论文笔记understanding black-box predictions via influence functions

原创 2017年09月15日 15:34:57

Purpose

To formalize the impact of a training point on a prediction, we ask the counterfactual: what would happen if we did not have this training point, or if the values of this training point were changed slightly?

Main method

Answering this question by perturbing the data and retraining the model can be prohibitively expensive. To overcome this problem, we use influence functions, a classic technique from robust statistics (Cook & Weisberg, 1980) that tells us **how the model parameters change as we upweight a training point** by an infinitesimal amount.

Approach

Consider a prediction problem from some input space $\mathcal{X}$ (e.g., images) to an output space $\mathcal{Y}$ (e.g., labels). We are given training points $z_1, \ldots, z_n$, where $z_i = (x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$. For a point z and parameters $\theta \in \Theta$, let $L(z, \theta)$ be the loss, and let $\frac{1}{n}\sum_{i=1}^{n} L(z_i, \theta)$ be the empirical risk. The empirical risk minimizer is given by $\hat{\theta} \stackrel{\text{def}}{=} \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} L(z_i, \theta)$.[1] Assume that the empirical risk is twice-differentiable and strictly convex in θ;

1.Upweighting a training point

how would the model’ s predictions change if we did not have this training point?

1.1删去一个样本后，模型参数的变化（参数增大了多少）

将某个样本z从训练集中删去后，模型新参数与原来参数相比，变化了：

$$\hat{\theta}_{-z} - \hat{\theta},$$

其中:

$$\hat{\theta}_{-z} \stackrel{\text{def}}{=} \arg\min_{\theta\in\Theta} \sum_{z_i \neq z} L(z_i, \theta)$$

如果把每个样本都去掉后各自训练一个新模型，再分别计算参数变化，显然需要很大的计算量。influence fucntions可以帮助我们解决这个问题。

首先，考虑将某一个样本z的权重增加一点点，计算新的模型的参数：

$$\hat{\theta}_{\epsilon,z} \stackrel{\text{def}}{=} \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} L(z_i,\theta) + \epsilon L(z,\theta)$$

那么，样本z的权重增加的大小对模型参数变化的影响可以用下式表示表示：

$$\mathcal{I}_{\text{up,params}}(z) \stackrel{\text{def}}{=} \left.\frac{d\hat{\theta}_{\epsilon,z}}{d\epsilon}\right|_{\epsilon=0} = -H_{\hat{\theta}}^{-1}\nabla_\theta L(z,\hat{\theta}), \quad (1)$$

其中:

$$H_{\hat{\theta}} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n}\nabla_\theta^2 L(z_i,\hat{\theta})$$

为Hessian矩阵，而且满足正定的假设。

删去一个样本，就等同于将这个样本的权重增加（-1/n），n为总样本数目。

因此，删除样本z后模型参数的变化（新参数-原来参数）就可以用下式估计：

$$-\frac{1}{n}\mathcal{I}_{\text{up,params}}(z)$$

1.2 删去一个样本z后，模型在测试样本test上预测误差的变化（误差增加了多少）

首先，样本z权重增加的大小，对模型在测试样本上loss变化的影响程度为：

$$\begin{aligned}\mathcal{I}_{\text{up,loss}}(z,z_{\text{test}}) &\stackrel{\text{def}}{=} \left.\frac{dL(z_{\text{test}},\hat{\theta}_{\epsilon,z})}{d\epsilon}\right|_{\epsilon=0} \quad (2)\\ &= \nabla_\theta L(z_{\text{test}},\hat{\theta})^\top \left.\frac{d\hat{\theta}_{\epsilon,z}}{d\epsilon}\right|_{\epsilon=0}\\ &= -\nabla_\theta L(z_{\text{test}},\hat{\theta})^\top H_{\hat{\theta}}^{-1}\nabla_\theta L(z,\hat{\theta}).\end{aligned}$$

那么，由于删去一个样本等同于权重增加（-1/n），因此，删除一个样本z后，模型在测试样本上的loss会增加(-1/n)*Iup,loss(z，test)

2.Perturbing a training point

how would the model’ s predictions change if a training input were modified?

