4_Evaluation_students

December 25, 2020

1 Hands-On Machine Learning

1.1 Session 4: Evaluation, Validation and Model Assessment

by Laxmi Gupta, Christoph Haarburger

1.1.1 Goals of this Session

In this session you will… * learn to assess a model's performance * get to know the most common evaluation measures * learn how to tune a model's hyperparameters

1.1.2 Iris Dataset

You already know the Iris dataset from you first session. Anyway, here's a quick recap: The dataset consists of three species (=classes) of the Iris plant (Setosa, Versicolor and Virginica). For each sample, four features have been measured (sepal length, sepal width, petal length, petal width).

The goal ist to train a model that can classify a given sample into one of the three species. Based on your knowledge of classifiers from the 1st session, today we'll focus more on the evaluation of performance and tuning of hyperparameters rather than the classifier itself.

[1]:
```
%matplotlib inline
```

[2]:
```
import seaborn as sns
from evaluation import load_iris
data = load_iris()
```

Before training any kind of model it is always a good idea to make ourselves familiar with the raw data and explore basic properties of the dataset. The dataset has already been loaded as a `pandas.DataFrame` which is a very handy data type for our exploration.

[3]:
```
data[:15]
```

[3]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa

3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa
5	5.4	3.9	1.7	0.4	setosa
6	4.6	3.4	1.4	0.3	setosa
7	5.0	3.4	1.5	0.2	setosa
8	4.4	2.9	1.4	0.2	setosa
9	4.9	3.1	1.5	0.1	setosa
10	5.4	3.7	1.5	0.2	setosa
11	4.8	3.4	1.6	0.2	setosa
12	4.8	3.0	1.4	0.1	setosa
13	4.3	3.0	1.1	0.1	setosa
14	5.8	4.0	1.2	0.2	setosa

You my find it interesting to look at the raw numbers, but with our goal of building a classifier in mind we particularly want to explore statistical properties of the dataset.

[4]:
```
data.describe()
```

[4]:

	sepal_length	sepal_width	petal_length	petal_width
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.057333	3.758000	1.199333
std	0.828066	0.435866	1.765298	0.762238
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

Task: Find out if the prevalence of the three classes is equal!

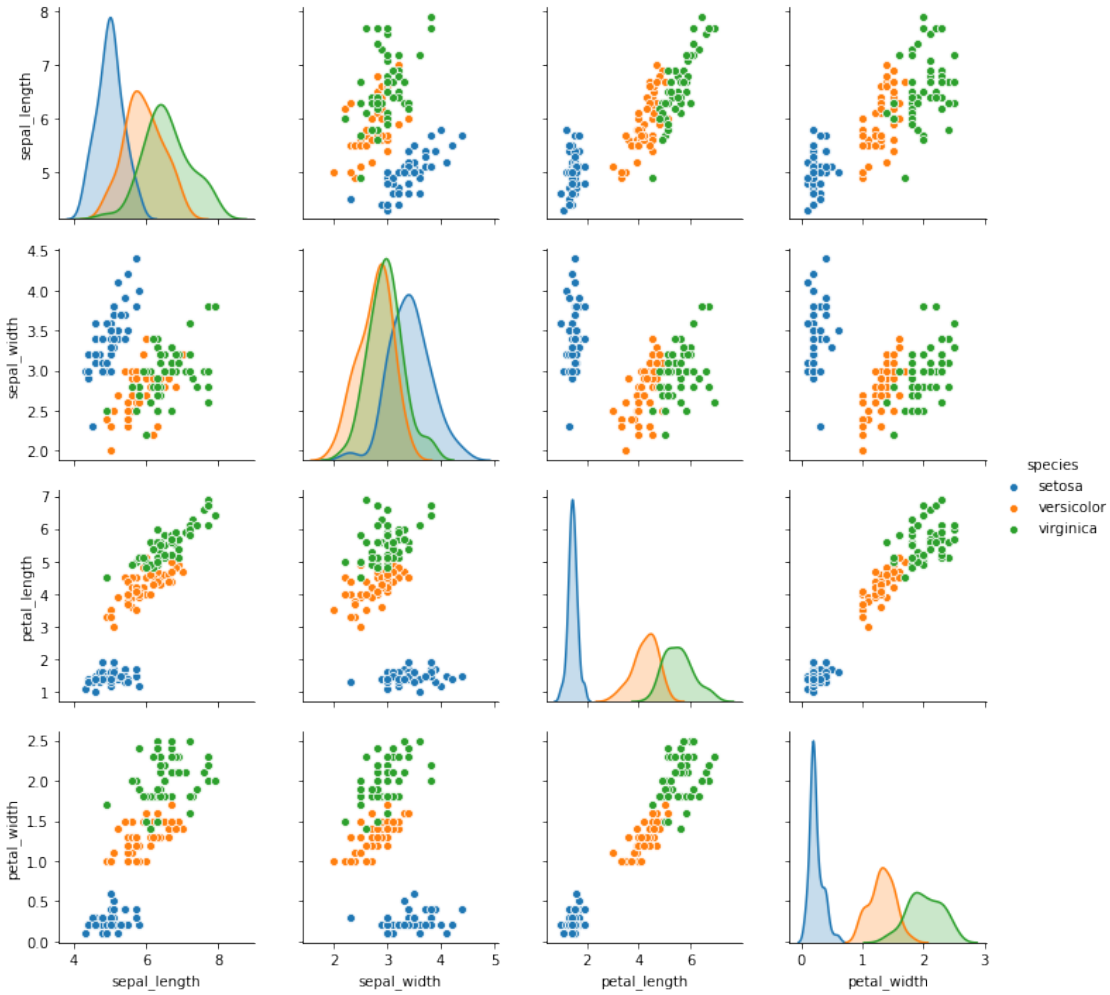
[6]:
```
import numpy as np
y = data['species'].values # turn pandas.DataFrame into np.ndarray
occ = np.unique(y, return_counts=True)
print(occ)
```

```
(array(['setosa', 'versicolor', 'virginica'], dtype=object), array([50, 50,
50]))
```

