作业理论题T1

2019年10月22日 16:05

T1：对于 $y=e^{w^Tx+b}$，$w^T=[\cdots 1]$，$x=\begin{bmatrix}\vdots\\ \vdots\end{bmatrix}$，

$\because y=g(w^Tx+b)$ 其中 $g(t)=e^t$，$g(t)$单调可逆

$\therefore \tilde{y}=g^{-1}(y)=w^Tx+b$，可用最小二乘法求解 w,b。设 $f(x)=\tilde{y}$

构建均方误差损失函数：

$$J(\hat{w})=\sum_{i=1}^{n}(f(x_i)-y_i)^2=(f-y)^T(f-y)$$

$$=(X\hat{w}-y)^T(X\hat{w}-y)$$

其中：$\hat{w}=\begin{bmatrix}w\\ b\end{bmatrix}$ $\hat{x}=\begin{bmatrix}x\\ 1\end{bmatrix}$

$$y=\begin{bmatrix}g^{-1}(y_1)\\ g^{-1}(y_2)\\ \vdots\\ g^{-1}(y_n)\end{bmatrix}=\begin{bmatrix}\ln y_1\\ \ln y_2\\ \vdots\\ \ln y_n\end{bmatrix}$$

设 $X=\begin{bmatrix}\hat{x}_1^T\\ \hat{x}_2^T\\ \vdots\\ \hat{x}_n^T\end{bmatrix}=\begin{bmatrix}x_1^T & 1\\ x_2^T & 1\\ \vdots & \vdots\\ x_n^T & 1\end{bmatrix}$

$$f=\begin{bmatrix}w^Tx_1+b\\ w^Tx_2+b\\ \vdots\\ w^Tx_n+b\end{bmatrix}$$

$$\therefore \frac{\partial J}{\partial \hat{w}}=2X^T(X\hat{w}-y)=0$$

$$X^TX\hat{w}-X^Ty=0$$

当 X 是满秩时，X^TX 可逆，此时有唯一解

$$\hat{w}^*=(X^TX)^{-1}X^Ty$$

作业理论题T2，T3

2019年10月22日 16:43

T2：产品为次品的概率：$P(\text{次品}) = P(\text{次品}|A)\times P(A) + P(\text{次品}|B)\times P(B) + P(\text{次品}|C)\times P(C)$

$= 0.015\times 0.35 + 0.010\times 0.35 + 0.020\times 0.30$

$= 0.01475$

产品为次品，来自A、B、C的概率：

$$P(A|\text{次品}) = \frac{P(\text{次品}|A)\times P(A)}{P(\text{次品})} = \frac{0.015\times 0.35}{0.01475} = 0.356$$

$$P(B|\text{次品}) = \frac{P(\text{次品}|B)\times P(B)}{P(\text{次品})} = \frac{0.010\times 0.35}{0.01475} = 0.237$$

$$P(C|\text{次品}) = \frac{P(\text{次品}|C)\times P(C)}{P(\text{次品})} = \frac{0.020\times 0.30}{0.01475} = 0.407$$

