小作业3

1、试比较LDA、logistic回归、softmax回归三种模型在学习任务、损失函数两方面的异同。（4分）

LDA是一种线性分类模型，同时是一种监督式降维方法。它将给定训练集的样本投影到最佳鉴别矢量空间，使得同类样本的投影点尽可能接近（类内协方差尽可能小），异类样本的投影点尽可能远离（类间距离尽可能大）。对二分类问题，须最大化的目标函数是类内散度矩阵$\boldsymbol{S}_{\mathrm{w}}$和类间散度矩阵$\boldsymbol{S}_{\mathrm{b}}$的“广义瑞利商”，即 $J=\frac{\boldsymbol{w}^T\boldsymbol{S}_{\mathrm{b}}\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}}=\frac{\boldsymbol{w}^T(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)^T\boldsymbol{w}}{\boldsymbol{w}^T(\boldsymbol{\Sigma}_0+\boldsymbol{\Sigma}_1)\boldsymbol{w}}$。

logistic回归模型将线性回归推广到分类问题，用对数几率函数 $y=\frac{1}{1+e^{-z}}$ 将线性回归预测值映射为$[0,1]$内的实值，输出样本为正标记的概率$h_\theta(\boldsymbol{x})=\frac{1}{1+e^{-\theta^T x}}$。对二分类问题，最小化损失函数：$J(\boldsymbol{\theta})=-\sum_{i=1}^{N}\left[(y^i log\left(h_\theta(\boldsymbol{x}^i)\right)+(1-y^i)log\left(1-h_\theta(\boldsymbol{x}^i)\right)\right]$，根据最大对数似然导出。

softmax 回 归 是 将 二 分 类 逻 辑 回 归 推 广 到 多 分 类 问 题 ， 用 归 一 化 指 数 函 数 $[\sigma(\boldsymbol{z})]_j=\frac{e^{z_j}}{\sum_{l=1}^{K}e^{z_l}}$，$j=1,2,\cdots,K$表示类别，输出样本在每一种分类结果上的概率。最小化损失函数：$L(\theta)=-\sum_{i=1}^{N}\left[\sum_{j=1}^{K}\delta\{y^i=j\}\log\frac{e^{\theta_j^T x^i}}{\sum_{l=1}^{K}e^{\theta_l^T x^i}}\right]$，$\boldsymbol{X}^i=[\boldsymbol{x}^i;1]\in\mathbb{R}^{(d+1)\times 1},\boldsymbol{\theta}\in\mathbb{R}^{K\times(d+1)}$，根据最大对数似然导出。

综上

(1)三种模型都是可用于分类的学习模型，LDA还是常用的监督式降维方法。

(2)LDA的损失函数为样本类间、类内散度矩阵的“广义瑞利商”，logistic回归和softmax回归都是由最大对数似然导出损失函数，似然函数为所有训练样本预测正确的概率。

小作业3

2、 若有一个点能被正确分类且远离决策边界。如果将该点加入到训练集，SVM的决策是否会受到影响，如果采用 logistic 回归进行决策会受到影响吗？为什么？（2分）

SVM不受到影响，因为SVM的结果仅与支持向量相关，其他样本的权重为0；而logistic回归相对来说会受到些许影响，因为它的损失函数遍历所有样本点。

小作业3

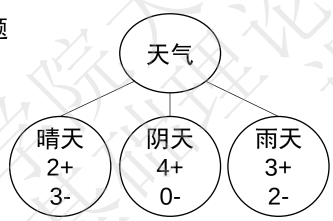
3、对于表中的数据，请基于信息增益的方法判断属性天气和风力，哪一个应作为决策树的根节点。（2分）

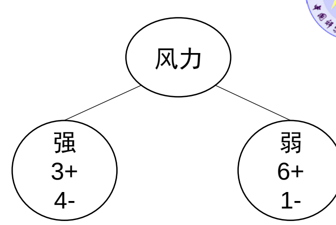
4、对于表中的数据，给一个新实例{天气是阴天, 温度高, 湿度高, 风速强}，请基于朴素贝叶斯方法决策是否去打球，并考虑是否平滑训练样本。（2分）

