Лабораторная работа №5. Решение задач классификации с помощью байесовского классификатора и метода k-ближайших соседей.

Часть 1. Гауссов наивный байесовский классификатор.

Используемый набор данных: Wine (https://archive.ics.uci.edu/ml/datasets/Wine)

In [1]:

```
from IPython.display import display
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.preprocessing import label_binarize
from sklearn.model_selection import train_test_split
from sklearn.multiclass import OneVsRestClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import classification_report, roc_curve, roc_auc_score
from itertools import cycle
import os
import requests

%matplotlib inline

pd.options.display.max_columns = None
```

In [2]:

```
def downloadFile(url, filePath):
    if not os.path.exists(filePath):
        req = requests.get(url)
        f = open(filePath, "wb")
        f.write(req.content)
        f.close

url = "https://archive.ics.uci.edu/ml/machine-learning-databases/wine"
downloadFile(url + "/wine.data", "dataset/wine.data")
downloadFile(url + "/wine.names", "dataset/wine.names")
```

In [3]:

```
headers = ["Class", "Alcohol", "Malic acid", "Ash", "Alcalinity of ash", "Magnesium",
"Total phenols", "Flavanoids",
           "Nonflavanoid phenols", "Proanthocyanins", "Color intensity", "Hue", "OD280/
OD315 of diluted wines", "Proline"]
data = pd.read_csv("dataset/wine.data", names=headers)
data["Class"] = data["Class"].astype("category")
data.sample(40)
```

Out[3]:

	Class	Alcohol	Malic acid	Ash	Alcalinity of ash	Magnesium	Total phenols	Flavanoids	Nonflavanoid phenols
131	3	12.88	2.99	2.40	20.0	104	1.30	1.22	0.24
105	2	12.42	2.55	2.27	22.0	90	1.68	1.84	0.66
118	2	12.77	3.43	1.98	16.0	80	1.63	1.25	0.43
90	2	12.08	1.83	2.32	18.5	81	1.60	1.50	0.52
57	1	13.29	1.97	2.68	16.8	102	3.00	3.23	0.31
34	1	13.51	1.80	2.65	19.0	110	2.35	2.53	0.29
50	1	13.05	1.73	2.04	12.4	92	2.72	3.27	0.17
135	3	12.60	2.46	2.20	18.5	94	1.62	0.66	0.63
115	2	11.03	1.51	2.20	21.5	85	2.46	2.17	0.52
97	2	12.29	1.41	1.98	16.0	85	2.55	2.50	0.29
138	3	13.49	3.59	2.19	19.5	88	1.62	0.48	0.58
121	2	11.56	2.05	3.23	28.5	119	3.18	5.08	0.47
51	1	13.83	1.65	2.60	17.2	94	2.45	2.99	0.22
71	2	13.86	1.51	2.67	25.0	86	2.95	2.86	0.21
21	1	12.93	3.80	2.65	18.6	102	2.41	2.41	0.25
104	2	12.51	1.73	1.98	20.5	85	2.20	1.92	0.32
46	1	14.38	3.59	2.28	16.0	102	3.25	3.17	0.27
72	2	13.49	1.66	2.24	24.0	87	1.88	1.84	0.27
20	1	14.06	1.63	2.28	16.0	126	3.00	3.17	0.24
156	3	13.84	4.12	2.38	19.5	89	1.80	0.83	0.48
160	3	12.36	3.83	2.38	21.0	88	2.30	0.92	0.50
163	3	12.96	3.45	2.35	18.5	106	1.39	0.70	0.40
85	2	12.67	0.98	2.24	18.0	99	2.20	1.94	0.30
164	3	13.78	2.76	2.30	22.0	90	1.35	0.68	0.41
47	1	13.90	1.68	2.12	16.0	101	3.10	3.39	0.21
64	2	12.17	1.45	2.53	19.0	104	1.89	1.75	0.45
162	3	12.85	3.27	2.58	22.0	106	1.65	0.60	0.60
98	2	12.37	1.07	2.10	18.5	88	3.52	3.75	0.24
78	2	12.33	0.99	1.95	14.8	136	1.90	1.85	0.35
92	2	12.69	1.53	2.26	20.7	80	1.38	1.46	0.58
177	3	14.13	4.10	2.74	24.5	96	2.05	0.76	0.56
142	3	13.52	3.17	2.72	23.5	97	1.55	0.52	0.50
149	3	13.08	3.90	2.36	21.5	113	1.41	1.39	0.34
154	3	12.58	1.29	2.10	20.0	103	1.48	0.58	0.53
54	1	13.74	1.67	2.25	16.4	118	2.60	2.90	0.21
133	3	12.70	3.55	2.36	21.5	106	1.70	1.20	0.17

