shap\_values[0].base\_values` - 22.53，所有特征未应用的平均 f(x) 值，所以所有特征都不起作用时模型在数学期望上的输出为22.53
* `shap\_values[0].data` - 所有特征值信息，该数据形状为(506, 13)，第0个样本13个特征的具体值
* `shap\_values[0].values` - 第0个样本13个特征的shap值

**代码4.2.4.2 样本特征影响图force plot的Python实现**

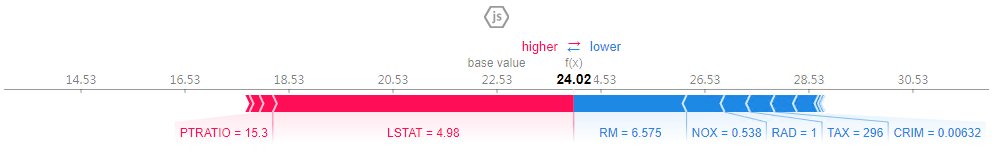


图 18

如图 18：

* 所有样本基础平均f(x)值:22.53,该样本shap值加总之后为f(x) -> 24.02
* 红色代表:正向影响，LSTAT这个特征，正向影响，影响最大；PTRATIO也是正向
* 蓝色代表：负向影响，RM特征，负向影响

**（3）特征密度散点图：beeswarm**

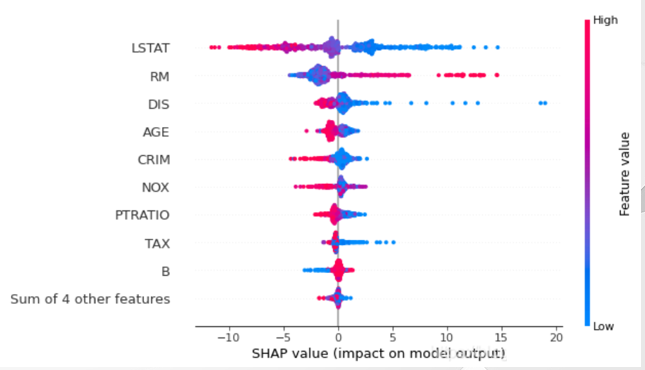


图 19

如图 19：

* 一行代表一个特征，横坐标为Shap值。
* 一个点代表一个样本，越红说明特征值越大，越蓝说明特征值越小。
* 特征重要性的排列图
* LSTAT对模型非常重要，而且LSTAT特征值越小，shap大于0，正向影响
* LSTAT这个特征，样本分布较为分散，表明该特征影响越大

**（4）特征密度散点图：summary\_plot**

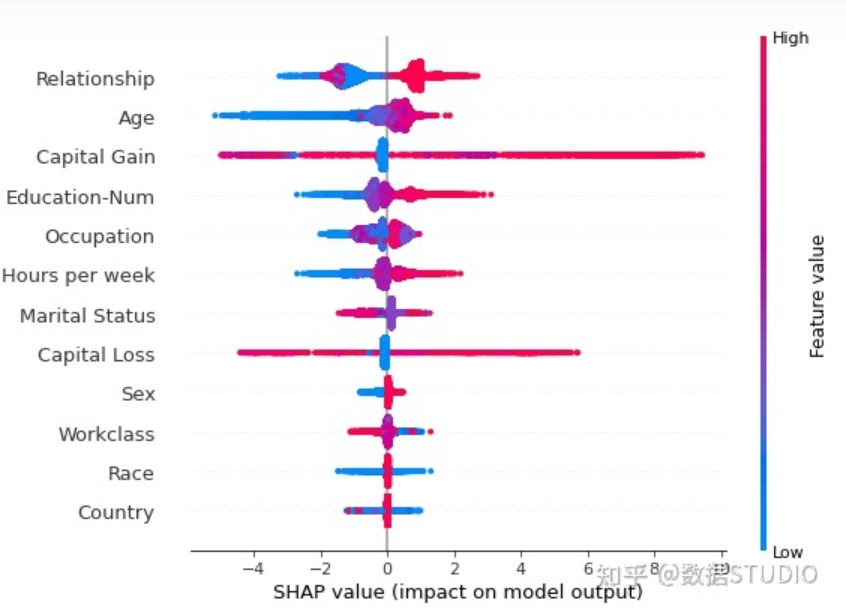


图 20

**（5）特征重要性SHAP值**

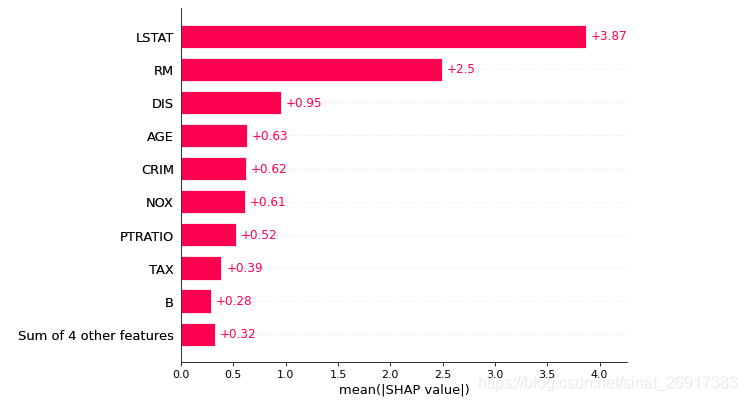


图 21

# 可解释机器学习的未来

## 重点将放在与模型无关的可解释性工具上

* 当可解释性与机器学习模型分离时，自动化会变得容易的多。
* 模型训练的自动化已是一个显而易见的趋势，包括特征选择（自动识别特征是否与预测相关）、参数优化、不同模型的比较（检测模型是否对现实的最好近似），结果的展示（自动检测关系的表示方法）等。
* 与模型无关的可解释方法便于实现可解释的自动化，可以自动应用于机器学习过程中出现的任何模型。

## 分析模型重于分析数据

相比关心数据，更关心数据中包含的知识。可解释的机器学习是从数据中提取知识的好办法。

## 机器人和程序会自我解释

为需要大量使用机器学习的机器和程序配备更直观的界面。例如：信用违约程序向银行员解释为何拒绝信用申请（申请人的信用卡过多，并且从事不稳定的工作）。

## 可解释性可以促进机器智能研究

改善对智能的理解，并更好地创造智能机器。

1. Z C Lipton. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery[J]. Communications of the ACM, 2018, 61(3):31-57. https://doi.org/10.1145/3236386.3241340. [↑](#footnote-ref-1)
2. 词嵌入(Word Embedding)：一种将文本中的词转换成数字向量的方法,把一个维数为所有词的数量的高维空间嵌入到一个维数低得多的连续向量空间中，每个单词或词组被映射为实数域上的向量 [↑](#footnote-ref-2)
3. J A McDermid, et al. Artificial intelligence explainability: the technical and ethical dimensions.

   Phylosophical Transactions A, 2020, 379:20200363. https://doi.org/10.1098/rsta.2020.0363 [↑](#footnote-ref-3)
4. archive.ics.uci.edu/ml/ [↑](#footnote-ref-4)
5. kaggle.com/c/bike-sharing-demand [↑](#footnote-ref-5)
6. 95%置信区间：指95%的值在这个范围 [↑](#footnote-ref-6)
7. L Breiman, Random forests. Machine learning, 2001. [↑](#footnote-ref-7)
8. A Fisher, et al., All models are wrong, but many are useful: Learning a variable’s importance by studying an entire class of prediction models simultaneously. Journal of Machine Learning Research, 2019. [↑](#footnote-ref-8)