We're lucky, all classes are represented in equal ratios .

If the number of features is relatively small as in our case, it is also a good idea to plot features pairswise against each other and inspect the feature scores for all classes as histograms.

[7]:
```
_ = sns.pairplot(data, hue='species')
```



Task: Can the classification problem be solved by a linear classifier?

Answer: Various features have a clear boundary. Therefore, we can easily separate the samples into three classes using two hyperplanes defined through a linear classifier.

Now that we have gained some intuitution on the problem to be solved und the underlying data, let's get to what we are actually up to: Classification! As a first shot, it is always a good idea to utilize a classifier that is easy to understand and easy to interpret. K-Nearest-Neighbors does the job...

[9]:
```
from sklearn.neighbors import KNeighborsClassifier as KNN

X = data[['sepal_length', 'sepal_width', 'petal_length', 'petal_width']].
 ↪as_matrix() # turn pandas.DataFrame into np.ndarray

model = KNN(n_neighbors=2)
model.fit(X, y)
```

```
y_pred = model.predict(X)
```

```
/opt/conda/lib/python3.7/site-packages/ipykernel_launcher.py:3: FutureWarning:
Method .as_matrix will be removed in a future version. Use .values instead.
  This is separate from the ipykernel package so we can avoid doing imports
until
```

Task: To interpret the result, implement the accuracy measure and calculate the accuracy for our predictions. Accuracy is defined as $\frac{\text{true}_{\text{positive}}+\text{true}_{\text{negative}}}{\text{positive}+\text{negative}}$.

[10]:
```
from __future__ import division
def accuracy_score(y_true, y_pred):
    true_combined = len(np.where(y_true == y_pred)[0])
    accuray_score = true_combined / len(y_true)
    return accuray_score

print(accuracy_score(y, y_pred))
```

```
0.98
```

Task: Why is the classification accuracy nearly perfect?

Answer: As the classifier is trained and tested both on the same data set.

1.2 Holdout Sets

A much better way to evaluate performance is to split all available data into a training and test set, where the training set is used for training the model's parameters and the test set is utilized for assessing performance.

Task: Implement the same pipeline as in the last code cell but with a separate training and test set (50/50)!

[11]:
```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.5,␣
 ↪random_state=0)

model2 = KNN(n_neighbors=2)
model2.fit(X_train, y_train)
y_pred = model2.predict(X_test)
print(accuracy_score(y_test, y_pred))
```

```
0.8933333333333333
```

That looks a lot more realistic!

Task: When using `train_test_split`, the samples are *randomly* drawn to one of the sets. What happens, if we use a different split? Determine the 95% confidence intervals for random 100 splits. *Hint:* `np.percentile`

[23]:
```
acc_scores = []
for num_iter in range(1, 101):
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.5)
    model2.fit(X_train, y_train)
    y_pred = model2.predict(X_test)
    acc_scores.append(accuracy_score(y_test, y_pred))

np.percentile(np.array(acc_scores), 95)
```

[23]:
```
0.9733333333333334
```

Task: What properties should a 'good' split have with respect to the dataset?

Answer: the dataset should be large enough, randomly split the data, Test set should avoid having high variance compared to the training set.

1.3 Cross Validation

One disadvantage of the holdout set method is that half the dataset does not contribute to the training of the model. Especially if the initial set of training data is small this is suboptimal since statistical learning methods tend to perform worse when trained on fewer observations.

