交叉熵

覃雄派

交叉熵

提纲

- 分类模型的预测效果
  - 分类错误率
  - 均方误差
- 交叉熵损失函数的提出
  - 二值分类
  - 多类别分类
- 交叉熵的应用
- 信息熵、交叉熵、和KL散度的关系

交叉熵

- 分类模型的预测效果
  - 我们希望通过一个人的若干独立特征，包括年龄、性别、年收入等，来预测一个人的政治倾向
    - 有三种可能的预测结果，即民主党、共和党、和其他党
    - 假设我们当前有两个逻辑斯蒂回归模型(两个模型的参数不同)
    - 这两个模型都是通过sigmoid函数的方式，得到对于每个预测结果的概率值

模型1

模型2

交叉熵

- 分类模型的预测效果
  - 我们希望通过一个人的若干独立特征，包括年龄、性别、年收入等，来预测一个人的政治倾向
    - 有三种可能的预测结果，即民主党、共和党、和其他党

模型1

模型2

Computed	Targets	Correct?
0.3 0.3 **0.4**	001(民主党)	Yes
0.3 **0.4** 0.3	010(共和党)	Yes
0.1 0.2 **0.7**	100(其他党)	No

可见，模型1对于样本1和样本2以非常微弱的优势判断正确；对于样本3的判断则彻底错误

交叉熵

- 分类模型的预测效果
  - 我们希望通过一个人的若干独立特征，包括年龄、性别、年收入等，来预测一个人的政治倾向
    - 有三种可能的预测结果，即民主党、共和党、和其他党

模型1

模型2

Computed	Targets	Correct?
0.1 0.2 **0.7**	001(民主党)	Yes
0.1 **0.7** 0.2	010(共和党)	Yes
0.3 **0.4** 0.3	100(其他党)	No

可见，模型2对于样本1和样本2判断相当准确；对于样本3判断错误，但是相对于模型1来说好一点，对于实际值来说没有错得太离谱

交叉熵

交叉熵

- **回顾分类错误率Classification Error**
  - 分类错误率的定义为 Classification Error = (count of error items)/(count of all items)
    - 模型1的分类错误率=1/3
    - 模型2的分类错误率=1/3
    - 模型1和模型2的分类错误率是一样的，看不出差别来
  - 而通过上述分析，我们知道两者是有较大的差别的

- 模型1和模型2虽然都是预测错了1个样本，但是相对来说模型2表现得更好，损失函数值照理来说应该更小
- 所以把分类错误率作为损失函数不合适

Computed	Targets	Correct?
0.3 0.3 0.4	001(民主党)	Yes
0.3 0.4 0.3	010(共和党)	Yes
0.1 0.2 0.7	100(其他党)	No

Computed	Targets	Correct?
0.1 0.2 0.7	001(民主党)	Yes
0.1 0.7 0.2	010(共和党)	Yes
0.3 0.4 0.3	100(其他党)	No

交叉熵

- **回顾均方误差(Mean Squared Error)**
  - 均方误差的公式为MSE=$\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$
  - 模型1的均方误差计算如下：
    - Sample 1 loss = $(0.3-0)^2 + (0.3-0)^2 + (0.4-1)^2 = 0.54$;
    - Sample 2 loss = $(0.3-0)^2 + (0.4-1)^2 + (0.3-0)^2 = 0.54$;
    - Sample 3 loss = $(0.1-1)^2 + (0.2-0)^2 + (0.7-0)^2 = 1.34$;
    - 对所有样本的loss求平均，有MSE = (0.54+0.54+1.34)/3=0.807
  - 同理，对于模型2，其均方误差计算如下：
    - Sample 1 loss = $(0.1-0)^2 + (0.2-0)^2 + (0.7-1)^2 = 0.14$;
    - Sample 2 loss = $(0.1-0)^2 + (0.7-1)^2 + (0.2-0)^2 = 0.14$;
    - Sample 3 loss = $(0.3-1)^2 + (0.4-0)^2 + (0.3-0)^2 = 0.74$;
    - 对所有样本的loss求平均，有MSE = (0.14+0.14+0.74)/3=0.34
  - 可以看出，模型2的MSE loss优于模型1
  - 可以MSE作为损失函数

