

Лабораторная работа № 3.

Подготовка обучающей и тестовой выборки, кросс-валидация и подбор гиперпараметров на примере метода ближайших соседей.

Цель лабораторной работы: изучение способов подготовки выборки и подбора гиперпараметров на примере метода ближайших соседей.

Задание:

1. Выбрать набор данных (датасет) для решения задачи классификации или регрессии.
2. С использованием метода `train_test_split` разделить выборку на обучающую и тестовую.
3. Обучить модель ближайших соседей для произвольно заданного гиперпараметра `K`. Оценить качество модели с помощью подходящих для задачи метрик.
4. Произвести подбор гиперпараметра `K` с использованием `GridSearchCV` и/или `RandomizedSearchCV` и кросс-валидации, оценить качество оптимальной модели. Желательно использование нескольких стратегий кросс-валидации.
5. Сравнить метрики качества исходной и оптимальной моделей.

Выполнение:

В качестве набора данных будем использовать датасет про вино: https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_wine.html#sklearn.datasets.load_wine

Импорт библиотек

```
import numpy as np
import pandas as pd
from sklearn.datasets import load_wine
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import MinMaxScaler
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score
from sklearn.metrics import confusion_matrix, plot_confusion_matrix
from sklearn.metrics import recall_score, precision_score
from sklearn.model_selection import cross_validate
from sklearn.model_selection import KFold, LeaveOneOut, LeavePOut
from sklearn.model_selection import GridSearchCV, RandomizedSearchCV
from typing import Dict
import seaborn as sns
import matplotlib.pyplot as plt
```

In [86]:

Загрузка данных

```
wine = load_wine()
```

In [2]:

```
for x in wine:
    print(x)
```

In [3]:

```
data
target
frame
target_names
DESCR
feature_names
```

In [4]:

```
# Наименование значений целевого признака
wine['target_names']
```

```
array(['class_0', 'class_1', 'class_2'], dtype='<U7')
```

Out[4]:

```
# Признаки датасета
wine['feature_names']
```

In [5]:

Out[5]:

```
['alcohol',
'malic_acid',
'ash',
'alcalinity_of_ash',
'magnesium',
'total_phenols',
'flavanoids',
'nonflavanoid_phenols',
'proanthocyanins',
'color_intensity',
'hue',
'od280/od315_of_diluted_wines',
'proline']
```

In [6]:

```
wine['data'].shape
```

Out[6]:

```
(178, 13)
```

In [7]:

```
# Преобразование в pandas DataFrame
w_data = pd.DataFrame(data=np.c_[wine['data'], wine['target']], columns = wine['feature_names']+['target'])
```

In [8]:

```
w_data
```

Out[8]:

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonflavanoid_phenols	proanthocyanins	color_intensity
0	14.23	1.71	2.43	15.6	127.0	2.80	3.06	0.28	2.29	5.64
1	13.20	1.78	2.14	11.2	100.0	2.65	2.76	0.26	1.28	4.38
2	13.16	2.36	2.67	18.6	101.0	2.80	3.24	0.30	2.81	5.68
3	14.37	1.95	2.50	16.8	113.0	3.85	3.49	0.24	2.18	7.80
4	13.24	2.59	2.87	21.0	118.0	2.80	2.69	0.39	1.82	4.32
...
173	13.71	5.65	2.45	20.5	95.0	1.68	0.61	0.52	1.06	7.70
174	13.40	3.91	2.48	23.0	102.0	1.80	0.75	0.43	1.41	7.30
175	13.27	4.28	2.26	20.0	120.0	1.59	0.69	0.43	1.35	10.20
176	13.17	2.59	2.37	20.0	120.0	1.65	0.68	0.53	1.46	9.30
177	14.13	4.10	2.74	24.5	96.0	2.05	0.76	0.56	1.35	9.20

178 rows × 14 columns

In [9]:

```
# Статистические характеристики датасета
w_data.describe()
```

Out[9]:

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonflavanoid_phenols	proanthocyanins	color_intensity
count	178.000000	178.000000	178.000000	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	0.361854	1.590899	6.411834
std	0.811827	1.117146	0.274344	3.339564	14.282484	0.625851	0.998859	0.124453	0.572359	1.059296
min	11.030000	0.740000	1.360000	10.600000	70.000000	0.980000	0.340000	0.130000	0.410000	3.760000
25%	12.362500	1.602500	2.210000	17.200000	88.000000	1.742500	1.205000	0.270000	1.250000	5.950000
50%	13.050000	1.865000	2.360000	19.500000	98.000000	2.355000	2.135000	0.340000	1.555000	6.595000
75%	13.677500	3.082500	2.557500	21.500000	107.000000	2.800000	2.875000	0.437500	1.950000	7.700000
max	14.830000	5.800000	3.230000	30.000000	162.000000	3.880000	5.080000	0.660000	3.580000	10.200000