T3：支持向量机中松弛变量ξ不同取值范围对应的含义：

ξ_i表示第i个样本不满足分类约束条件的程度，其中

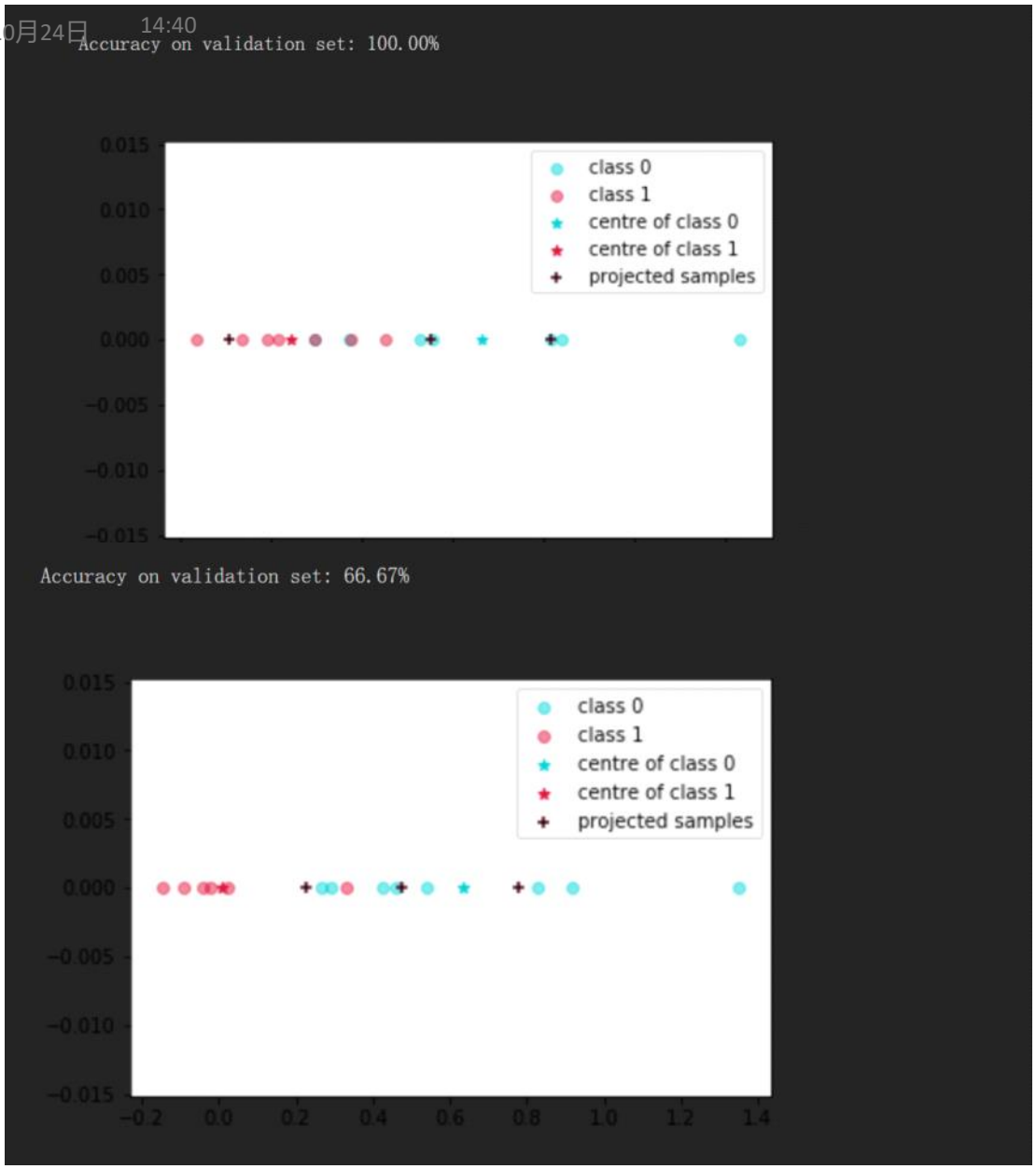
$\xi_i < 1$ 表示样本i分类正确，且$|\xi_i|$越大，正确裕量越大

$\xi_i > 1$ 表示样本i分类错误，且$|\xi_i|$越大，错误裕量越大。

实践题结果与分析T1：LDA

2019年10月24日 14:40

T1：



西瓜数据集中包含离散的特征，如西瓜颜色分为青绿、乌黑、浅白三种，不能用0和1表示。在调查资料后得知，用one-hot 独热编码，对数据集进行预处理，可以避免由于数据离散导致同一属性的不同特征的间距不一致，使得分类更加科学。因此我用独热编码处理了西瓜数据集。

LDA算法原理为将高维数据投影到直线上，使同类数据间尽可能接近，异类数据尽可能相互远离。为了直观地看出分类的结果，我将投影后的数据点、数据中心等信息绘制在数轴上，可以直观地看出分类的好坏。具体图标含义见右上角图释。

我对数据进行了随机乱序处理，因此每次训练集与验证集的数据都有差异。因此每次分类的结果也有所区别。由于西瓜数据集数据过少，导致分类结果并不稳定，容易受到验证集与训练集选取的影响，最终的正确率在60%~100%。

实践题结果与分析T2：Naïve Bayes

2019年10月24日 14:40

T2：

```
Bayes_watermelon = Naive_Bayes(dataset)

print('Accuracy with Cross Validation:',"%.2f%%"  %(Bayes_watermelon.Cross_validation()*100))
```

```
Accuracy with Cross Validation: 66.67%
```

西瓜数据集中包含离散的特征，如西瓜颜色分为青绿、乌黑、浅白三种，不能用0和1表示。在调查资料后得知，用one-hot 独热编码，对数据集进行预处理，可以避免由于数据离散导致同一属性的不同特征的间距不一致，使得分类更加科学。因此我用独热编码处理了西瓜数据集。

而对于连续特征，在此处为含糖量与密度两个属性，我用极大似然估计法对含糖量与密度属性进行了回归分析，认为它们的分布各自满足正态分布。在估计出正态分布的参数后，就可以得到两个属性的概率分布，可用于朴素贝叶斯的分类中。

