Curs 6: Optimizarea modelelor, preprocesare, pipelines

Optimizarea modelelor

In cursul anterior s-a aratat cum se poate folosi k-fold cross validation pentru estimarea performantei unui model. Totodata, s-a aratat o maniera simpla de cautare a valorilor celor mai potrivite pentru hiperparametri - in cazul respectiv. valoarea adecvata a numarului de vecini.

Vom continua aceasta idee pentru mai multi hiperparametri, apoi folosim facilitatile bibliotecii `sklearn` pentru automatizarea procesului.

K-fold cross validation (asigura ca fiecare din cele k partitii ale setului de date initial este pe rand folosit ca subset de testare:

In [1]:
```
from sklearn.model_selection import KFold
!pip install prettytable --upgrade
from prettytable import PrettyTable
```

```
Requirement already up-to-date: prettytable in c:\anaconda3\envs\data-science
\lib\site-packages
```

In [2]:
```
kf = KFold(n_splits=3)
splits = kf.split(range(30))
t = PrettyTable(['Iter', 'Train', 'Test'])
t.align = 'l'
for i, data in enumerate(splits):
    t.add_row([i+1, data[0], data[1]])
print(t)
```

```
+------+------------------------------------------------------------------+-----
-----------------------------+
| Iter | Train                                                            | Test
                             |
+------+------------------------------------------------------------------+-----
-----------------------------+
| 1    | [10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29] | [0 1
2 3 4 5 6 7 8 9]            |
| 2    | [ 0  1  2  3  4  5  6  7  8  9 20 21 22 23 24 25 26 27 28 29] | [10
11 12 13 14 15 16 17 18 19] |
| 3    | [ 0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19] | [20
21 22 23 24 25 26 27 28 29] |
+------+------------------------------------------------------------------+-----
-----------------------------+
```

Pentru calculul performantei prin k-fold CV se asigura ca esantionarea se face in mod stratificat: fiecare fold are aceeasi proportie a claselor ca si in setul originar.

Folosim k-fold cross validation pentru a face evaluarea de modele pentru diferite valori ale hiperparametrilor.

In [3]:
```
import numpy as np
import pandas as pd
print ('numpy: ', np.__version__)
print ('pandas: ', pd.__version__)
```

```
numpy:  1.14.2
pandas:  0.22.0
```

In [4]:
```
from sklearn.model_selection import cross_val_score, train_test_split
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score

from sklearn.datasets import load_iris

iris = load_iris()
X = iris.data
y = iris.target
```

Pentru k-nearest neighbors vom cauta valorile optime pentru:

- numarul de vecini, $k \in \{1, \ldots, 31\}$
- putere corespunzatoare metricii Minkowski:

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^{n} \left| x_i - y_i \right|^p \right)^{1/p}$$

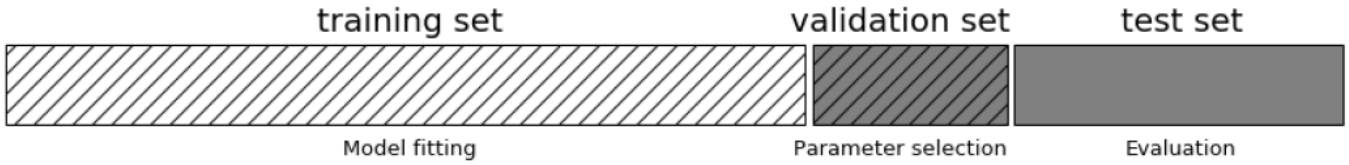
In [5]:
```
best_score = 0
for k in range(1, 31):
    for p in [1, 2, 3, 4.7]:
        model = KNeighborsClassifier(n_neighbors=k, p=p)
        score = np.mean(cross_val_score(model, X, y, cv=10))
        if score >= best_score:
            best_score = score
            best_params = {'n_neighbors':k, 'p':p}
print('Best score:', best_score)
print('Best params:', best_params)
model = KNeighborsClassifier(n_neighbors=best_params['n_neighbors'], p=best_pa
rams['p'])
model.fit(X, y)
y_predicted = model.predict(X)
print('Accuracy on whole set:', accuracy_score(y, y_predicted))
```

```
Best score: 0.9800000000000001
Best params: {'n_neighbors': 20, 'p': 2}
Accuracy on whole set: 0.98
```

Pentru procesul de mai sus urmatoarele comentarii sunt necesare:

1. strategia implementata se numeste grid search: se cauta peste toate combinatiile de 30*4 variante si sa retine cea mai buna; este consumatoare de resurse, dar o prima varianta de lucru acceptabila
2. am dori sa avem o modalitate automatizata de considerare a tuturor combinatiilor de parametri din multimea de valori candidat. Codul devine greu de scris cand sunt 4 hiperparametri, fiecare cu multimea proprie de valori candidat
3. estimarea efectuata in final este de cele mai multe ori optimista: optimizarea parametrilor s-a facut peste niste date, care date in final sunt cele folosite pentru evaluarea finala; am ajuns practic sa facem evaluare pe setul de antrenare, ceea ce e o idee proasta. Estimarea finala a performantelor modelului trebuie facuta peste un set de date aparte, care nu a fost folosit nici pentru antrenare, nici pentru validarea modelelor candidat.