2.1 对样本z增加扰动后，模型参数会变化多少？

给训练样本z增加扰动，使其变成z_delta:

For a training point $z = (x, y)$, define $z_\delta \stackrel{\text{def}}{=} (x + \delta, y)$. Consider the perturbation $z \mapsto z_\delta$, and let $\hat{\theta}_{z_\delta,-z}$ be the empirical risk minimizer on the training points with z_δ in place of z. To approximate its effects, define the parameters resulting from moving ϵ mass from z onto z_δ: $\hat{\theta}_{\epsilon,z_\delta,-z} \stackrel{\text{def}}{=} \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} L(z_i,\theta) + \epsilon L(z_\delta,\theta) - \epsilon L(z,\theta)$. An

将样本z的权重移动一点点到新样本上，权重移动的大小对模型参数变化的影响为：

$$\left.\frac{d\hat{\theta}_{\epsilon,z_\delta,-z}}{d\epsilon}\right|_{\epsilon=0} = \mathcal{I}_{\text{up,params}}(z_\delta) - \mathcal{I}_{\text{up,params}}(z)$$
$$= -H_{\hat{\theta}}^{-1}\left(\nabla_\theta L(z_\delta,\hat{\theta}) - \nabla_\theta L(z,\hat{\theta})\right). \quad (3)$$

当样本z被删去，取而代之的是样本z_delta时，就相当于把样本z的权重都移动到新样本上了，此时，模型参数的变化量为：

$$\hat{\theta}_{z_\delta,-z} - \hat{\theta} \approx -\frac{1}{n}H_{\hat{\theta}}^{-1}[\nabla_x \nabla_\theta L(z,\hat{\theta})]\delta$$

2.2 样本z增加扰动后，模型在测试样本上的预测loss会增大多少？

如果样本数据x连续，扰动量delta很小很小，公式3可以进一步写成：

$$\left.\frac{d\hat{\theta}_{\epsilon,z_\delta,-z}}{d\epsilon}\right|_{\epsilon=0} \approx -H_{\hat{\theta}}^{-1}[\nabla_x \nabla_\theta L(z,\hat{\theta})]\delta. \quad (4)$$

此时，若将样本z替换成扰动后的样本z_delta,模型的参数会有如下的变化：

$$\hat{\theta}_{z_\delta,-z} - \hat{\theta} \approx -\frac{1}{n}H_{\hat{\theta}}^{-1}[\nabla_x \nabla_\theta L(z,\hat{\theta})]\delta.$$

扰动量delta的大小对新模型在测试样本z_test上面的loss变化量的影响可以用如下方式衡量：

$$\mathcal{I}_{\text{pert,loss}}(z,z_{\text{test}})^\top \stackrel{\text{def}}{=} \nabla_\delta L(z_{\text{test}},\hat{\theta}_{z_\delta,-z})^\top\Big|_{\delta=0} \quad (5)$$
$$= -\nabla_\theta L(z_{\text{test}},\hat{\theta})^\top H_{\hat{\theta}}^{-1}\nabla_x \nabla_\theta L(z,\hat{\theta}).$$

那么，当扰动量大小为delta时，模型在z_test上面的预测loss会增大这么多：

$$\mathcal{I}_{\text{pert,loss}}(z,z_{\text{test}})^\top\delta$$

因此，沿着公式5的方向调整delta的大小，就可以构建出使模型对z_test的预测loss最大的z的扰动样本。

分析影响influence的因素

Let $p(y \mid x) = \sigma(y\theta^\top x)$, with $y \in \{-1,1\}$ and $\sigma(t) = \frac{1}{1+\exp(-t)}$. We seek to maximize the probability of the training set. For a training point $z = (x,y)$, $L(z,\theta) = \log(1+\exp(-y\theta^\top x))$, $\nabla_\theta L(z,\theta) = -\sigma(-y\theta^\top x)yx$, and $H_\theta = \frac{1}{n}\sum_{i=1}^n \sigma(\theta^\top x_i)\sigma(-\theta^\top x_i)x_i x_i^\top$. From (2), $\mathcal{I}_{\text{up,loss}}(z,z_{\text{test}})$ is:

$$-y_{\text{test}}y \cdot \sigma(-y_{\text{test}}\theta^\top x_{\text{test}}) \cdot \sigma(-y\theta^\top x) \cdot x_{\text{test}}^\top H_{\hat{\theta}}^{-1}x.$$

1.sigma(-y*theta^T*x)越大——>L(z,theta)越大——>训练误差越大的样本，I_up,loss越大，即影响越大。因此，外点（outliers）也可以dominate模型的参数。

