链滴

李航《统计学习方法》第六章逻辑回归 python 实现（蘑菇分类数据集）

作者：hailangjiang

原文链接：https://ld246.com/article/1587100745226

来源网站：链滴

1.逻辑回归简介

逻辑回归名为回归实际上是分类算法，通常我们所说的是二项逻辑斯蒂回归简称逻辑回归。逻辑回归条件概率分布：$P(Y=1|x)=\frac{exp(wx+b)}{1+exp(wx+b)}$,即在给定x的条件下Y=1的概率。标记$P(Y=1|x)=\pi(x),P(Y=0|x)=1-\pi(x)$,我们希望最大化如下概率即似然函数$\prod_{i=1}^N[\pi(x_i)]^{y_i}[1-\pi(x)]^{1-yi}$，为方便计算对似然函数取对数得$L(w)=\sum_{i=1}^N[y_ilog\pi(x_i)+(1-y_i)log(1-pi(x_i)]$，可以看出对L(W)取负号就是交叉熵损失函数，通过梯度下降极小化损失函数或梯度上升极大对数似然函数就可以得到w,b的极大似然估计值$\tilde{w},\tilde{b}$,则预测阶段有$P(Y=1|x)=\frac{exp(\tilde{w}+\tilde{b})}{1+exp(\tilde{w}+\tilde{b})}$。

2.实验

本文依旧采用蘑菇分类数据集进行建模，对数据集随机七三划分为训练集、测试集，学习率设为0.1，迭代次数20000次，最后测试集正确率为98.7。因为kaggle蘑菇分类数据集其实只给出了训练集，这人为对训练集划分，但问题不大，关键是了解学习算法。

```
Iterations:2000,Loss:0.258565
Iterations:4000,Loss:0.200004
Iterations:6000,Loss:0.172559
Iterations:8000,Loss:0.155533
Iterations:10000,Loss:0.143558
Iterations:12000,Loss:0.134504
Iterations:14000,Loss:0.127326
Iterations:16000,Loss:0.121440
Iterations:18000,Loss:0.116491
Iterations:20000,Loss:0.112247
测试集正确率: 98.666667 %
```

原文链接：李航《统计学习方法》第六章逻辑回归 python 实现（蘑菇分类数据集）