	Class	Alcohol	Malic acid	Ash	Alcalinity of ash	Magnesium	Total phenols	Flavanoids	Nonflavanoid phenols
36	1	13.28	1.64	2.84	15.5	110	2.60	2.68	0.34
17	1	13.83	1.57	2.62	20.0	115	2.95	3.40	0.40
33	1	13.76	1.53	2.70	19.5	132	2.95	2.74	0.50
13	1	14.75	1.73	2.39	11.4	91	3.10	3.69	0.43

In [4]:

```
display(data.isna().sum())
display(data.describe())
```

```
Class                           0
Alcohol                         0
Malic acid                      0
Ash                             0
Alcalinity of ash               0
Magnesium                       0
Total phenols                   0
Flavanoids                      0
Nonflavanoid phenols            0
Proanthocyanins                 0
Color intensity                 0
Hue                             0
OD280/OD315 of diluted wines    0
Proline                         0
dtype: int64
```

	Alcohol	Malic acid	Ash	Alcalinity of ash	Magnesium	Total phenols	Flavanoids
count	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000
mean	13.000618	2.336348	2.366517	19.494944	99.741573	2.295112	2.029270
std	0.811827	1.117146	0.274344	3.339564	14.282484	0.625851	0.998859
min	11.030000	0.740000	1.360000	10.600000	70.000000	0.980000	0.340000
25%	12.362500	1.602500	2.210000	17.200000	88.000000	1.742500	1.205000
50%	13.050000	1.865000	2.360000	19.500000	98.000000	2.355000	2.135000
75%	13.677500	3.082500	2.557500	21.500000	107.000000	2.800000	2.875000
max	14.830000	5.800000	3.230000	30.000000	162.000000	3.880000	5.080000

Пропусков в данных нет.

Подготовим данные для классификации: выберем признаки и метки и сформируем тренировочные и тестовые наборы.

In [5]:

```
classes = data["Class"].unique()
n_classes = len(classes)
y = label_binarize(data["Class"], classes=classes)
X = data.drop(columns=["Class"]).copy()

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=2
7)
```

Создадим классификатор, обучим его, а затем выполним классификацию.

In [6]:

```
y_score = OneVsRestClassifier(GaussianNB()).fit(X_train, y_train).predict(X_test)
```

In [7]:

```
fpr, tpr, auc = dict(), dict(), dict()
for i in range(n_classes):
    y_test_cl = y_test[:,i]
    y_score_cl = y_score[:,i]
    fpr[i], tpr[i], _ = roc_curve(y_test_cl, y_score_cl)
    auc[i] = roc_auc_score(y_test_cl, y_score_cl)

lw = 2
colors = cycle(['aqua', 'darkorange', 'cornflowerblue'])
for i, color in zip(range(n_classes), colors):
    plt.plot(fpr[i], tpr[i], color=color, lw=lw, label='ROC of class {0} (AUC = {1:0.2
f})'.format(i, auc[i]))
plt.plot([0, 1], [0, 1], 'k--', lw=lw)
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.1])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver operating characteristic for multi-class classification')
plt.legend(loc="lower right")
plt.show()
```