数据科学基础

作业报告

院　　系：信息与通信工程学院

班　　级：2019211112

学　　号：2021523016

类　　别：交流生

姓　　名：徐川峰

指导老师：刘芳

2021 年 11 月 02 日

1.复现课件中线性 SVM、决策树、朴素贝叶斯分类的示例，并相对课件代码作出如下作图修改（必做）

-设定支持向量分类器的惩罚为 0.05

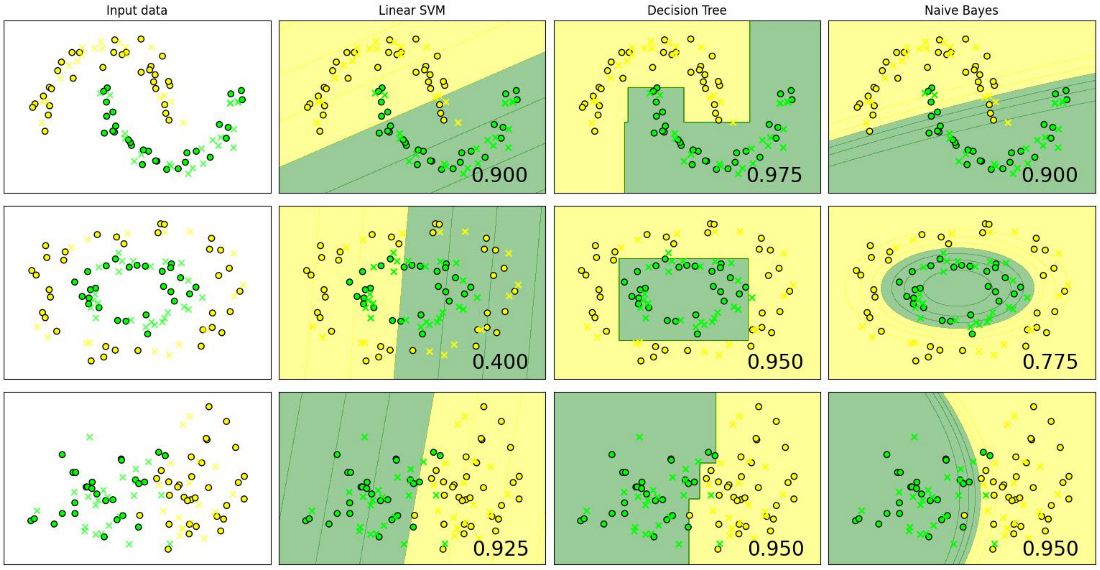
-对朴素贝叶斯分类器的先验概率进行设定（可随机设定）

-在每张结果图上展示图例

-修改散点颜色为黄和绿

-测试结果的正确率保留三位小数展示

```python
import numpy as np
import matplotlib.pyplot as plt
from matplotlib.colors import ListedColormap
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.datasets import make_moons,make_circles,make_classification
from sklearn.svm import SVC
from sklearn.tree import DecisionTreeClassifier
from sklearn.naive_bayes import GaussianNB

names = ["Linear SVM","Decision Tree","Naive Bayes"]
classifiers = [
    SVC(kernel="linear",C=0.05),
    DecisionTreeClassifier(max_depth=5),
    GaussianNB()]

X,y=make_classification(n_features=2,n_redundant=0,n_informative=2,random_state
=1,n_clusters_per_class=1)
rng = np.random.RandomState(2)
X += 2 * rng.uniform(size=X.shape)
linearly_separable = (X,y)

datasets=[make_moons(noise=0.1,random_state=0),make_circles(noise=0.1,factor=0.
5,random_state=
1),linearly_separable]
figure = plt.figure(figsize=(27,9))
i = 1

for ds_cnt,ds in enumerate(datasets):
    X,y = ds
    X = StandardScaler().fit_transform(X)
    X_train,X_test,y_train,y_test = \
        train_test_split(X,y,test_size=.4,random_state=42)

    x_min,x_max = X[:,0].min() - .5,X[:,0].max() + .5
    y_min,y_max = X[:,1].min() - .5,X[:,1].max() + .5
    h= .02
    xx,yy = np.meshgrid(np.arange(x_min,x_max,h),np.arange(y_min,y_max,h))
    cm = ListedColormap((['yellow','green']))