In [13]:

```
# Значения признаков
wine.data[:5]
```

Out[13]:

```
array([[1.423e+01, 1.710e+00, 2.430e+00, 1.560e+01, 1.270e+02, 2.800e+00,
        3.060e+00, 2.800e-01, 2.290e+00, 5.640e+00, 1.040e+00, 3.920e+00,
        1.065e+03],
       [1.320e+01, 1.780e+00, 2.140e+00, 1.120e+01, 1.000e+02, 2.650e+00,
        2.760e+00, 2.600e-01, 1.280e+00, 4.380e+00, 1.050e+00, 3.400e+00,
        1.050e+03],
       [1.316e+01, 2.360e+00, 2.670e+00, 1.860e+01, 1.010e+02, 2.800e+00,
        3.240e+00, 3.000e-01, 2.810e+00, 5.680e+00, 1.030e+00, 3.170e+00,
        1.185e+03],
       [1.437e+01, 1.950e+00, 2.500e+00, 1.680e+01, 1.130e+02, 3.850e+00,
        3.490e+00, 2.400e-01, 2.180e+00, 7.800e+00, 8.600e-01, 3.450e+00,
        1.480e+03],
       [1.324e+01, 2.590e+00, 2.870e+00, 2.100e+01, 1.180e+02, 2.800e+00,
        2.690e+00, 3.900e-01, 1.820e+00, 4.320e+00, 1.040e+00, 2.930e+00,
        7.350e+02]])
```

In [15]:

```
# Отмасштабируем признаки, т.к. min и max у некоторых признаков находятся не в одном диапазоне
sc = MinMaxScaler()
sc_wine = sc.fit_transform(wine.data)
```

Разделение выборки на обучающую и тестовую

In [16]:

```
X_train, X_test, Y_train, Y_test = train_test_split(sc_wine, wine.target, test_size=0.3, random_state=1)
```

In [17]:

```
# Размер обучающей выборки
X_train.shape, Y_train.shape
```

Out[17]:

```
((124, 13), (124,))
```

In [18]:

```
# Размер тестовой выборки
X_test.shape, Y_test.shape
```

Out[18]:

```
((54, 13), (54,))
Функция train_test_split делит выборку так, чтобы сохранились все классы
```

In [19]:

```
np.unique(Y_train)
```

Out[19]:

```
array([0, 1, 2])
```

In [20]:

```
np.unique(Y_test)
```

Out[20]:

```
array([0, 1, 2])
```

Построим базовую модель на основе ближайших соседей с произвольно заданным гиперпараметром K

In [94]:

```
# 4 ближайших соседа
cl1 = KNeighborsClassifier(n_neighbors=4)
cl1.fit(X_train, Y_train)
target1 = cl1.predict(X_test)
target11 = cl1.predict(X_train)
len(target1), target1
```

Out[94]:

```
(54,
 array([2, 1, 0, 1, 0, 2, 1, 0, 2, 1, 0, 0, 1, 0, 1, 1, 2, 0, 1, 0, 0, 1,
        2, 0, 0, 2, 0, 0, 0, 2, 1, 2, 2, 0, 1, 1, 1, 1, 1, 0, 0, 1, 2, 0,
        0, 0, 0, 0, 0, 0, 1, 2, 2, 0]))
```

In [36]:

```
# 2 ближайших соседа
cl2 = KNeighborsClassifier(n_neighbors=2)
cl2.fit(X_train, Y_train)
target2 = cl2.predict(X_test)
len(target2), target2
```

Out[36]:

```
(54,
 array([2, 1, 0, 1, 0, 1, 1, 0, 2, 1, 0, 0, 1, 0, 1, 1, 2, 0, 1, 0, 0, 1,
        2, 0, 0, 2, 0, 0, 0, 2, 1, 2, 2, 0, 1, 1, 1, 1, 1, 0, 0, 1, 2, 0,
        0, 0, 1, 0, 0, 0, 1, 2, 2, 0]))
```

Метрика Accurasy

In [34]:

```
accuracy_score(Y_test, target1)
```

```
0.9629629629629629
```

Out[34]:

In [37]:

```
accuracy_score(Y_test, target2)
```

Out[37]:

```
0.9629629629629629
```

Точность в случае двух и четырех ближайших соседей составляет 96%. Однако эта метрика показывает точность по всем классам в целом, поэтому точность в каждом отдельном классе может отличаться.