2.如果 $\nabla_\theta L(z,\hat{\theta})$ 指向一个变化很小的方向，意味着L(z,theta^)不怎么被theta的变化影响，那么这个样本z就是一个具有higher influence的样本，因为沿着这个方向移动不会显著增加其他训练样本的loss。

实例

1.understanding model behavior

we wanna find a training sample that is the most responsible for a given prediction on a test image

将每个样本都删去后，都计算一下新模型在test image上的预测loss的变化，若I_up,loss为正，则说明删去这个样本后预测误差增大，即这个样本z对test image是helpful的。反之则为harmful。

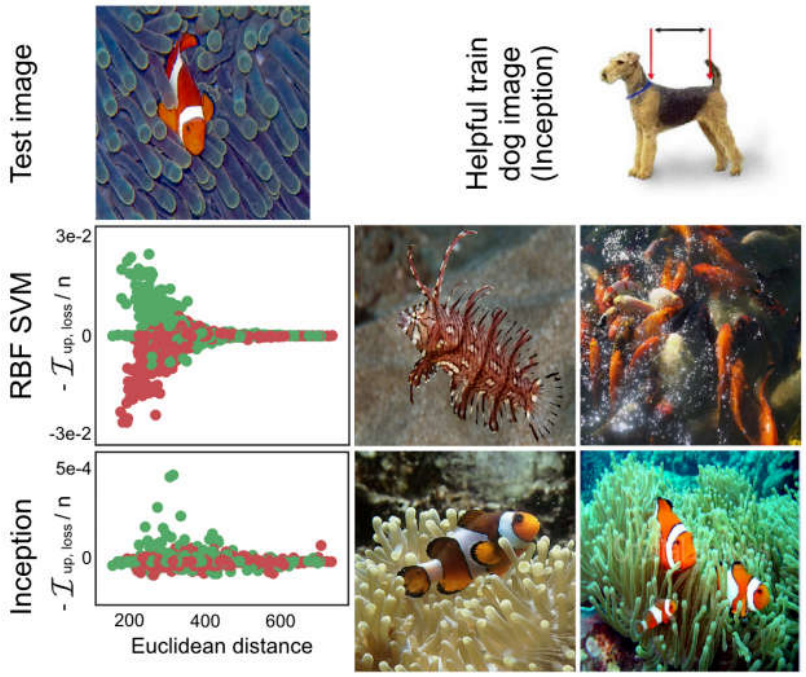

Figure 4. **Inception vs. RBF SVM. Bottom left:** $-\mathcal{I}_{\text{up,loss}}(z, z_{\text{test}})$ vs. $\|z - z_{\text{test}}\|_2^2$. Green dots are fish and red dots are dogs. **Bottom right:** The two most helpful training images, for each model, on the test. **Top right:** An image of a dog in the training set that helped the Inception model correctly classify the test image as a fish.

用RBF-SVM 和 Inception分别构建一个dog-fish二分类分类器，可以看出：
对RBF-SVM，与test image的欧几里得距离越大的train image，influence越小；反之，与test image欧几里得距离最小的train image，删去后模型在test image上的预测loss变化越大。（正方向上的代表loss变大，即为helpful training images，如绿色的训练样本；反方向为harmful training images，如红色的训练样本）
对Inception，模型在test image上的预测loss基本上与训练样本的种类关系不大，dog样本也会对模型正确预测出鱼有帮助。

2.对抗训练样本

对抗测试样本：在视觉上难以明显区分，且具有不同label，但可以完全愚弄分类器的测试样本。

对抗训练样本：视觉上难以明显区分，但却可以使分类器在同一个test样本上的预测结果完全相反的训练样本。
如何利用influence function生成对抗训练样本呢？

由于$\mathcal{I}_{\text{pert,loss}}(z, z_{\text{test}})$指的是loss对扰动的梯度，我们就知道了如何调整扰动量，可以使得在test上的预测loss增长最快啦。
对于每个test image，都可以给training images的几个image加入扰动，从而使得新模型在这个test image的预测结果完全颠倒。

经过attacks的training sets虽然在像素上变化很小，但是Inception 的feature layer上面的区别确实很大。说明inception model可以capture到这种attacks，可以通过检测feature vector来检测样本是否被attacked。另外，通过检查$\mathcal{I}_{\text{pert,loss}}(z, z_{\text{test}})$的数量级也可以量化模型对training sets attacks的抗击能力：越小，代表对training sets加入扰动后的变化越小，即在test image上的预测loss变化越小，模型就越难被training sets attacked。