    cm_bright = ListedColormap(['#FFFF00','#00FF00'])
    ax = plt.subplot(len(datasets),len(classifiers) + 1,i)
    if ds_cnt == 0:
        ax.set_title("Input data")
    ax.scatter(X_train[:,0],X_train[:,1],c=y_train,
               cmap=cm_bright,edgecolors='k',marker='o',label='train set')
    ax.scatter(X_test[:,0],X_test[:,1],c=y_test,
               cmap=cm_bright,alpha=0.6,edgecolors='k',marker='x',label='test
set')
    ax.set_xlim(xx.min(),xx.max())
    ax.set_ylim(yy.min(),yy.max())
    ax.set_xticks(())
    ax.set_yticks(())
    i += 1
    for name,clf in zip(names,classifiers):
        ax = plt.subplot(len(datasets),len(classifiers) + 1,i)
        clf.fit(X_train,y_train)
        score = clf.score(X_test,y_test)
        if hasattr(clf,"decision_function"):
            Z = clf.decision_function(np.c_[xx.ravel(),yy.ravel()])
        else:
            Z = clf.predict_proba(np.c_[xx.ravel(),yy.ravel()])[:,1]

        Z = Z.reshape(xx.shape)
        ax.contourf(xx,yy,Z,cmap=cm,alpha=.4)
        ax.scatter(X_train[:,0],X_train[:,1],c=y_train,
                   cmap=cm_bright,edgecolors='k',marker='o',label='train set')
        ax.scatter(X_test[:,0],X_test[:,1],c=y_test,
                   cmap=cm_bright,edgecolors='k',marker='x',label='test set')

        ax.set_xlim(xx.min(),xx.max())
        ax.set_ylim(yy.min(),yy.max())
        ax.set_xticks(())
        ax.set_yticks(())
        if ds_cnt == 0:
            ax.set_title(name)
        ax.text(xx.max() -.3,yy.min() +.3,('%.3f' % score).lstrip('o'),
                size = 20,horizontalalignment='right')
        i += 1
plt.tight_layout()
plt.show()
```