In [31]:

```
def accuracy_score_for_classes(
    y_true: np.ndarray,
    y_pred: np.ndarray) -> Dict[int, float]:
    """
    Вычисление метрики accuracy для каждого класса
    Возвращает словарь: ключ - метка класса,
    значение - Accuracy для данного класса
    """
    # Для удобства фильтрации сформируем Pandas DataFrame
    d = {'t': y_true, 'p': y_pred}
    df = pd.DataFrame(data=d)
    # Метки классов
    classes = np.unique(y_true)
    # Результирующий словарь
    res = dict()
    # Перебор меток классов
    for c in classes:
        # отфильтруем данные, которые соответствуют текущей метке класса в истинных значениях
        temp_dataflt = df[df['t']==c]
        # расчет accuracy для заданной метки класса
        temp_acc = accuracy_score(
            temp_dataflt['t'].values,
            temp_dataflt['p'].values)
        # сохранение результата в словарь
        res[c] = temp_acc
    return res

def print_accuracy_score_for_classes(
    y_true: np.ndarray,
    y_pred: np.ndarray):
    """
    Вывод метрики accuracy для каждого класса
    """
    accs = accuracy_score_for_classes(y_true, y_pred)
    if len(accs)>0:
        print('Метка \t Accuracy')
    for i in accs:
        print('{ } \t {}'.format(i, accs[i]))
```

In [35]:

```
print_accuracy_score_for_classes(Y_test, target1)
```

```
Метка  Accuracy
0      1.0
1      0.8947368421052632
2      1.0
```

Accuracy для класса 0 и 2 составляет 100%, но для класса 1 проседает до 89%.

In [38]:

```
print_accuracy_score_for_classes(Y_test, target2)
```

```
Метка  Accuracy
0      1.0
1      0.9473684210526315
2      0.9166666666666666
```

Accuracy для класса 0 составляет 100%, но для классов 1 и 2 - 95% и 92% соответственно.

In [39]:

```
# Конвертация целевого признака в бинарный
def convert_target_to_binary(array: np.ndarray, target: int) -> np.ndarray:
    # Если целевой признак совпадает с указанным, то 1 иначе 0
    res = [1 if x==target else 0 for x in array]
    return res
```

In [40]:

```
# Допустим целевой признак == 2, будем считать этот случай = 1 в бинарном признаке
bin_Y_train = convert_target_to_binary(Y_train, 2)
list(zip(Y_train, bin_Y_train))[:10]
```

Out[40]:

```
[(1, 0),
 (1, 0),
 (0, 0),
 (1, 0),
 (2, 1),
 (1, 0),
 (1, 0),
 (0, 0),
 (2, 1),
 (1, 0)]
```

In [41]:

```
bin_Y_test = convert_target_to_binary(Y_test, 2)
list(zip(Y_test, bin_Y_test))[:10]
```

Out[41]:

```
[(2, 1),
 (1, 0),
 (0, 0),
 (1, 0),
 (0, 0),
 (2, 1),
 (1, 0),
 (0, 0),
 (2, 1),
 (1, 0)]
```

In [51]:

```
bin_target1=convert_target_to_binary(target1, 2)
bin_target2=convert_target_to_binary(target2, 2)
```

Матрица ошибок

In [60]:

```
# для 2 ближайших соседей с целевым признаком == 2
confusion_matrix(bin_Y_test, bin_target2, labels = [0, 1])
```

Out[60]:

```
array([[42,  0],
       [ 1, 11]])
```

In [44]:

```
# для 4 ближайших соседей
confusion_matrix(Y_test, target1, labels = [0, 1, 2])
```

Out[44]:

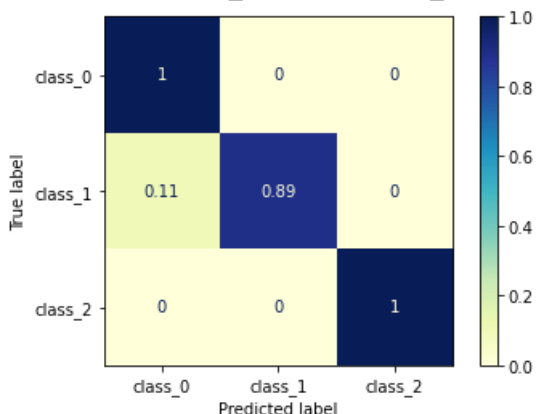
```
array([[23,  0,  0],
       [ 2, 17,  0],
       [ 0,  0, 12]])
```

