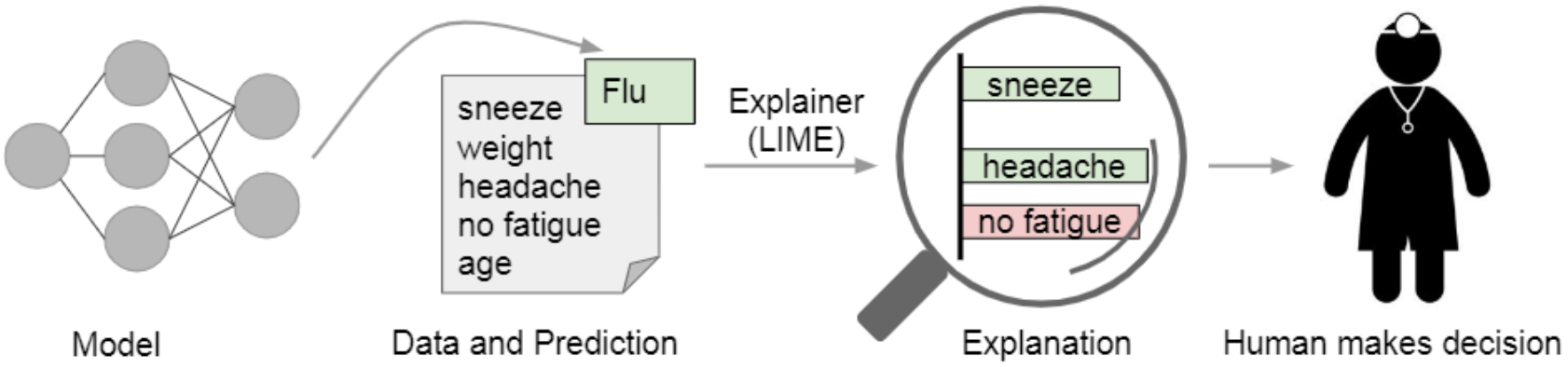
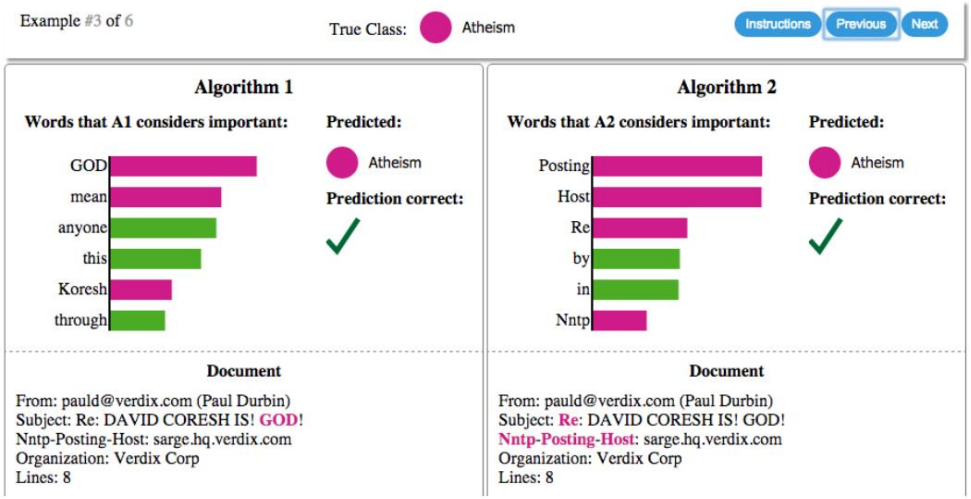
LIME: Local Interpretable Model-Agnostic Explanation