- 如果逻辑斯蒂回归配合MSE损失函数，采用梯度下降法进行学习的时候，刚开始训练时会出现学习速率非常慢的情况
- 对于分类问题来讲，分类错误率和均方误差都不是好的损失函数

Computed	Targets	Correct?
0.3 0.3 0.4	001(民主党)	Yes
0.3 0.4 0.3	010(共和党)	Yes
0.1 0.2 0.7	100(其他党)	No

Computed	Targets	Correct?
0.1 0.2 0.7	001(民主党)	Yes
0.1 0.7 0.2	010(共和党)	Yes
0.3 0.4 0.3	100(其他党)	No

交叉熵

- 交叉熵
  - 一般地，在线性回归问题中，使用MSE(Mean Squared Error)作为loss函数；而在分类问题中常常使用交叉熵作为loss函数
  - 交叉熵能够衡量同一个随机变量的两个不同概率分布的差异程度，在机器学习中一般表示为真实概率分布与预测概率分布之间的差异
    - 交叉熵的值越小，模型预测效果就越好
  - 看个例子

*	猫	狗	马
Label	0	1	0
Predict	0.2	0.7	0.1

$$Loss = -(0*\log(0.2)+1*\log(0.7)+0*\log(0.1))=0.515$$

交叉熵在分类问题中常常与Softmax一起使用，Softmax将输出的结果进行处理，使其多个分类的预测值(概率)的和为1，再通过交叉熵来计算损失

交叉熵

交叉熵

- 二值分类
  - 在二值分类问题里，模型最后需要预测的结果有两种情况，对于每个类别我们预测得到的概率为p或者1-p
  - 交叉熵损失函数的形式如下：
    - $\mathrm{L} = \frac{1}{N}\sum_{i=1}^{N} -[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)]$
    - y_i表示样本i的label，正例为1，负例为0；p_i表示样本i预测为正的概率

交叉熵

- 二值分类
  - 在二值分类问题里，模型最后需要预测的结果有两种情况，对于每个类别我们预测得到的概率为p或者1-p
  - 交叉熵损失函数的形式如下：
    - $\mathrm{L} = \frac{1}{N}\sum_{i=1}^{N} -[y_i \log(p_i) + (1-y_i)\log(1-p_i)]$
    - y_i表示样本i的label，正例为1，负例为0；p_i表示样本i预测为正的概率

看一个构造的实例

	B	C	D	E	F	G	H	I	J
		真实分布			模型1预测效果			模型2预测效果	
	样本1	0	1		0.4	0.6		0.2	0.8
	样本2	1	0		0.6	0.4		0.8	0.2
				交叉熵	0.736966		交叉熵	0.321928	

模型1
Sample 1 loss = -(0*log0.4 +1*log0.6) = 0.737；
Sample 2 loss = -(1*log0.6 +0*log0.4) = 0.737；
对所有样本的loss求平均，得到L=(0.737+0.737)/2=0.737

模型2
Sample 1 loss = -(0*log0.2 +1*log0.8) = 0.322；
Sample 2 loss = -(1*log0.8 +0*log0.2) = 0.322；
对所有样本的loss求平均，得到L=(0.322+0.322)/2=0.322

交叉熵

- 多类别分类
  - 多类别分类是对二值分类的扩展，损失函数形式如下：
  - $\mathrm{L} = \frac{1}{N}\sum_{i=1}^{N} -[\sum_{c=1}^{M} y_{ic} \log(p_{ic})]$
    - N为样本数量
    - M为类别的数量；y_{ic}为指示变量，如果样本i的类别和类别c相同，y_{ic}为1，否则为0；p_{ic}表示对于样本i预测为类别c的预测概率

针对前文的例子，算一算试试看
(下页继续)

交叉熵

- 多类别分类
  - 多类别分类是对二值分类的扩展，损失函数形式如下：
  - $L = \frac{1}{N}\sum_{i=1}^{N} -[\sum_{c=1}^{M} y_{ic} \log(p_{ic})]$
  - 对于模型1，我们计算交叉熵如下：
    - Sample 1 loss = -(0*log0.3 +0*log0.3+1*log0.4) = 1.322;
    - Sample 2 loss = -(0*log0.3 +1*log0.4+0*log0.3) = 1.322;
    - Sample 3 loss = -(1*log0.1 +0*log0.2+0*log0.7) = 3.322;
    - 对所有样本的loss求平均，得到L=(1.322+1.322+3.322)/3=**1.989**
  - 对于模型2，我们计算交叉熵如下：
    - Sample 1 loss = -(0*log0.1 +0*log0.2+1*log0.7) = 0.515;
    - Sample 2 loss = -(0*log0.1 +1*log0.7+0*log0.2) = 0.515;
    - Sample 3 loss = -(1*log0.3 +0*log0.4+0*log0.4) = 1.737;
    - 对所有样本的loss求平均，得到L=(0.515+0.515+1.737)/3=**0.922**