In [48]:

```
# для 4 ближайших соседей
plot_confusion_matrix(cl1, X_test, Y_test,
                      display_labels=wine.target_names, cmap=plt.cm.YlGnBu, normalize='true')
```

Out[48]:

<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x7fc3f80510a0>

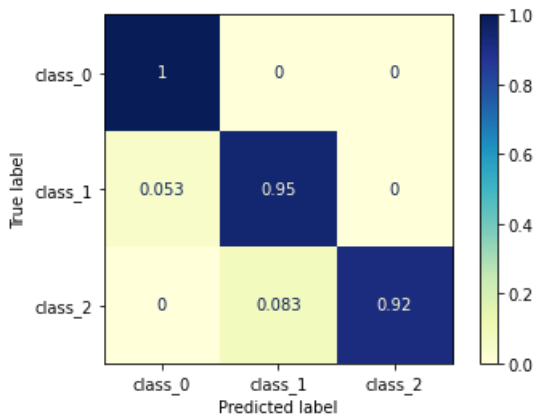


In [49]:

```
# для 2 ближайших соседей
plot_confusion_matrix(cl2, X_test, Y_test,
                      display_labels=wine.target_names, cmap=plt.cm.YlGnBu, normalize='true')
```

Out[49]:

```
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x7fc3f8051d90>
```



Метрика recall (полнота), precision

1) Доля верно предсказанных классификатором положительных объектов, из всех действительно положительных объектов. (TP/(TP+FN))

In [53]:

```
recall_score(bin_Y_test, bin_target1), recall_score(bin_Y_test, bin_target2)
```

Out[53]:

```
(1.0, 0.9166666666666666)
```

2) Доля верно предсказанных классификатором положительных объектов, из всех объектов, которые классификатор верно или неверно определил как положительные. (TP/(TP+FP))

In [55]:

```
precision_score(bin_Y_test, bin_target1), precision_score(bin_Y_test, bin_target2)
```

Out[55]:

```
(1.0, 1.0)
```

Кросс-валидация

In [62]:

```
# Словарь метрик качества
scoring = {'precision': 'precision_weighted',
           'recall': 'recall_weighted',
           'f1': 'f1_weighted'}
```

In [65]:

```
# Кросс-валидация для 5 фолдов, на которые мы разделили обучающую и тестовую выборки
scores = cross_validate(KNeighborsClassifier(n_neighbors = 2),
                        sc_wine, wine.target, scoring = scoring, cv = 5,
                        return_train_score = True)

scores
```

Out[65]:

```
{'fit_time': array([0.00216603, 0.00057101, 0.00048399, 0.00048113, 0.00061607]),
 'score_time': array([0.00609899, 0.00359392, 0.003232 , 0.00280595, 0.00400901]),
 'test_precision': array([0.80886752, 0.94910645, 1. , 1. , 0.92593407]),
 'train_precision': array([0.97404032, 0.96817443, 0.97404032, 0.98027972, 0.98025504]),
 'test_recall': array([0.80555556, 0.94444444, 1. , 1. , 0.91428571]),
 'train_recall': array([0.97183099, 0.96478873, 0.97183099, 0.97902098, 0.97902098]),
 'test_f1': array([0.80041152, 0.94352462, 1. , 1. , 0.91265664]),
 'train_f1': array([0.9718937 , 0.96487031, 0.9718937 , 0.9790619 , 0.97905014])}
```

Значения метрик precision, recall и f1 на обучающей выборке стабильнее и ближе к 100%, чем на тестовой выборке.

In [67]:

```
# Кросс-валидация методом LeaveOneOut: в тестовой выборке 1 элемент, а все остальные в обучающей
loo = LeaveOneOut()
```