与模型无关的局部可解释性的解释

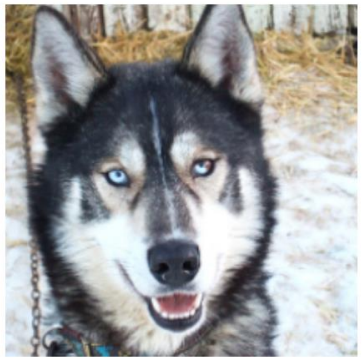
20211201



对“黑盒子”的可解释性，其实就是体现在feature importance中，TOP N个重要的特征就能很好的对结果进行解释，如图中对一个人是否有流感的预测。



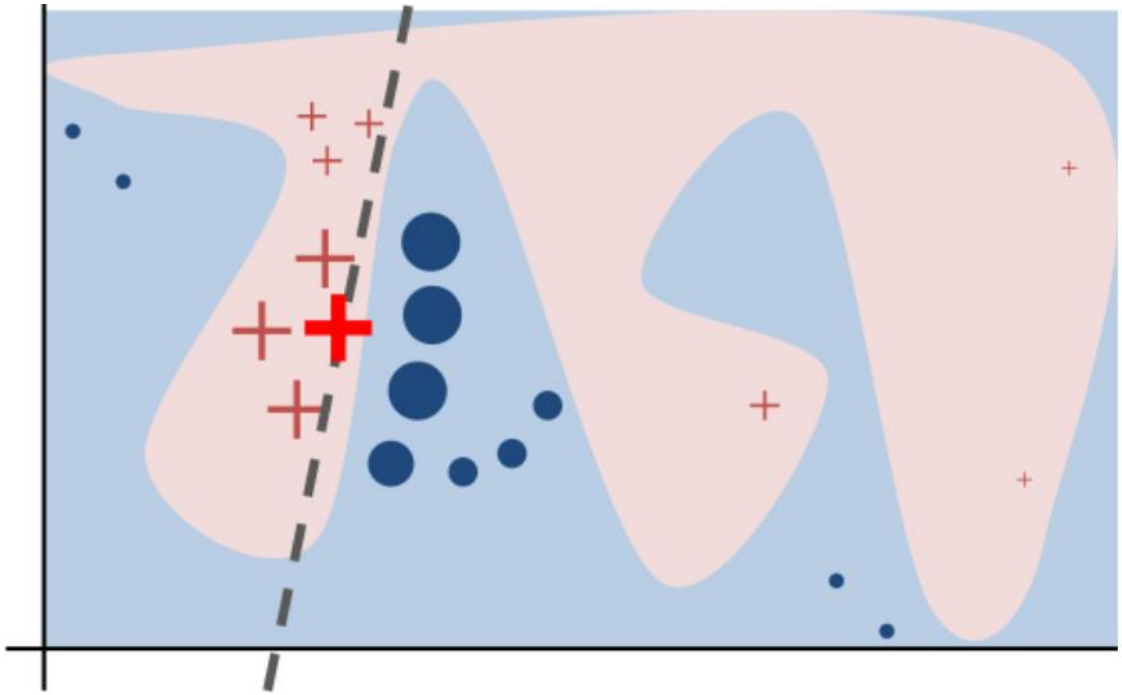
Posting（邮件标头的一部分）在无神论文章中出现的频次很高



(a) Husky classified as wolf

(b) Explanation

背景是白色的雪地得到是狼的结论

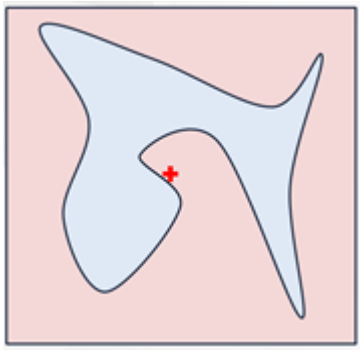


红色和蓝色区域表示一个复杂的分类模型（黑盒），图中加粗的红色十字表示需要解释的样本，显然，这是很难从全局用一个可解释的模型（例如线性模型）去逼近拟合它。但是，当把关注点从全局放到局部时，可以看到在某些局部是可以用线性模型去拟合的。

LIME通过扰动输入样本（**Perturb the input**），来判断哪些特征的存在与否，对于输出结果有着最大的影响。而扰动的精髓在于这些绕扰动必须是人类可以理解的。比如，在一张图片中将部分区域进行遮盖。