Pentru ultimul punct se recomanda ca setul sa fie impartit ca mai jos:



Ca atare, va trebui sa rescriem codul astfel:

In [6]:
```
X_trainval, X_test, y_trainval, y_test = train_test_split(X, y, test_size=1/
5)
best_score = 0
for k in range(1, 31):
    for p in [1, 2, 3, 4.7]:
        model = KNeighborsClassifier(n_neighbors=k, p=p)
        score = np.mean(cross_val_score(model, X_trainval, y_trainval, cv=10
))
        if score >= best_score:
            best_score = score
            best_params = {'n_neighbors':k, 'p':p}
print('Best score:', best_score)
print('Best params:', best_params)

model = KNeighborsClassifier(n_neighbors=best_params['n_neighbors'], p=best_
params['p'])
model.fit(X_trainval, y_trainval)
y_predicted = model.predict(X_test)
print(accuracy_score(y_test, y_predicted))
```

```
Best score: 0.9825757575757577
Best params: {'n_neighbors': 15, 'p': 3}
0.9666666666666667
```

Desigur, si implementarea de mai sus e criticabila: s-a facut evaluare pe un singur set de testare, anume cel rezultat dupa impartirea initiala in partitiile *_trainvalid si *_test. Este totusi o estimare mai corect facuta decat cea precedenta. In realitate, acest stil de lucru este frecvent intalnit: exista un set de testare unic, dar necunoscut la inceput. Singurele date disponibile sunt impartite in *training set* si *validation set* (eventual mai multe) pentru a obtine un model care se spera ca generalizeaza bine = se comporta bine pe setul de testare.

Varianta anterioara se numeste **grid search with cross validation**. Exista clasa `sklearn.model_selection.GridSearchCV` care automatizeaza procesul:

In [7]:
```
from sklearn.model_selection import GridSearchCV
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=1/5)
parameter_grid = {'n_neighbors': list(range(1, 10)), 'p': [1, 2, 3, 4.7]}
grid_search = GridSearchCV(estimator = KNeighborsClassifier(), param_grid=para
meter_grid, scoring='accuracy', cv=5,
                           return_train_score=True)
grid_search.fit(X_train, y_train)
```

Out[7]:
```
GridSearchCV(cv=5, error_score='raise',
       estimator=KNeighborsClassifier(algorithm='auto', leaf_size=30, metric
='minkowski',
           metric_params=None, n_jobs=1, n_neighbors=5, p=2,
           weights='uniform'),
       fit_params=None, iid=True, n_jobs=1,
       param_grid={'n_neighbors': [1, 2, 3, 4, 5, 6, 7, 8, 9], 'p': [1, 2, 3,
4.7]},
       pre_dispatch='2*n_jobs', refit=True, return_train_score=True,
       scoring='accuracy', verbose=0)
```

In [8]:
```
y_estimated = grid_search.predict(X_test)
print(accuracy_score(y_test, y_estimated))
```

```
0.9666666666666667
```

In codul anterior denumirile cheilor din dictionarul `parameter_grid` nu sunt intamplatoare: ele coincid cu numele parametrilor modelului vizat. Instantierea `estimator = KNeighborsClassifier()` se face cu valorile implicite ale parametrilor, apoi insa se ruleaza metode de tip `set_` care seteaza parametrii dati in dictionarul `parameter_grid`.

Pentru cei interesati, valorile de performanta pentru fiecare fold se pot inspecta. Pentru ca acestea sa fie disponibile, este obligatorie setarea parametrului `return_train_score=True` din clasa `GridSearchCV`.

