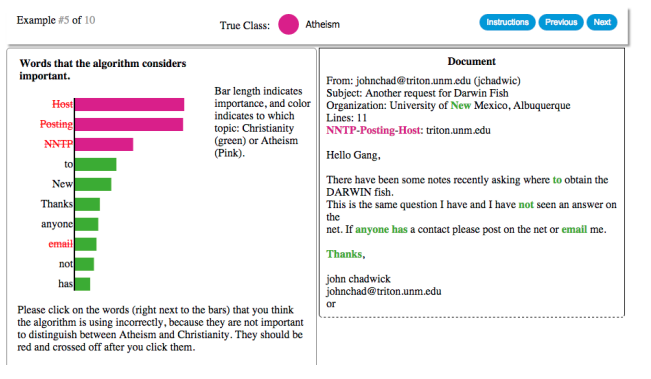
LIME实验

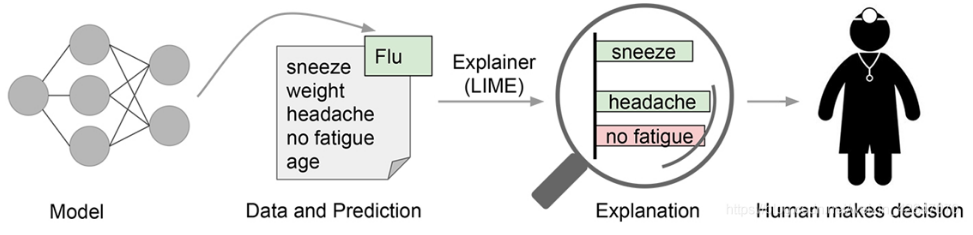
实验:文本分类器(左边是预测的备选项[无神论和基督教]和可能的概率,中间是它被预测成christian的原因分析,右边是这些原因在文本中的位置)

本例说明的问题是:有时候分类器分类的理由是不正确的.例如在本文中分类Atheism的特性是根据Posting来进行分类,但显然这个词不是分类原因.

模型可解释性:

在算法建模过程中,我们一般会用测试集的准确率与召回率衡量一个模型的好坏.但在和客户的实际沟通时,单抛出一个数字就想要客户信任我们,那肯定是不够的,这就要求我们摆出规则,解释模型.但不是所有的模型都是规则模型,一些黑盒模型(比如神经网络)有着更高的准确率,但是无法给出具体的规则,无法让普通人理解和信任模型的预测结果.因此,我们需要一个模型解释器对模型进行解释.

可解释性的目的:

图中,病人的症状(输入)产生了flu的结果(输出),现在对这一个样本进行解释.可以看到,sneeze和headache(喷嚏和头痛)是主要部分,no fatigue(无疲劳)则是产生负结果的部分.

1.帮助客户解决模型的可解释性问题,即为什么模型是对的.

2.帮助模型设计者排除不必要的因素,提高分类器效率.

结构: 需要一个解释器,需要一个判定解释器好坏的拟合器(将解释拟合到样本中). (需要原始模型?--黑盒)

解释器的要求:

1. 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。

2. 局部性: 既然我们已经使用了可解释的模型与特征，就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。

3. 模型无关性: 这里所指的是与复杂模型无关，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。

LIME算法(2016):

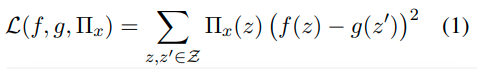
给出一个简单模型g(例如g=Wx).

给出复杂性度量Ω(g)(例如文本分类中,K是分类词分类数量,那么

复杂度应当最低,f和g应当足够相似,因此给出loss函数的定义:



具体的,根据样本点有L的定义和相似度\pi的定义:





基于定义,计算待解释的样本x在可解释维度上(即可被简单模型拟合的维度)x’的K个待解释特征的扰动z’.将z’还原到原始维度,计算预测值与相似度,再距离加权,利用回归模型分析其影响系数.

举例(判断一个用户是否会违约):



模型:GBDT分类模型

采用LIME进行解释:

from lime.lime\_tabular import LimeTabularExplainer