由于西瓜数据集数据过少，导致分类结果并不稳定，容易受到验证集与训练集选取的影响，这在朴素贝叶斯分类器上表现非常明显。最终的正确率较差。

在此处应用了交叉验证的方法，具体为，将数据分为5组，每次用其中一组作为测试集，不参与训练，而其他的4个组作为该测试集的分类标准。

实践题结果与分析T3：SVM

2019年10月24日 14:40

T3：

```
SVM_model = My_SVM(dataset)
SVM_model.comparasion()
```

```
Accuracy with Kernel:linear: 66.67%
Accuracy with Kernel:poly: 100.00%
Accuracy with Kernel:rbf: 100.00%
Accuracy with Kernel:sigmoid: 66.67%
```

西瓜数据集中包含离散的特征，如西瓜颜色分为青绿、乌黑、浅白三种，不能用0和1表示。在调查资料后得知，用one-hot 独热编码，对数据集进行预处理，可以避免由于数据离散导致同一属性的不同特征的间距不一致，使得分类更加科学。因此我用独热编码处理了西瓜数据集。

在线性不可分的情况下，SVM首先在低维空间中完成计算，然后通过核函数将输入空间映射到高维特征空间，最终在高维特征空间中构造出最优分离超平面，从而把平面上本身不好分的非线性数据分开。

本程序对比了以下几种核函数：

linear kernel：应用于线性可分且特征数量多时；

Polynomial kernel：多项式核函数，实现将低维的输入空间映射到高纬的特征空间

Gaussian radial basis function (RBF)：应用于线性不可分时，特征维数少，在没有先验知识时用；

Sigmoid kernel：生成神经网络；

不同的核函数在不同的数据集上的表现可能各有优劣，且需要相应地调整参数。由于西瓜数据集样本数量太小，导致调整参数时容易出现过拟合情况，因此本处我未进行调参，仅用默认参数对比不同核函数对分类器效果的影响

实践题结果与分析T4：Iris数据集对比

2019年10月24日 14:40

Iris 数据集与西瓜数据集最大的区别在于，是一个包含三类的数据集，因此在分类时我们面临的是一个多分类问题。

对于多分类问题，SVM算法和Naïve Bayes算法不会受到影响。但是对于LDA算法，由于二分类LDA算法原理是，将两类的中心点投影到直线上，比较样本投影点与两类中心投影点的距离，从而对样本实现分类。这样的思想无法直接运用在超过2个的分类问题上。

查询资料后得知，所有二分类的机器学习算法都可使用OvO方法或者OvR方法进行改造，从而处理多分类问题。OvR（One vs Rest），是一对剩余的意思。n 种类型的样本进行分类时，分别取一种样本作为一类，将剩余的所有类型的样本看做另一类，这样就形成了 n 个二分类问题，使用逻辑回归算法对 n 个数据集训练出 n 个模型，将待预测的样本传入这 n 个模型中，所得概率最高的那个模型对应的样本类型即认为是该预测样本的类型；OvO（One vs One），是一对一的意思，n 类样本中，每次挑出 2 种类型，两两结合，一共有 C_n^2种二分类情况，使用 C_n^2 种模型预测样本类型，有 C_n^2个预测结果，种类最多的那种样本类型，就认为是该样本最终的预测类型；

我最终选用了OvO方法，即：使用C_n^2个分类器，选出预测结果中最多的种类。OvO 用时较多，但其分类结果更准确，因为每一次二分类时都用真实的类型进行比较，没有混淆其它的类别

Iris数据集样本数量远大于西瓜数据集，这也导致分类表现要远远好于西瓜数据集，分类效果也更加稳定。这说明了提高样本数量更有利于模型的训练。

```
Using LDA method:
Accuracy on Watermelon validation set: 80.00%
Accuracy on Iris validation set: 92.67%

Using Naive Bayes method:
Accuracy on Watermelon validation set: 73.33%
Accuracy on Iris validation set: 94.67%

Using SVM method:
On watermelon dataset:
('Accuracy with Kernel:linear:', '73.33%')
('Accuracy with Kernel:poly:', '26.67%')
('Accuracy with Kernel:rbf:', '33.33%')
('Accuracy with Kernel:sigmoid:', '20.00%')
On Iris dataset:
('Accuracy with Kernel:linear:', '94.67%')
('Accuracy with Kernel:poly:', '94.67%')
('Accuracy with Kernel:rbf:', '97.33%')
('Accuracy with Kernel:sigmoid:', '19.33%')

Using Logistic Regression method:
Accuracy on Watermelon validation set: 73.33%
Accuracy on Iris validation set: 97.33%
```