日期	天气	温度	湿度	风速	是否打球
1	晴天	高	高	弱	否
2	晴天	高	高	强	否
3	阴天	高	高	弱	是
4	雨天	中	高	弱	是
5	雨天	低	中	弱	是
6	雨天	低	中	强	否
7	阴天	低	中	强	是
8	晴天	中	高	强	否
9	晴天	低	中	弱	是
10	雨天	中	中	弱	是
11	晴天	中	中	强	是
12	阴天	中	高	强	是
13	阴天	高	中	弱	是
14	雨天	中	高	强	否

小作业3

第3题





$H(\text{晴天}) = -0.4\log_2(0.4) - 0.6\log_2(0.6) = 0.9710$

$H(\text{阴天}) = -1\log_2(1) - 0 = 0$

$H(\text{雨天}) = -0.6\log_2(0.6) - 0.4\log_2(0.4) = 0.9710$

$H(\text{强}) = 0.9852$

$H(\text{弱}) = 0.5917$

$$\frac{5}{14} * 0.9710 + \frac{4}{14} * 0 + \frac{5}{14} * 0.9710 = 0.6936$$

$$\frac{7}{14} * 0.9852 + \frac{7}{14} * 0.5917 = 0.7884$$

$$H(\text{原始}) - 0.6936 > H(\text{原始}) - 0.7884$$

天气的信息增益大

天气属性作为根节点

4

小作业3

第4题

若 $x=$(阴天，高温，湿度中，风速强)

先验概率：

$$P(y_1)=9/14,\ P(y_2)=5/14$$

条件概率：

天气	y_1	y_2	$P(x_1\|y_1)$	$P(x_1\|y_2)$
晴天	2	3	2/9	3/5
阴天	4	0	4/9	0
雨天	3	2	3/9	2/5
总计	9	5	100%	100%

温度	y_1	y_2	$P(x_2\|y_1)$	$P(x_2\|y_2)$
高	2	2	2/9	2/5
中	4	2	4/9	2/5
低	3	1	3/9	1/5
总计	9	5	100%	100%

湿度	y_1	y_2	$P(x_3\|y_1)$	$P(x_3\|y_2)$
高	3	4	3/9	4/5
中	6	1	6/9	1/5
总计	9	5	100%	100%

风力	y_1	y_2	$P(x_4\|y_1)$	$P(x_4\|y_2)$
强	3	4	3/9	4/5
弱	6	1	6/9	1/5
总计	9	5	100%	100%

$$P(y_1)\prod_{j=1}^{n}P(x_j|y_1)=\frac{9}{14}\times\frac{4}{9}\times\frac{2}{9}\times\frac{6}{9}\times\frac{3}{9}=0.0141\qquad P(y_2)\prod_{j=1}^{n}P(x_j|y_2)=\frac{5}{14}\times\frac{0}{5}\times\frac{2}{5}\times\frac{1}{5}\times\frac{4}{5}=0$$

去打球？

小作业3

第4题

存在概率为0的条件概率，会直接抹灭其他的信息，所以进行拉普拉斯平滑，平滑后

天气	y_1	y_2	$P(x_1\|y_1)$	$P(x_1\|y_2)$
晴天	2+1	3+1	3/12	4/8
阴天	4+1	0+1	5/12	1/8
雨天	3+1	2+1	4/12	3/8
总计	9+3	5+3	100%	100%

湿度	y_1	y_2	$P(x_3\|y_1)$	$P(x_3\|y_2)$
高	3+1	4+1	4/11	5/7
中	6+1	1+1	7/11	2/7
总计	9+2	5+2	100%	100%

先验概率和温度风力属性做同样平滑

$$\hat{P}(y_1)\prod_{j=1}^{4}\hat{P}(x_j|y_1)=\frac{9+1}{14+2}\times\frac{4+1}{9+3}\times\frac{2+1}{9+3}\times\frac{6+1}{9+2}\times\frac{3+1}{9+2}=0.0151$$

√ 去打球

$$\hat{P}(y_2)\prod_{j=1}^{4}\hat{P}(x_j|y_2)=\frac{5+1}{14+2}\times\frac{0+1}{5+3}\times\frac{2+1}{5+3}\times\frac{1+1}{5+2}\times\frac{4+1}{5+2}=0.0036$$