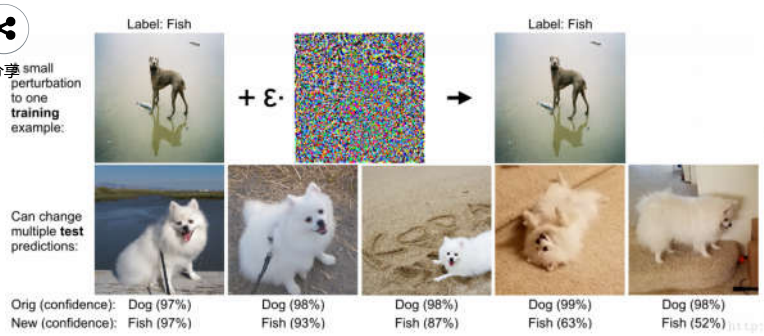

Figure 5. **Training-set attacks.** We targeted a set of 30 test images featuring the first author's dog in a variety of poses and backgrounds. By maximizing the average loss over these 30 images, we created a visually-imperceptible change to the particular training image (shown on top) that flipped predictions on 16 test images.

3.检查领域误匹配

Domain mismatch：训练样本与测试样本不同分布时，会造成训练loss小，测试loss大。我们可以通过influence function找到使得test error最大的training example。
分别把每个training example删去，计算$-\mathcal{I}_{\text{up,loss}}(z_i, z_{\text{test}})$，若计算结果为正，则次training example对此test example为helpful；否则harmful。

4.修正错误标签

关键是找出对模型产生最大influence$-\mathcal{I}_{\text{up,loss}}(z_i, z_{\text{test}})$（可能为正 helpful，可能为负harmful）的训练样本。由于我们没有测试样本，就计算$\mathcal{I}_{\text{up,loss}}(z_i, z_i)$,估计出将第i个训练样本z_i删去后，在z_i上的预测loss的变化。即用这个z_i来当测试样本。

Black Box (/junjie435/article/details/39934535)

Black Box Time Limit: 1000MS Memory Limit: 10000K Total Submissions: 7592 Acce...

junjie435 (http://blog.csdn.net/junjie435) 2014-10-09 17:28 287

Visualizing and Understanding Convolutional Networks论文笔记 (/bailufeiyan/article/details/50575343)

本文是Matthew D.Zeiler 和Rob Fergus 13年的论文，主要通过Deconvnet（反卷积）来可视化卷积网络，来理解卷积网络并进行分析和调优；本文通过反卷积，将Alex-net...

bailufeiyan (http://blog.csdn.net/bailufeiyan) 2016-01-24 20:51 588

如何成为一名机器学习的大咖？ (http://edu.csdn.net/huiyiCourse/series_detail/61?utm_source=blog10)

对于机器学习，很多人的观点是：机器学习技术是今后所有技术人员都绕不过的一个门槛。 那么，普通程序员该学习机器学作为一名对机器学习心有向往的程序员，我该以什么样的姿势开始呢？

Understanding the SIP Via Header (/hzhsan/article/details/47809083)

Understanding the SIP Via Header

hzhsan (http://blog.csdn.net/hzhsan) 2015-08-20 15:58 504

神经性风格化过程的特征控制 (/github_39502869/article/details/75127742)

Controlling Perceptual Factors in Neural Style Transfer相当于在第一篇上的扩展，是一种对风格化方法的优化。介绍对神经元风格化方法的概括：分别从两张...

github_39502869 (http://blog.csdn.net/github_39502869) 2017-07-14 18:59 39

论文笔记之：Let there be Color!: Joint End-to-end Learning of Global and Local Image Priors for Automatic (/qq_20513495/article/details/51635635)

前言：不知道你是否被这张实验效果图所震撼？Yes, I do. 那么他是怎么做到的呢？本文提出了一种给灰度图像自动上色的框架，结合了图像的局部和全局先验知识 (both global priors a...

qq_20513495 (http://blog.csdn.net/qq_20513495) 2016-06-11 10:33 1022

The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation_2016 (/sunyao_123/article/details/75449116)

作者：Simon Jégou Michal Drozdzal David VazquezAbstract 典型的语义分割架构由（a）一条下采样路径，负责提取粗的语义特征，再加一个（b）一...

sunyao_123 (http://blog.csdn.net/sunyao_123) 2017-07-19 20:31 191

论文笔记---小论文 (/skewrain/article/details/20046403)

初稿4月初。我们主要关注云应用和云工程方向。工具的类别---》归类、提炼。关于工具：前面主要写：原理、综述总结。后面主要写：给出一种迁移方法。关于云迁移工具：我们主要强调：从物理环境向云端...