In cross validation we can overcome this by partitioning the data set in `k` "folds" of approximately equal size. The first fold is used as a test set and the model is trained on the remaining `k` - 1 folds. The test error is then co puted on the test set and the process is repeated `k` times, resulting in `k` estimates of the test error.

Task: Implement the `CrossValidation` class as sketched in the following cell! Hint: This is advanced task that is not mandatory to be solved.

[11]:
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.50,␣
↪random_state=0)

class CrossValidation(object):

    def __init__(self, model, k):
        self.model = model
        self.k = k
        self.k_size = int(len(X_train) / k)


    def evaluate(self, X_train, y_train, X_test, y_test):

        acc_val_mean = []
        acc_test_mean = []

        for k_idx in range(0, len(X_train), self.k_size):
```

```
            X_train_k = np.concatenate((X_train[:k_idx], X_train[k_idx + k_size:
↪]))
            y_train_k = np.concatenate((y_train[:k_idx], y_train[k_idx + k_size:
↪]))
            X_valid_k = X_train[k_idx:(k_idx + k_size)]
            y_valid_k = y_train[k_idx:(k_idx + k_size)]

            # print(y_train_k.shape)
            # print(X_train_k.shape)

            self.model.fit(X_train_k, y_train_k)
            y_valid_k_pred = model.predict(X_valid_k)

            acc_val = accuracy_score(y_valid_k, y_valid_k_pred)
            acc_val_mean.append(acc_val)

            y_test_k_pred = model.predict(X_test)
            acc_test = accuracy_score(y_test, y_test_k_pred)
            acc_test_mean.append(acc_test)

        mu_val = sum(acc_val_mean) / len(acc_val_mean)
        mu_test = sum(acc_test_mean) / len(acc_test_mean)

        return mu_val, mu_test
```

Task: Use the `CrossValidation` class to evaluate the test error of the same KNN model as above but with 10 fold cross validation. What's the mean test error? If you did not solve the previous task on your own, you can use `KFold` from `scikit-learn`.

[12]:
```
from sklearn.model_selection import KFold, cross_val_score

k_fold = KFold(n_splits=10)
cross_val_val = []

model3 = KNN(n_neighbors=2)

for train_indices, test_indices in k_fold.split(X):
    model3.fit(X[train_indices], y[train_indices])
    y_pred = model3.predict(X[test_indices])
    acc = accuracy_score(y[test_indices], y_pred)
    cross_val_val.append(1 - acc)

np.mean(cross_val_val)
```

[12]:
```
0.06666666666666665
```

Task: Why did the accuracy improve compared to the holdout split? What's a disadvantage of the cross validation approach?

Answer: By using k-fold cross-validation a larger training set is used.

With less validation set the variance increase, more computation and increased training time.

1.4 Hyperparameter Optimization

KNN is well known for being easy to use and interpret, howerever to achieve top performance it is not the best choice. In the first Session you got to know the Support Vector Machine (SVM) classifier. In contrast to KNN it has quite a few hyperparameters that need to optimized to achieve top performance. In the next few steps we will optimize SVM hyperparameters by a grid search.

SVMs are not scale invariant, i.e. all features need to be scaled to the same range, for example [-1, 1] or [0, 1]. The most common approach to this is the subtraction of the mean and division by the standard deviation of each feature.

It is one of the most common errors in machine learning to carry out scaling on the whole dataset rather than on training- and test set independently. If this is not considered, information about the distribution of the test set (mean and standard deviation) leaks into the training set and the cross validation is flawed.

A very handy way to overcome this is using the `sklearn.pipeline.Pipeline` class for combining consecutive processing pipelines such as scaling and classification.

Task: Combine an SVM classifier (default parameters) and a `StandardScaler` in a `Pipeline`.

[23]:
```
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler
from sklearn.svm import SVC

X_train, X_test, y_train, y_test = train_test_split(X, y,random_state=0,␣
↪test_size=0.5)

pipe = Pipeline([('scaler', StandardScaler()), ('svc', SVC())])

pipe.fit(X_train, y_train)
pipe.score(X_test, y_test)
```

[23]:
```
0.92
```

If you now call `pipe.fit(X_train, y_train)`, the pipeline calculcates mean and standard deviation based on the training data, applies the scaling and trains the SVM.

It's time to tune the SVM's hyperparameters, namely the `C` parameter, the type of kernel and the `gamma` parameter. The naive approach to hyperparameter tuning lies in exhaustively testing all possible combinations of parameters, which is called grid search.

Task: Implement a parameter grid for our three parameters to tune and perform the grid search in a 10-fold cross validation!

Hint: Keep in mind that exhaustive grid search results in high computation times. Use parallelization with the `n_jobs` argument and keep track of time with the `%time` magic command.

[25]:
```
%time
from sklearn.model_selection import GridSearchCV
from sklearn.preprocessing import StandardScaler