交叉熵损失函数可以捕抓到模型1和模型2的预测效果的差异，**模型2的预测效果要好**

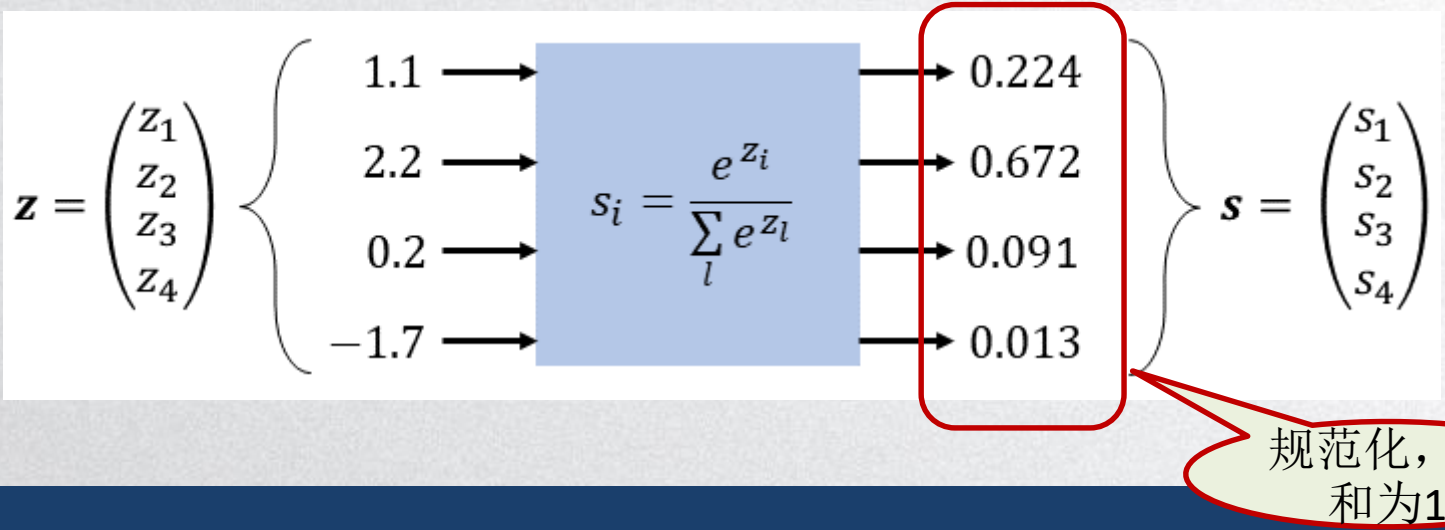
Computed	Targets	Correct?
0.3 0.3 **0.4**	001(民主党)	Yes
0.3 **0.4** 0.3	010(共和党)	Yes
0.1 0.2 **0.7**	100(其他党)	No

Computed	Targets	Correct?
0.1 0.2 **0.7**	001(民主党)	Yes
0.1 **0.7** 0.2	010(共和党)	Yes
0.3 **0.4** 0.3	100(其他党)	No

交叉熵

交叉熵

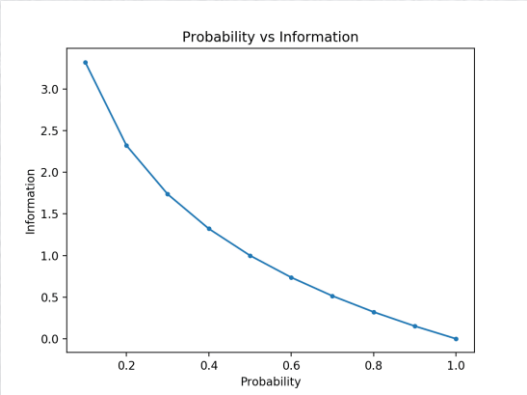
- 交叉熵的应用
  - 交叉熵损失函数经常用于分类问题中，特别是在神经网络做分类问题时，经常使用交叉熵作为损失函数
  - 此外，由于交叉熵涉及到计算每个类别的概率，所以交叉熵一般都和sigmoid(或Softmax)函数一起出现