SkewRain (http://blog.csdn.net/SkewRain) 2014-02-27 14:10 563

White matter hyperintensity and stroke lesion segmentation and differentiation using cnn_part1_2017 (/sunyao_123/article/details/76707380)

作者：R. Guerrero ⁎1 , C. Qin ⁎1 , O. Oktay 1 Abstract 最近，一些方法被提出来自动分割WMH区域。这些方法大多分割

MRI图像的WHM，但是不能区分WM...

sunyao_123 (http://blog.csdn.net/sunyao_123) 2017-08-05 11:01 87

White matter hyperintensity and stroke lesion segmentation and differentiation using cnn_part2_2017 (/sunyao_123/article/details/76944034)

4.2 Comparison to state-of-the-art 作者和当前最好的方法进行了比较，LST (http://www.statistical-modelling.de/lst....

sunyao_123 (http://blog.csdn.net/sunyao_123) 2017-08-08 21:03 27

The Importance of Skip Connections in Biomedical Image segmentation_2016 (/sunyao_123/article/details/75950330)

作者：Michal Drozdzal， Eugene Vorontsov， Gabriel Chartrand， Samuel Kadoury and Chris Pal Abstract论...

sunyao_123 (http://blog.csdn.net/sunyao_123) 2017-07-24 09:06 64

论文笔记：[ACL2016]End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF (/youngdreamnju/article/details/54346658)

文章：Ma X, Hovy E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF[J]. 2016. 发表在ACL2016...

YoungDreamNJU (http://blog.csdn.net/YoungDreamNJU) 2017-01-11 19:07 868

POJ1442——Black Box (/smile_kai/article/details/52199231)

Black Box Time Limit: 1000MS Memory Limit: 10000K Total Submissions: 10621 Acc...

smile_kai (http://blog.csdn.net/smile_kai) 2016-08-13 14:46 90

POJ 1442 Black Box (/liang654213/article/details/52166286)

Black Box Time Limit: 1000MS Memory Limit: 10000K Total Submissions: 10302 Acc...

liang654213 (http://blog.csdn.net/liang654213) 2016-08-09 21:26 35

OSWatcher Black Box (/suyishuai/article/details/32939225)

很多时候，我们要监视Orac

suyishuai (http://blog.csdn.net/suyishuai) 2014-06-21 17:46 1020

POJ 1442 Black Box (/houserabbit/article/details/38071187)

题意：给你个序列和一串询问 询问前a[i]个数字第i小的是几思路：动态的第k值问题 由于区间只增不减所以是水题利用平衡树解决这类问题 treap是方便编写的类似平衡树的产...

u013351160 (http://blog.csdn.net/u013351160) 2014-07-23 21:55 689

POJ 1442 Black Box (/sssogs/article/details/8731028)

题意：给你n个数的数列，然后再给你m次查询，第i (i 题解：根据Discuss说的，这题可以用各种数据结构AC...为了练习一下红黑树，我写了个红黑树求解的。240行代码，结果居然1Y了,,,这.....

sssogs (http://blog.csdn.net/sssogs) 2013-03-28 13:38 1420

poj 1442 Black Box (/zhangxiaoxiang123/article/details/47787409)

Black Box Time Limit: 1000MS Memory Limit: 10000K Total Submissions: 8847 Acce...

zhangxiaoxiang123 (http://blog.csdn.net/zhangxiaoxiang123) 2015-08-19 20:41 454

POJ 1442 - Black Box (/sureina/article/details/52432045)

Description Our Black Box represents a primitive database. It can save an integer array and has a s...

Sureina (http://blog.csdn.net/Sureina) 2016-09-04 13:52 56

POJ-1442-Black Box (/fengkuangdewoniudada/article/details/53117635)

Black Box Time Limit: 1000MS Memory Limit: 10000K Total Submissions: 11225 Acc...

fengkuangdewoniudada (http://blog.csdn.net/fengkuangdewoniudada) 2016-11-10 17:22 189

POJ 1442 Black Box (/q779160073/article/details/47101999)

Black Box Time Limit: 1000MS Memory Limit: 10000K Total Submissions: 8627 Acce...

q779160073 (http://blog.csdn.net/q779160073) 2015-07-28 10:47 159