In [68]:

```
scores1 = cross_validate(KNeighborsClassifier(n_neighbors = 2),
                        sc_wine, wine.target, scoring = scoring, cv = loo,
                        return_train_score = True)

scores1
```

```
/Users/kalashnikova/opt/anaconda3/lib/python3.8/site-packages/sklearn/metrics/_classification.py:1221:
UndefinedMetricWarning: Precision is ill-defined and being set to 0.0 in labels with no predicted
samples. Use `zero_division` parameter to control this behavior.
  warn_prf(average, modifier, msg_start, len(result))
/Users/kalashnikova/opt/anaconda3/lib/python3.8/site-packages/sklearn/metrics/_classification.py:1221:
UndefinedMetricWarning: Recall is ill-defined and being set to 0.0 in labels with no true samples. Use `
zero_division` parameter to control this behavior.
  warn_prf(average, modifier, msg_start, len(result))
/Users/kalashnikova/opt/anaconda3/lib/python3.8/site-packages/sklearn/metrics/_classification.py:1221:
UndefinedMetricWarning: Precision is ill-defined and being set to 0.0 in labels with no predicted
```

```

samples. Use `zero_division` parameter to control this behavior.
    warn_prf(average, modifier, msg_start, len(result))
/Users/kalashnikova/opt/anaconda3/lib/python3.8/site-packages/sklearn/metrics/_classification.py:1221:
UndefinedMetricWarning: Recall is ill-defined and being set to 0.0 in labels with no true samples. Use
zero_division` parameter to control this behavior.
    warn_prf(average, modifier, msg_start, len(result))
/Users/kalashnikova/opt/anaconda3/lib/python3.8/site-packages/sklearn/metrics/_classification.py:1221:
UndefinedMetricWarning: Precision is ill-defined and being set to 0.0 in labels with no predicted
samples. Use `zero_division` parameter to control this behavior.
    warn_prf(average, modifier, msg_start, len(result))
/Users/kalashnikova/opt/anaconda3/lib/python3.8/site-packages/sklearn/metrics/_classification.py:1221:
UndefinedMetricWarning: Recall is ill-defined and being set to 0.0 in labels with no true samples. Use
zero_division` parameter to control this behavior.
    warn_prf(average, modifier, msg_start, len(result))
/Users/kalashnikova/opt/anaconda3/lib/python3.8/site-packages/sklearn/metrics/_classification.py:1221:
UndefinedMetricWarning: Precision is ill-defined and being set to 0.0 in labels with no predicted
samples. Use `zero_division` parameter to control this behavior.
    warn_prf(average, modifier, msg_start, len(result))
/Users/kalashnikova/opt/anaconda3/lib/python3.8/site-packages/sklearn/metrics/_classification.py:1221:
UndefinedMetricWarning: Recall is ill-defined and being set to 0.0 in labels with no true samples. Use
zero_division` parameter to control this behavior.
    warn_prf(average, modifier, msg_start, len(result))
/Users/kalashnikova/opt/anaconda3/lib/python3.8/site-packages/sklearn/metrics/_classification.py:1221:
UndefinedMetricWarning: Precision is ill-defined and being set to 0.0 in labels with no predicted
samples. Use `zero_division` parameter to control this behavior.
    warn_prf(average, modifier, msg_start, len(result))
/Users/kalashnikova/opt/anaconda3/lib/python3.8/site-packages/sklearn/metrics/_classification.py:1221:
UndefinedMetricWarning: Recall is ill-defined and being set to 0.0 in labels with no true samples. Use
zero_division` parameter to control this behavior.
    warn_prf(average, modifier, msg_start, len(result))
/Users/kalashnikova/opt/anaconda3/lib/python3.8/site-packages/sklearn/metrics/_classification.py:1221:
UndefinedMetricWarning: Precision is ill-defined and being set to 0.0 in labels with no predicted
samples. Use `zero_division` parameter to control this behavior.
    warn_prf(average, modifier, msg_start, len(result))
/Users/kalashnikova/opt/anaconda3/lib/python3.8/site-packages/sklearn/metrics/_classification.py:1221:
UndefinedMetricWarning: Recall is ill-defined and being set to 0.0 in labels with no true samples. Use
zero_division` parameter to control this behavior.
    warn_prf(average, modifier, msg_start, len(result))
/Users/kalashnikova/opt/anaconda3/lib/python3.8/site-packages/sklearn/metrics/_classification.py:1221:
UndefinedMetricWarning: Precision is ill-defined and being set to 0.0 in labels with no predicted
samples. Use `zero_division` parameter to control this behavior.
    warn_prf(average, modifier, msg_start, len(result))
/Users/kalashnikova/opt/anaconda3/lib/python3.8/site-packages/sklearn/metrics/_classification.py:1221:
UndefinedMetricWarning: Recall is ill-defined and being set to 0.0 in labels with no true samples. Use
zero_division` parameter to control this behavior.
    warn_prf(average, modifier, msg_start, len(result))