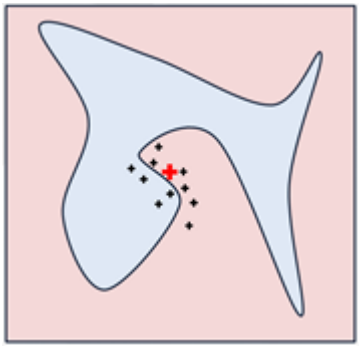
Quickshit超像素分割算法

分类器

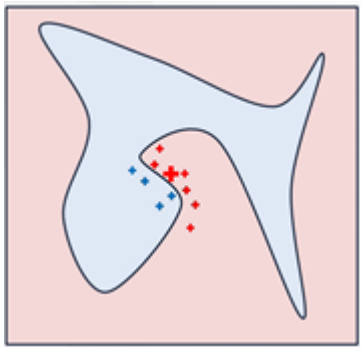


选定一个要解释的样本x，以及在可解释纬度上的x'

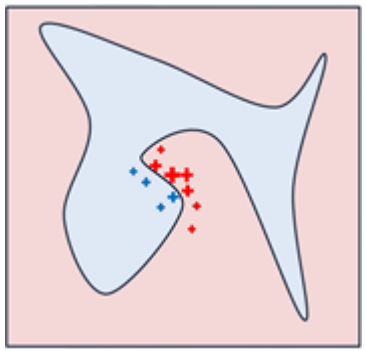
选取的K个特征来解释



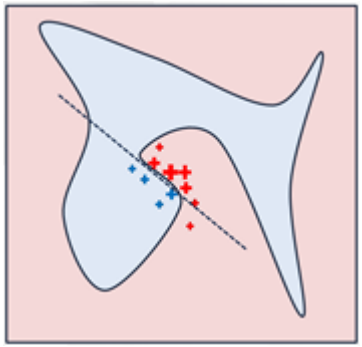
进行N次perturb扰动，z'，从x'扰动而来



将z'还原到d维度，并计算预测值f(z)以及相似度



根据距离进行加权



收集到N次扰动的样本后，利用岭回归取得对这个样本有影响力的系数

Algorithm 1 Sparse Linear Explanations using LIME

Require: Classifier f, Number of samples N
Require: Instance x, and its interpretable version x'
Require: Similarity kernel π_x, Length of explanation K

$\mathcal{Z} \leftarrow \{\}$
for $i \in \{1, 2, 3, ..., N\}$ **do**
  $z'_i \leftarrow sample\_around(x')$
  $\mathcal{Z} \leftarrow \mathcal{Z} \cup \langle z'_i, f(z_i), \pi_x(z_i) \rangle$
end for
$w \leftarrow$ K-Lasso$(\mathcal{Z}, K)$ ▷ with z'_i as features, $f(z)$ as target
return w

LIME伪代码

（1）目标函数：

$$\xi = \underset{g}{\operatorname{argmin}} L(f, g, \pi_x) + \Omega(g)$$

$g \in G$: 定义的解释模型g

$\pi_x(z)$: 实例z与x之间的接近度

L: π_x局部定义下，g如何逼近f（复杂模型）

$\Omega(g)$: 解释模型复杂度

（2）引入相似度后的目标函数：

$\pi_x(z) = \exp(-D(x,z)^2/\sigma^2)$　　文本是余弦相似性，图像是L2范数距离

扰动前后的样本相似度的距离

（3）最终函数：

$$\xi = \underset{g}{\operatorname{argmin}} L(f, g, \pi_x) + \Omega(g) = \sum_{z', z \in Z} \pi_x(z)\big(f(z) - g(z')\big)^2$$

$f(z)$: 扰动样本在d维空间（原始特征）上的预测值

$g(z')$: 扰动样本在d'维空间（可解释特征）上的预测值

对来自 Inception 的预测的解释。
前三个预测类别是“树蛙”、
“台球桌”和“气球”。

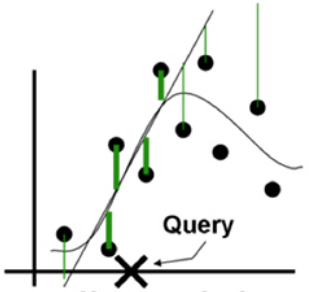
P() = 0.54

P() = 0.07

P() = 0.05

Original Image
P(tree frog) = 0.54

Perturbed Instances	P(tree frog)
	0.85
	0.00001
	0.52



Locally weighted regression

Explanation

P() = 0.54

P() = 0.07

P() = 0.05