交叉熵

交叉熵

- 信息熵、交叉熵、和KL散度
- **信息量**
  - 一个事件x的信息量，可以用公式 $I(x) = -\log(p(x))$进行计算
  - 这个公式表达的意思是
    - 一个事件发生的概率越大，那么信息量就越小
    - 如果事件发生的概率是1，即事件100%发生，那么信息量为0



交叉熵

- 信息熵、交叉熵、和KL散度
- **熵**
  - 熵是对信息量求期望值(平均值)，计算公式为
    - $H(x) = E(I(x)) = -\sum_{x \in X} p(x) \log p(x)$
  - 假设有一个考生参加了10次考试，有9次不及格，1次及格，x=A表示及格事件
    - 那么这个事件的熵，通过上述公式进行计算
    - 有$H_A(x) = -(0.9 * \log 0.9 + 0.1 * \log 0.1) = 0.469$
    - 注意这里的log是以2作为底数的

期望值就是平均值

	x1	x2	…
概率	P(x1)	P(x2)	…
信息量	$-\log p(x1)$	$-\log p(x2)$	…

	语文	数学
概率	80	90
分数	1/2	1/2

按列乘，按行累加

交叉熵

- 信息熵、交叉熵、和KL散度
- **交叉熵**
  - 假设有两个分布，那么它们在给定样本集上的交叉熵(Cross Entropy)定义如下:
  - $\mathrm{CrossEntropy}(p, q) = -\sum_{x \in X} p(x) \log q(x)$

	x1	x2	…
p	P(x1)	P(x2)	…
q	q(x1)	q(x2)	…

交叉熵

- 信息熵、交叉熵、和KL散度
- **相对熵(KL散度)**
  - 相对熵(Relative Entropy)即KL散度(Kullback-Leibler Divergence)，也称为KL距离，是两个随机分布之间的距离的度量，记为$D_{KL}(p||q)$。其计算公式如下：
  - $D_{KL}(p||q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} = \sum_{x \in X} p(x) \log p(x) - \sum_{x \in X} p(x) \log q(x) = -H(p) - \sum_{x \in X} p(x) \log q(x)$
    - 当p==q的时候，KL散度为0
    - 注意，KL散度是非对称的，即$D_{KL}(p||q) \neq D_{KL}(q||p)$

交叉熵

- 信息熵、交叉熵、和KL散度
- **相对熵(KL散度)**
  - 相对熵(Relative Entropy)即KL散度(Kullback-Leibler Divergence)，也称为KL距离，是两个随机分布之间的距离的度量，记为$D_{KL}(p||q)$。其计算公式如下：
  - $D_{KL}(p||q) = \sum_{x\in X} p(x)\log\frac{p(x)}{q(x)} = \sum_{x\in X} p(x)\log p(x) - \sum_{x\in X} p(x)\log q(x) = -H(p) - \sum_{x\in X} p(x)\log q(x)$
    - 当p==q的时候，KL散度为0
    - 注意，KL散度是非对称的，即$D_{KL}(p||q) \neq D_{KL}(q||p)$

KL散度 = - 信息熵 + 交叉熵=交叉熵-信息熵

交叉熵

- 信息熵、交叉熵、和KL散度
- **相对熵(KL散度)**
  - 相对熵(Relative Entropy)即KL散度(Kullback-Leibler Divergence)，也称为KL距离
  - 看看一个简单的KL散度的例子
    - 假设P和Q分别为离散的分布。具体为P{ x1:0.75, x2: 0.25}，Q{ x1:0.50, x2: 0.50}。
    - $D_{KL}(p||q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} = 0.75 log \frac{0.75}{0.5} + 0.25 log \frac{0.25}{0.5} = 0.1887$
    - $D_{KL}(q||p) = \sum_{x \in X} q(x) \log \frac{q(x)}{p(x)} = 0.50 log \frac{0.50}{0.75} + 0.50 log \frac{0.50}{0.25} = 0.2075$
    - 可以看到$D_{KL}(p||q) \neq D_{KL}(q||p)$