```

Out[68]:

```

{'fit_time': array([0.00428391, 0.00054288, 0.00053024, 0.00052381, 0.00046587,
0.00046206, 0.00046301, 0.00051832, 0.00052094, 0.00049686,
0.00048685, 0.00075078, 0.00058413, 0.00166392, 0.00051618,
0.00044298, 0.00045013, 0.00040388, 0.00042892, 0.00043511,
0.00043201, 0.00060582, 0.00048304, 0.00047469, 0.00041389,
0.00046682, 0.00039101, 0.00042486, 0.00039315, 0.00039411,
0.00038099, 0.0003798 , 0.00037718, 0.00037694, 0.00038719,
0.00039387, 0.00037789, 0.00037503, 0.00037694, 0.00038385,
0.00038624, 0.00038409, 0.00038505, 0.00038004, 0.00039697,
0.00037694, 0.00038195, 0.00037599, 0.00037789, 0.00038624,
0.00039506, 0.00039887, 0.00037813, 0.00037718, 0.00037599,
0.0003829 , 0.00037479, 0.00037885, 0.00040698, 0.00041509,
0.00040317, 0.00041294, 0.00041199, 0.00047207, 0.00044608,
0.00042701, 0.00043201, 0.00044727, 0.00046301, 0.00049305,
0.00046587, 0.00046277, 0.00071812, 0.00048804, 0.00046706,
0.00046062, 0.00070286, 0.0004878 , 0.00046492, 0.00046706,
0.00073409, 0.00048709, 0.00047088, 0.00049782, 0.00108004,
0.00062895, 0.00048494, 0.00046515, 0.0004437 , 0.00041199,
0.00059199, 0.00044203, 0.00043702, 0.00039387, 0.00043488,
0.00039101, 0.00043178, 0.00037408, 0.00037718, 0.00037217,
0.0003767 , 0.00037503, 0.00037527, 0.00037408, 0.00037193,
0.00037098, 0.00038195, 0.00052714, 0.00058508, 0.00081801,
0.00047684, 0.00045896, 0.00044918, 0.00037885, 0.00042892,
0.00037694, 0.00047421, 0.00041723, 0.00037789, 0.00049782,
0.00051999, 0.0004921 , 0.00039601, 0.00038099, 0.000458 ,
0.00043702, 0.00041699, 0.00037909, 0.00038028, 0.00037885,
0.00037384, 0.00054407, 0.00042009, 0.00037622, 0.00037599,
0.00037503, 0.00038624, 0.00037408, 0.00037193, 0.00037098,
0.00044799, 0.00037503, 0.00037289, 0.00059009, 0.00046515,
0.00195694, 0.00046325, 0.00075102, 0.00046802, 0.00054097,
0.00048399, 0.00055313, 0.00045013, 0.00049496, 0.000489 ,
0.00051308, 0.0004499 , 0.00047708, 0.00045085, 0.00046611,
0.0004909 , 0.0004518 , 0.00047779, 0.00045109, 0.00046682,
0.0004549 , 0.00044799, 0.00045729, 0.00041485, 0.00045586,
0.00052404, 0.00052214, 0.00044918, 0.00046492, 0.00060487,
0.00048923, 0.00045896, 0.00069809]),
'score_time': array([0.0070622 , 0.00250316, 0.00240588, 0.00216222, 0.00206423,
0.00209498, 0.00207996, 0.00300789, 0.00216389, 0.00213695,

```


[illegible]

Подбор гиперпараметра K с использованием GridSearch и кросс-валидации

```
# создаем массив со значениями гиперпараметра (кол-во ближайших соседей)
n_range = np.array(range(5,55,5))
tuned_params = [{'n_neighbors': n_range}]
tuned_params
```

Out[73]:

```
{'n_neighbors': array([ 5, 10, 15, 20, 25, 30, 35, 40, 45, 50])}]
```

In [75]:

```
# Подбор гиперпараметра с использованием метода кросс-валидации LeavePOut(2) - 2 параметра в тестовой вы  
# остальные в обучающей, и метрики качества accuracy
```

```
lpo = LeavePOut(2)
```

```
clf_gs = GridSearchCV(KNeighborsClassifier(), tuned_params, cv = lpo, scoring = 'accuracy')
```

```
clf_gs.fit(X_train, Y_train)
```

Out[75]:

```
GridSearchCV(cv=LeavePOut(p=2), estimator=KNeighborsClassifier(),  
             param_grid=[{'n_neighbors': array([ 5, 10, 15, 20, 25, 30, 35, 40, 45, 50])}],  
             scoring='accuracy')
```