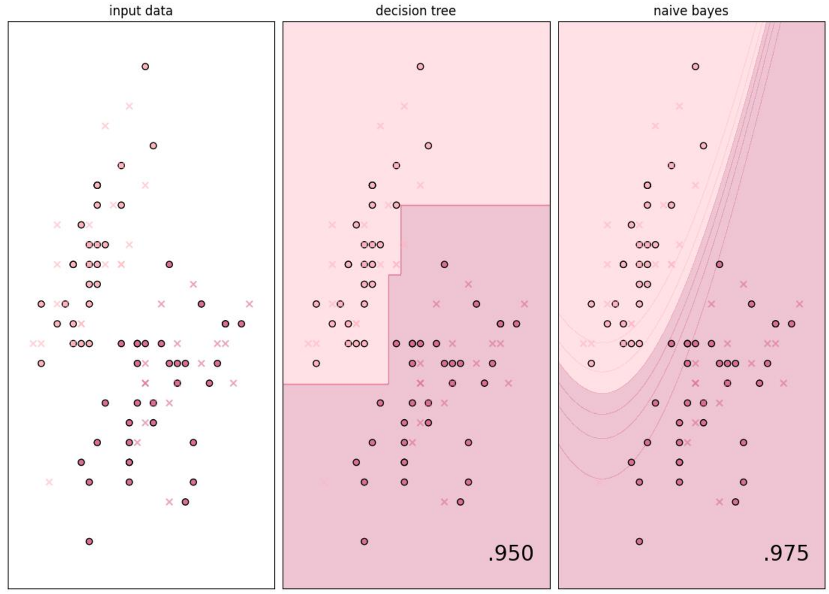


2.创新与拓展（选作）：

-自主选取其他的数据集，采用上述三类分类器进行分类，展示分类结果

```
import numpy as np
import matplotlib.pyplot as plt
from matplotlib.colors import ListedColormap
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.svm import SVC
from sklearn.tree import DecisionTreeClassifier
from sklearn.naive_bayes import GaussianNB
#自选数据集鸢尾花

from sklearn.datasets import load_iris
iris = load_iris() #载入数据集
datasets = [iris]
names =["linear svm","decision tree","naive bayes"]
classifiers = [
    SVC(kernel="linear",C=0.025),
    DecisionTreeClassifier(max_depth=5),
    GaussianNB()]
figure =plt.figure(figsize=(27,9))
i=1

for ds_cnt,ds in enumerate(datasets):
    X=ds.data[0:100,[0,1]]
    Y=ds.target[0:100]
    X=StandardScaler().fit_transform(X)
    X_train,X_test,Y_train,Y_test=train_test_split(X,Y,test_size=.4)
    x_min,x_max=X[:,0].min()-.5,X[:,0].max()+.5
    y_min,y_max=X[:,1].min()-.5,X[:,1].max()+.5
    h=.02
    xx,yy=np.meshgrid(np.arange(x_min,x_max,h),np.arange(y_min,y_max,h))
    cm=ListedColormap((['LightPink','PaleVioletRed']))
    cm_bright=ListedColormap(['#FFB6C1','#DB7093'])
    ax=plt.subplot(len(datasets),len(classifiers)+1,i)
    if ds_cnt==0:
        ax.set_title("input data")

ax.scatter(X_train[:,0],X_train[:,1],c=Y_train,cmap=cm_bright,edgecolors='k',ma
rker='o',label='train set')

ax.scatter(X_test[:,0],X_test[:,1],c=Y_test,cmap=cm_bright,alpha=0.6,edgecolors
        ='k',marker='x',label='test set')
        ax.set_xlim(xx.min(),xx.max())
        ax.set_ylim(yy.min(),yy.max())
        ax.set_xticks(())
        ax.set_yticks(())
        i+=1
        for name,clf in zip(names,classifiers):
            ax=plt.subplot(len(datasets),len(classifiers)+1,i)
            clf.fit(X_train,Y_train)
            score=clf.score(X_test,Y_test)
            if hasattr(clf,"decision_function"):
                Z=clf.decision_function(np.c_[xx.ravel(),yy.ravel()])
            else:
                Z=clf.predict_proba(np.c_[xx.ravel(),yy.ravel()])[:,1]
                Z=Z.reshape(xx.shape)
                ax.contourf(xx,yy,Z,cmap=cm,alpha=.4)

ax.scatter(X_train[:,0],X_train[:,1],c=Y_train,cmap=cm_bright,edgecolors='k',ma
rker='o',label='train set')

ax.scatter(X_test[:,0],X_test[:,1],c=Y_test,cmap=cm_bright,alpha=0.6,edgecolors
                ='k',marker='x',label='test set')
                ax.set_xlim(xx.min(),xx.max())
                ax.set_ylim(yy.min(),yy.max())
                ax.set_xticks(())
                ax.set_yticks(())
                if ds_cnt==0:
                    ax.set_title(name)

ax.text(xx.max()-.3,yy.min()+.3,('%.3f'%score).lstrip('0'),size=20,horizontalal
ignment='right')
                    i+=1
```

```
plt.tight_layout()
plt.show()
```



-探究分类器的参数对于分类结果的影响并进行文字分析（选做）

答：训练样本对分类精度的影响要大于分类器本身的影响，对于SVM来说：随着样本量的增加,平均分类精度随之增高,分类精度方差逐渐降低,对于SVM分类器，当低于30个样本时，分类精度比较稳定，SVM分类对样本数量不敏感，对于SVM分类器， 边缘训练样本构成的支持向量是决定最优超平面的挂念，输入有效的边缘样本能够提高SVM的分类精度。