In [9]:
```
df_grid_search = pd.DataFrame(grid_search.cv_results_)
df_grid_search.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 36 entries, 0 to 35
Data columns (total 22 columns):
mean_fit_time         36 non-null float64
mean_score_time       36 non-null float64
mean_test_score       36 non-null float64
mean_train_score      36 non-null float64
param_n_neighbors     36 non-null object
param_p               36 non-null object
params                36 non-null object
rank_test_score       36 non-null int32
split0_test_score     36 non-null float64
split0_train_score    36 non-null float64
split1_test_score     36 non-null float64
split1_train_score    36 non-null float64
split2_test_score     36 non-null float64
split2_train_score    36 non-null float64
split3_test_score     36 non-null float64
split3_train_score    36 non-null float64
split4_test_score     36 non-null float64
split4_train_score    36 non-null float64
std_fit_time          36 non-null float64
std_score_time        36 non-null float64
std_test_score        36 non-null float64
std_train_score       36 non-null float64
dtypes: float64(18), int32(1), object(3)
memory usage: 6.1+ KB
```

In [10]:
```
df_grid_search.head()
```

Out[10]:

	mean_fit_time	mean_score_time	mean_test_score	mean_train_score	param_n_neigl
0	0.002407	0.003008	0.958333	1.000000	1
1	0.003611	0.005013	0.958333	1.000000	1
2	0.002605	0.003811	0.958333	1.000000	1
3	0.001606	0.004811	0.958333	1.000000	1
4	0.001202	0.002207	0.941667	0.974954	2

5 rows × 22 columns

Pentru situatia in care se doreste evaluarea nu doar pe un singur set de testare, ci in stil cross-validation, se poate face un *nested cross-validation*:

In [11]:
```
scores = cross_val_score(GridSearchCV(estimator = KNeighborsClassifier(), para
m_grid=parameter_grid,
                                      scoring='accuracy', cv=5), X, y, cv=10)
```

In [12]:
```
print(scores.mean())
```

```
0.9666666666666668
```

Metode de preprocesare

Uneori, inainte de aplicarea vreunui model, este nevoie ca datele de intrare sa fie supuse unor transformari. De exemplu, daca pentru algoritmul k-NN vreuna din trasaturi (fie ea *F*) are valori de ordinul sutelor si celelalte de ordinul unitatilor, atunci distanta dintre doi vectori ar fi dominata de diferenta pe dimensiunea *F*; celelalte dimensiuni nu ar conta prea mult.

Intr-o astfel de situatie se recomanda sa se faca o scalare in prealabil a datelor la intervale comparbile, de ex [0, 1].

In modulul `sklearn.preprocessing` se afla clasa `MinMaxScaler` care permite scalarea independenta a trasaturilor. Il vom demonstra pe un set de date care are trasaturi cu marimi disproportionate.

In [13]:
```
from sklearn.datasets import load_breast_cancer
medical = load_breast_cancer()
X, y = medical.data, medical.target
```

In [14]:
```
def print_ranges(X):
    for col_index in range(X.shape[1]):
        column = X[:, col_index]
        print(np.min(column), np.max(column))

print_ranges(X)
```

```
6.981 28.11
9.71 39.28
43.79 188.5
143.5 2501.0
0.05263 0.1634
0.01938 0.3454
0.0 0.4268
0.0 0.2012
0.106 0.304
0.04996 0.09744
0.1115 2.873
0.3602 4.885
0.757 21.98
6.802 542.2
0.001713 0.03113
0.002252 0.1354
0.0 0.396
0.0 0.05279
0.007882 0.07895
0.0008948 0.02984
7.93 36.04
12.02 49.54
50.41 251.2
185.2 4254.0
0.07117 0.2226
0.02729 1.058
0.0 1.252
0.0 0.291
0.1565 0.6638
0.05504 0.2075
```

In [15]:
```
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
scaler.fit(X)
X = scaler.transform(X)

print_ranges(X)
```

```
0.0 1.0
0.0 1.0
0.0 1.0
0.0 0.9999999999999999
0.0 1.0
0.0 0.9999999999999999
0.0 0.9999999999999999
0.0 0.9999999999999999
0.0 1.0
0.0 1.0
0.0 0.9999999999999999
0.0 1.0
0.0 1.0
0.0 1.0000000000000002
0.0 1.0
0.0 0.9999999999999999
0.0 1.0
0.0 0.9999999999999999
0.0 1.0
0.0 1.0
0.0 0.9999999999999999
0.0 1.0
0.0 1.0
0.0 1.0
0.0 0.9999999999999998
0.0 1.0
0.0 1.0
0.0 1.0
0.0 0.9999999999999999
0.0 1.0
```

De mentionat ca secventa `fit` si `transform` se poate apela intr-un singur pas:

In [16]:
```
X, y = medical.data, medical.target
X = scaler.fit_transform(X)
print_ranges(X)
```

```
0.0 1.0
0.0 1.0
0.0 1.0
0.0 0.9999999999999999
0.0 1.0
0.0 0.9999999999999999
0.0 0.9999999999999999
0.0 0.9999999999999999
0.0 1.0
0.0 1.0
0.0 0.9999999999999999
0.0 1.0
0.0 1.0
0.0 1.0000000000000002
0.0 1.0
0.0 0.9999999999999999
0.0 1.0
0.0 0.9999999999999999
0.0 1.0
0.0 1.0
0.0 0.9999999999999999
0.0 1.0
0.0 1.0
0.0 1.0
0.0 0.9999999999999998
0.0 1.0
0.0 1.0
0.0 1.0
0.0 0.9999999999999999
0.0 1.0
```

De regula, setul de date se imparte in doua (in modul naiv): set de antranare si set de testare. Se presupune ca setul de testare este cunoscut mult mai tarziu decat cel de antrenare. Ca atare, doar cel de antrenare se trece prin preprocesor, iar valorile 'invatate' via `fit` se pastreaza (obiectul de tip `MinMaxScaler` are stare). ele vor fi folosite pentru scalarea setului de test:

In [17]:
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=1/3)
scaler = MinMaxScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
print_ranges(X_test)
```

```
0.03659670238269972 1.033758990165859
0.07372336827866077 0.8150152181264796
0.02885891971210963 1.011180211026483
0.014973488865323448 0.9991516436903501
-0.09792843691148768 0.6917434830012885
0.02183915097233299 0.8957119195141405
0.0 0.7823336457357076
0.0 0.9169980119284297
0.0722222222222223 0.7621212121212122
0.00610783487784361 0.9644060657118789
-0.0011921400970155386 1.1340129902162293
0.0005746110325318271 0.7263967468175389
0.005108142849158894 1.1861062985525066
-0.0009796699783500406 0.9689646533576132
0.03824319271169731 0.6818166366386782
0.01122059662931475 0.6866644636044102
0.0 0.7671717171717172
0.0 0.6518279977268423
0.02331569764169529 0.5946136095007598
-0.002545745487796006 0.7260292951229059
0.03678406261117046 0.8210601209533975
0.07462686567164184 0.8899253731343283
0.033667015289606084 0.8032770556302604
0.014009044435705859 0.7269465198584348
0.11952679795138145 1.0923321070475367
0.015503875968992255 0.8834783789814787
0.0 1.07008547008547
0.0 0.9852233676975946
0.016361127537946 0.7871082199881728
0.0011150465695920486 0.6136691591237047
```

Se remarca faptul ca, folosindu-se parametrii de scalare din setul de antrenare, nu se poate garanta ca setul de testare este cuprins de asemenea in hipercubul unitate $[0, 1]^{X.shape[1]}$

Exista si alte metode de preprocesare in modulul `sklearn.preprocessing` (http://scikit-learn.org/stable/modules/preprocessing.html).

Pipelines

Se prefera inlantuirea intr-un proces a pasilor: preprocesare si aplicare de model. Exemplificam pentru cazul simplu in care exista un set de antrenare si unul de testare:

In [23]:
```
X, y = medical.data, medical.target
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=1/3)
```

In [25]:
```
from sklearn.pipeline import Pipeline
pipe = Pipeline([('scaler', MinMaxScaler()), ('knn', KNeighborsClassifier())])
pipe.fit(X_train, y_train)
y_predicted = pipe.predict(X_test)
print(accuracy_score(y_test, y_predicted))
```

```
0.9842105263157894
```

Pentru cazul in care se vrea k-fold cross validation pentru determinarea valorilor optime pentru hiperparametri, urmata de testare pe un set de testare:

In [29]:
```
X_trainval, X_test, y_trainval, y_test = train_test_split(X, y, test_size=1/3)
parameter_grid = {'knn__n_neighbors': list(range(1, 10)), 'knn__p': [1, 2, 3,
4.7]}
grid = GridSearchCV(pipe, param_grid = parameter_grid, scoring = 'accuracy', c
v=5)
grid.fit(X_trainval, y_trainval)
```

Out[29]:
```
GridSearchCV(cv=5, error_score='raise',
       estimator=Pipeline(memory=None,
     steps=[('scaler', MinMaxScaler(copy=True, feature_range=(0, 1))), ('kn
n', KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
           metric_params=None, n_jobs=1, n_neighbors=5, p=2,
           weights='uniform'))]),
       fit_params=None, iid=True, n_jobs=1,
       param_grid={'knn__n_neighbors': [1, 2, 3, 4, 5, 6, 7, 8, 9], 'knn__p':
[1, 2, 3, 4.7]},
       pre_dispatch='2*n_jobs', refit=True, return_train_score='warn',
       scoring='accuracy', verbose=0)
```

In [31]:
```
y_predicted = grid.predict(X_test)
print(accuracy_score(y_test, y_predicted))
```

```
0.9631578947368421
```