In [76]:

```
clf_gs.cv_results_
```

Out[76]:

```
{'mean_fit_time': array([0.00036827, 0.0003652 , 0.00036578, 0.00036527, 0.00037244,  
                        0.0004025 , 0.00038617, 0.00038437, 0.00040303, 0.00038359]),  
'std_fit_time': array([4.75352266e-05, 3.67897686e-05, 4.90215461e-05, 4.54180927e-05,  
                        5.98046029e-05, 8.55278823e-05, 4.55882250e-05, 4.40962847e-05,  
                        8.89617720e-05, 4.31835096e-05]),  
'mean_score_time': array([0.00061715, 0.00063268, 0.00061605, 0.00062711, 0.00063152,  
                        0.00070614, 0.00066285, 0.0006728 , 0.00070313, 0.00066211]),  
'std_score_time': array([8.77901857e-05, 1.62070956e-03, 6.98159128e-05, 8.17194101e-04,  
                        8.95293201e-05, 9.73635963e-04, 8.33577222e-05, 1.15841650e-03,  
                        6.98868543e-04, 7.86441299e-05]),  
'param_n_neighbors': masked_array(data=[5, 10, 15, 20, 25, 30, 35, 40, 45, 50],  
                                  mask=[False, False, False, False, False, False, False, False, False, False],  
                                  fill_value='?',  
                                  dtype=object),  
'params': [{'n_neighbors': 5},  
            {'n_neighbors': 10},  
            {'n_neighbors': 15},  
            {'n_neighbors': 20},  
            {'n_neighbors': 25},  
            {'n_neighbors': 30},  
            {'n_neighbors': 35},  
            {'n_neighbors': 40},  
            {'n_neighbors': 45},  
            {'n_neighbors': 50}],  
'split0_test_score': array([0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5]),  
'split1_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),  
'split2_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),  
'split3_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),  
'split4_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 0.5]),  
'split5_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),  
'split6_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),  
'split7_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),  
'split8_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),  
'split9_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),  
'split10_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),  
'split11_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),  
'split12_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),  
'split13_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),  
'split14_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),  
'split15_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),  
'split16_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),  
'split17_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),  
'split18_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),  
'split19_test_score': array([0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5]),  
'split20_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),  
'split21_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),  
'split22_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),  
'split23_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),  
'split24_test_score': array([0.5, 1., 0.5, 1., 1., 1., 1., 1., 1., 1.]),  
'split25_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),  
'split26_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),  
'split27_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),  
'split28_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),  
'split29_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),  
'split30_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),  
'split31_test_score': array([0.5, 0.5, 1., 1., 1., 1., 1., 1., 1., 1.]),  
'split32_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),  
'split33_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),  
'split34_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),  
'split35_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),  
'split36_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),  
'split37_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),  
'split38_test_score': array([1., 0.5, 1., 1., 1., 1., 1., 1., 1., 1.]),  
'split39_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),  
'split40_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),  
'split41_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),  
'split42_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),  
'split43_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),  
'split44_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),  
'split45_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),  
'split46_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),  
'split47_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),  
'split48_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.])}
```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

```
'split983_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),
'split984_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),
'split985_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),
'split986_test_score': array([1., 0.5, 1., 1., 1., 1., 1., 1., 1., 1.]),
'split987_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),
'split988_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),
'split989_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),
'split990_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),
'split991_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),
'split992_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),
'split993_test_score': array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1.]),
...}
```

In [77]:

```
# Лучшая модель
clf_gs.best_estimator_
```

Out[77]:

```
KNeighborsClassifier(n_neighbors=20)
```

In [78]:

```
# Лучшее значение метрики
clf_gs.best_score_
```

Out[78]:

```
0.9832153160241279
```

In [81]:

```
# Лучшее значение параметров
clf_gs.best_params_
```

Out[81]:

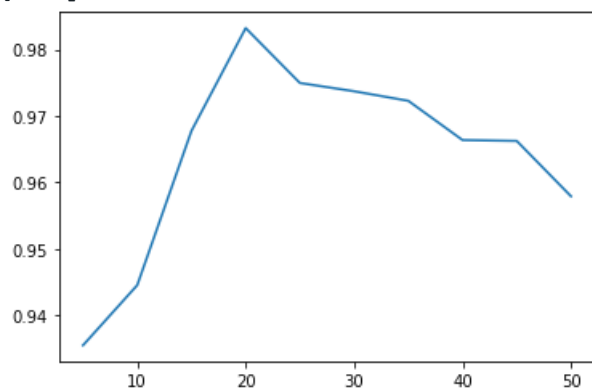
```
{'n_neighbors': 20}
```

In [82]:

```
# Изменение качества на тестовой выборке в зависимости от K-соседей
plt.plot(n_range, clf_gs.cv_results_['mean_test_score'])
```

