黑客派

为什么分类问题用 cross entropy，而回归问题用 MSE

作者：moloee

原文链接：https://hacpai.com/article/1516696419994

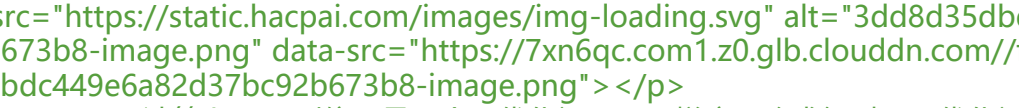
来源网站：黑客派

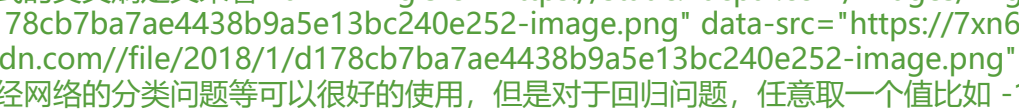
<h3 id="先说为什么分类问题用交叉熵">先说为什么分类问题用交叉熵</h3>

<h4 id="为啥不用-classification-error--">为啥不用 classification error ?</h4>
来两个模型
模型 1
<table>
 <thead>
  <tr>
   <th>COMPUTED</th>
   <th>TARGETS</th>
   <th>CORRECT?</th>
  </tr>
 </thead>
 <tbody>
  <tr>
   <td>0.3 0.3 0.4</td>
   <td>0 0 1 (democrat)</td>
   <td>yes</td>
  </tr>
  <tr>
   <td>0.3 0.4 0.3</td>
   <td>0 1 0 (republican)</td>
   <td>yes</td>
  </tr>
  <tr>
   <td>0.1 0.2 0.7</td>
   <td>1 0 0 (other)</td>
   <td>no</td>
  </tr>
 </tbody>
</table>
模型 2
<table>
 <thead>
  <tr>
   <th>COMPUTED</th>
   <th>TARGETS</th>
   <th>CORRECT?</th>
  </tr>
 </thead>
 <tbody>
  <tr>
   <td>0.1 0.2 0.7</td>
   <td>0 0 1 (democrat)</td>
   <td>yes</td>
  </tr>

<tr>
<td>0.1 0.7 0.2</td>
<td>0 1 0 (republican)</td>
<td>yes</td>
</tr>
<tr>
<td>0.3 0.4 0.3</td>
<td>1 0 0 (other)</td>
<td>no</td>
</tr>
</tbody>
</table>

 2 个模型的 classification error 相等，但模型 2 要明显优于模型 1.<code>classification error 很难精确描述模型与理想模型之间的距离。</code>
 计算交叉熵，用交叉熵则可以更准确的衡量模型之间的差距

<h4 id="为啥不用-MSE--">为啥不用 MSE ?</h4>
总的来说，分类问题需要用 one hot 的形式计算个 label 的概率，然后用 argmax 来决定分类
计算概率的时候通常用 softmax。参考流程：计算 loss-> 计算 softmax->argmax。
用 MSE 计算 loss 的问题在于，通过 Softmax 输出的曲线是波动的，有很多局部的极值点。 即
非凸优化问题 (non-convex)，如下图

2d37bc92b673b8-image.png" data-src="https://7xn6qc.com1.z0.glb.clouddn.com//file/2018/
/3dd8d35dbdc449e6a82d37bc92b673b8-image.png">
而 cross entropy 计算 loss，则依旧是一个凸优化问题，用梯度下降求解时，凸优化问题有很好
收敛特性。
<a href="https://link.hacpai.com/forward?goto=http%3A%2F%2Fjackon.me%2Fposts%2
why-use-cross-entropy-error-for-loss-function%2F" target="_blank" rel="nofollow ugc">参

<h3 id="为什么回归问题用MSE">为什么回归问题用 MSE</h3>
<h4 id="为啥不能用交叉熵-">为啥不能用交叉熵？ </h4>
拿二项式的交叉熵定义来看
 
vg" alt="d178cb7ba7ae4438b9a5e13bc240e252-image.png" data-src="https://7xn6qc.com1.
0.glb.clouddn.com//file/2018/1/d178cb7ba7ae4438b9a5e13bc240e252-image.png">
对于神经网络的分类问题等可以很好的使用，但是对于回归问题，任意取一个值比如 -1.5，就没
计算 log(-1.5)，所以一般不用交叉熵来优化回归问题。
<h4 id="为什么用MSE">为什么用 MSE</h4>
最小二乘是在欧氏距离为误差度量的情况下，由系数矩阵所张成的向量空
内对于观测向量的最佳逼近点。
为什么用欧式距离作为误差度量 （即 MSE），09 年 IEEE Signal Processing Magzine 的 《Me
n squared error: Love it or leave it?》这篇文章做了很好的讨论。链接：<a href="https://link.hac
ai.com/forward?goto=https%3A%2F%2Flink.zhihu.com%2F%3Ftarget%3Dhttp%253A%2F%2
www2.units.it%2Framponi%2Fteaching%2FDIP%2Fmateriale%2Fmse_bovik09.pdf" target="_b
ank" rel="nofollow ugc">文章
这篇文章在"WHY DO WE LOVE THE MSE?"中说，MSE：

 1. 它简单。
 2. 它提供了具有很好性质的相似度的度量。例如：
 1）它是非负的;
 2）唯一确定性。只有 x=y 的时候，d(x,y)=0；
 3）它是对称的，即 d(x,y)=d(y,x)；
 4）符合三角性质。即 d(x,z)<=d(x,y)+d(y,z).
 3. 物理性质明确，在不同的表示域变换后特性不变，例如帕萨瓦尔等式。

 4. 便于计算。通常所推导得到的问题是凸问题，具有对称性，可导性。通常具有解析解，外便于通过迭代的方式求解。
 5. 和统计和估计理论具有关联。在某些假设下，统计意义上是最优的。

然而，MSE 并非没有缺点。并不是所有的问题都可以套用该准则，在“IMLICIT ASSUMPTIONS WHEN USING THE MSE”说，它基于了以下几点对于信号的假设：

 1. 信号的保真度和该信号的空间和时间顺序无关。即，以同样的方法，改变两个待比较的号本身的空间或时间排列，它们之间的误差不变。例如，[1 2 3], [3 4 5]两组信号的 MSE 和[3 2 1],[5 3]的 MSE 一样。
 2. 误差信号和原信号无关。只要误差信号不变，无论原信号如何，MSE 均不变。例如，对固定误差[1 1 1]，无论加在[1 2 3]产生[2 3 4]还是加在[0 0 0]产生[1 1 1]，MSE 的计算结果不变。</>
 3. 信号的保真度和误差的符号无关。即对于信号[0 0 0]，与之相比较的两个信号[1 2 3]和[--2 -3]被认为和[0 0 0]具有同样的差别。
 4. 信号的不同采样点对于信号的保真度具有同样的重要性。

参考