Раздел I. Методики за класификация на множества от наблюдения

Пример 1.

Множеството “Framingham Heart Study Dataset” (framingham.csv)

Проучване на сърдечно-съдово заболяване сред жители на град Фрамингам, Масачузетс Множеството съдържа 4240 наблюдения, 16 колони и 15 характеристики. Целта на множеството е да се предвиди дали определен пациент има риск в следващите 10 години да развие коронарна болест на сърцето.

Линк към данните:

https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset

https://www.kaggle.com/dileep070/heart-disease-prediction-using-logistic-regression

Тип на анализа – класификационен анализ, т.е. създавене на модел, който да разделя наблюденията аналитично по класове. Класическа методология за анализ, приложена към моделите Логистична регресия, Найвен Бейс, Дърво на решенията (Logistic Regression, Naïve Bayes, Decision Tree) :

Example_framinghamLR.py and Example_framinghamLRKFold.py

Example_framinghamNB.py

Example_framinghamDT.py and Example_framinghamRF.py

За моделите Decision Tree and Random Forest може да се запознаете от pdf файла, приложен към материалите по темата.

Различните модели са приложени по различен начин – или с прилагане на train_test_split разделянето или чрез Kfold разделянето на тренировъчно и тестово подмножество.

Оценката на модела се извършва върху тестовото подмножество чрез confusion matrix and classification report.

Пример 2.

Множеството от наблюдения за качеството на водата.

Разглеждаме множеството от наблюдения „water_potability.csv“. То е съставено от 3276 наблюдения с по 10 характеристики. В него се съдържа информация за качеството на водата, базирано на различни нейни характеристики. Взимайки предвид

тези характеристики може да бъде определено, дали водата е питейна или не – информация, за което можем да видим в характеристиката „Potability“, която приема стойности 0 и 1 (0 – непитейна, 1 – питейна).

Линк към данните:

https://www.kaggle.com/adityakadiwal/water-potability

Класическа методология за анализ, приложена към моделите Логистична регресия, най-близки съседи (Logistic Regression, KNeighborsClassifier) :

Example_waterLR.py

Example_waterkNN.py

ПОДОБРЕНИЯ

Logistic Regression:

logreg = LogisticRegression(solver='lbfgs', max_iter=5000, class_weight='balanced')

logreg=LogisticRegression(solver='liblinear', class_weight="balanced")

Decision Tree:

logreg = DecisionTreeClassifier(class_weight='balanced', random_state=366)

KNeighborsClassifier Параметърът “class_weight='balanced'” не е приложим

Naïve Bayes Параметърът “class_weight='balanced'” не е приложим

Прилагане на методология за равен брой наблюдения в тренировъчното подмножество.
Програмни файлове:

ExampleEqNum_framinghamDT.py

ExampleEqNum_framinghamLR.py

За упражнение.

За студенти втори семестър.На базата на тези два примера, приложете различни методики за избор на тренировъчното подмножество за различни модели с цел да получите високи стойности на коефициента за чувствителност (recall) за всички класове за следващите примери.

За студенти първи семестър.На базата на тези два примера, модифицирайте програмните кодове за моделиране на наблюденията (на данните) за да получите съответни модели за данните от следващите примери. (Пояснявам – приложете същите програми кодове за данните от следващите примери).

Пример 1.

Haberman data set. Линк към файла с наблюдения

https://archive.ics.uci.edu/ml/datasets/haberman%27s+survival_

Четене на файла

```
import pandas as pd
data = pd.read_csv('haberman.csv', delimiter=',', header = None)
 X = data.values[:, :3]
y = data.values[:, 3]
```

Пример 2.

Pima_diabetes data set. Линк към файла с наблюдения

Линк към данните.

https://networkrepository.com/pima-indians-diabetes.php

Четене на данните

```
import pandas as pd
data = pd.read_csv('pima_diabetes.csv')

X = data.values[:, :8]
y = data.values[:, 8]
```

Пример 3.

Изследване на Центъра за кръвопреливане в Тайван. (transfusion.txt)

https://archive.ics.uci.edu/ml/datasets/Blood+Transfusion+Service+Center

За да се демонстрира маркетинговият модел RFMTC (модифицирана версия на RFM) е направено проучване на базата данни за донорите на Центъра за кръвопреливане в град Хсин-Чу в Тайван. Центърът позиционира свой специализиран автобус за кръвопреливане в един университет в град Хсин-Чу, за да събира кръв, дарена на всеки три месеца. За да бъде изграден модел FRMTC, са избрани произволно 748 донора от базата данни на донорите. Тези 748 наблюдения са изразени чрез няколко показателя. Следват имена на променливи, типа на променливите, мерната единица и кратко описание.

R (Рецензия - месеци от последното дарение),
F (Честота - общ брой на даренията)
M (Стойност на даренията - обща кръв, дарена в куб.см)
T (Време - месеци от първото дарение)
Целава променлива, представяща дали той / тя е дарил кръв през март 2007 г. (1-дарява кръв; 0 означава, че не е дарил/а кръв).

Четене на данните

```
from numpy import loadtxt
data = loadtxt('transfusion.txt', delimiter=',')
X = data[:, :4]
y = data[:, 4]
```

Пример 4 Анализ на банкови данни (bankrupt.csv)

https://archive.ics.uci.edu/ml/datasets/Taiwanese+Bankruptcy+Prediction

https://www.kaggle.com/datasets/fedesoriano/company-bankruptcy-prediction

Описани са 95 характеристики и една целева променлива в нулева колона.

Четене на данните

```
import pandas as pd
dataw = pd.read_csv('bankrup.csv')
X=dataw.values[:,1:]
y=dataw.values[:,0]
```

Файловете с данни са приложени !

2 март 2025
Проф. Иван Иванов