Out[82]:

```
[<matplotlib.lines.Line2D at 0x7fc3f8553ee0>]
```



In [85]:

```
lpo = LeavePOut(2)
clf_rs = RandomizedSearchCV(KNeighborsClassifier(), tuned_params, cv = lpo, scoring = 'accuracy')
clf_rs.fit(X_train, Y_train)
```

Out[85]:

```
RandomizedSearchCV(cv=LeavePOut(p=2), estimator=KNeighborsClassifier(),
                  param_distributions=[{'n_neighbors': array([ 5, 10, 15, 20, 25, 30, 35, 40, 45, 50])}],
                  scoring='accuracy')
```

In [87]:

```
clf_rs.best_score_, clf_rs.best_params_
```

Out[87]:

```
(0.9832153160241279, {'n_neighbors': 20})
```

In [88]:

```
# Подбор гиперпараметра с использованием метода кросс-валидации KFold - деление выборок на 5 фолдов,
# и метрики качества accuracy
clf_gs1 = GridSearchCV(KNeighborsClassifier(), tuned_params, cv = 5, scoring = 'accuracy')
clf_gs1.fit(X_train, Y_train)
```

Out[88]:

```
GridSearchCV(cv=5, estimator=KNeighborsClassifier(),
            param_grid=[{'n_neighbors': array([ 5, 10, 15, 20, 25, 30, 35, 40, 45, 50])}],
            scoring='accuracy')
```

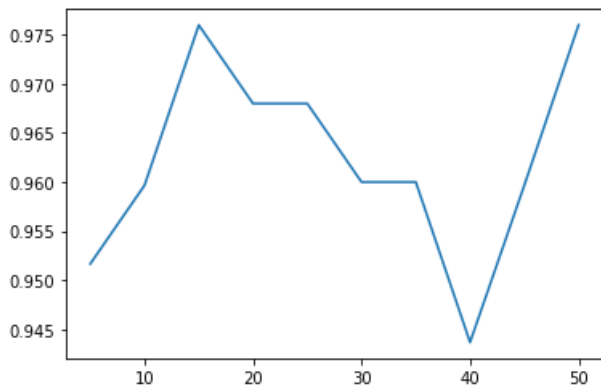
In [89]:

```
clf_gs1.best_score_, clf_gs1.best_params_
```

```
(0.976, {'n_neighbors': 15})

plt.plot(n_range, clf_gs1.cv_results_['mean_test_score'])
```

```
[<matplotlib.lines.Line2D at 0x7fc3fd244430>]
```



Сравнение метрик качества исходной и оптимальной моделей

```
clf_gs.best_estimator_.fit(X_train, Y_train)
opt_target_train = clf_gs.best_estimator_.predict(X_train)
opt_target_test = clf_gs.best_estimator_.predict(X_test)
```

```
# Точность для оптимальной модели
accuracy_score(Y_train, opt_target_train), accuracy_score(Y_test, opt_target_test)
```

```
(0.9838709677419355, 0.9814814814814815)
```

```
# Точность для исходной модели
accuracy_score(Y_train, target1), accuracy_score(Y_test, target1)
```

```
(0.967741935483871, 0.9629629629629629)
Точность у оптимальной модели выше и разница в точности между обучающей и тестовой выборками уменьшилась с 0.0048 до 0.0023
```

```
# Precision для оптимальной модели
precision_score(Y_train, opt_target_train, average='macro'), precision_score(Y_test, opt_target_test, average='macro')
```

```
(0.981981981981982, 0.9861111111111112)
```

```
# Precision для исходной модели
precision_score(Y_train, target1, average='macro'), precision_score(Y_test, target1, average='macro')
```

```
(0.9653499653499654, 0.9733333333333333)
Разница метрики между выборками уменьшилась с 0.0079 до 0.0041
```

```
# Полнота для оптимальной модели, учитывается вес классов
recall_score(Y_train, opt_target_train, average='weighted'), recall_score(Y_test, opt_target_test, average='weighted')
```

```
(0.9838709677419355, 0.9814814814814815)
```

```
# Полнота для исходной модели, учитывается вес классов
recall_score(Y_train, target1, average='weighted'), recall_score(Y_test, target1, average='weighted')
```

```
(0.967741935483871, 0.9629629629629629)
Разница полноты между выборками уменьшилась с 0.0047 до 0.0023
```