k_fold = KFold(n_splits=10)
parameters = {'svc__kernel':('linear', 'rbf', 'poly'), 'svc__C':[0.1, 0.5, 1,␣
↪5, 10], 'svc__gamma':('scale', 'auto')}

gcv = GridSearchCV(pipe, parameters, n_jobs = -1, cv = 10)
gcv.fit(X_train, y_train)
y_pred = gcv.predict(X_test)
acc = accuracy_score(y_test, y_pred)

print(acc)
```

```
CPU times: user 2 µs, sys: 1 µs, total: 3 µs
Wall time: 4.77 µs
0.9466666666666667

/opt/conda/lib/python3.7/site-packages/sklearn/model_selection/_search.py:813:
DeprecationWarning: The default of the `iid` parameter will change from True to
False in version 0.22 and will be removed in 0.24. This will change numeric
results when test-set sizes are unequal.
  DeprecationWarning)
```

Take a closer look at the grid search results. What mean test score can you achieve?

[26]:
```
from evaluation import show_grid_search_results
show_grid_search_results(gcv)[:5]
```

[26]:

	mean_fit_time	std_fit_time	mean_score_time	std_score_time	param_svc__C	\
15	0.000915	0.000162	0.000336	0.000057	1	
6	0.000877	0.000090	0.000309	0.000033	0.5	
21	0.001120	0.000030	0.000409	0.000005	5	
9	0.000783	0.000043	0.000312	0.000034	0.5	
12	0.000857	0.000052	0.000295	0.000013	1	

	param_svc__gamma	param_svc__kernel	\
15	auto	linear	
6	scale	linear	
21	auto	linear	
9	auto	linear	
12	scale	linear	

	params	split0_test_score	\
15	{'svc__C': 1, 'svc__gamma': 'auto', 'svc__kern…	1.0	
6	{'svc__C': 0.5, 'svc__gamma': 'scale', 'svc__k…	1.0	
21	{'svc__C': 5, 'svc__gamma': 'auto', 'svc__kern…	1.0	

```
9   {'svc__C': 0.5, 'svc__gamma': 'auto', 'svc__ke…                1.0
12  {'svc__C': 1, 'svc__gamma': 'scale', 'svc__ker…                1.0

    split1_test_score  …  split3_test_score  split4_test_score  \
15                1.0  …                1.0                1.0
6                 1.0  …                1.0                1.0
21                1.0  …                1.0                1.0
9                 1.0  …                1.0                1.0
12                1.0  …                1.0                1.0

    split5_test_score  split6_test_score  split7_test_score  \
15                1.0           0.857143                1.0
6                 1.0           0.857143                1.0
21                1.0           0.857143                1.0
9                 1.0           0.857143                1.0
12                1.0           0.857143                1.0

    split8_test_score  split9_test_score  mean_test_score  std_test_score  \
15           0.857143                1.0         0.973333        0.055663
6            0.857143                1.0         0.973333        0.055663
21           0.857143                1.0         0.973333        0.055663
9            0.857143                1.0         0.973333        0.055663
12           0.857143                1.0         0.973333        0.055663

    rank_test_score
15                1
6                 1
21                1
9                 1
12                1

[5 rows x 21 columns]
```

As the number of hyperparameter is usually much higher than here, a naive grid search is often not feasible in practice. A good alternative lies in a randomized search, in which the parameter space is sampled coarsely.

Task: Implement the same workflow as above withe the same parameter range but using `RandomizedSarchCV`. Can you achieve comparable performance in shorter time?

[27]:
```
from sklearn.model_selection import RandomizedSearchCV
from scipy.stats import uniform
%time

gcv = RandomizedSearchCV(pipe, parameters, random_state=0, cv = 10)
gcv.fit(X_train, y_train)
y_pred = gcv.predict(X_test)
acc = accuracy_score(y_test, y_pred)
```

```
print(acc)
```

```
CPU times: user 2 µs, sys: 0 ns, total: 2 µs
Wall time: 5.96 µs
0.9466666666666667
```

```
/opt/conda/lib/python3.7/site-packages/sklearn/model_selection/_search.py:813:
DeprecationWarning: The default of the `iid` parameter will change from True to
False in version 0.22 and will be removed in 0.24. This will change numeric
results when test-set sizes are unequal.
  DeprecationWarning)
```

You probably found out that it is possible to achieve literally the same mean test score using a randomizes search in just the fraction of the time needed for al full grid search.

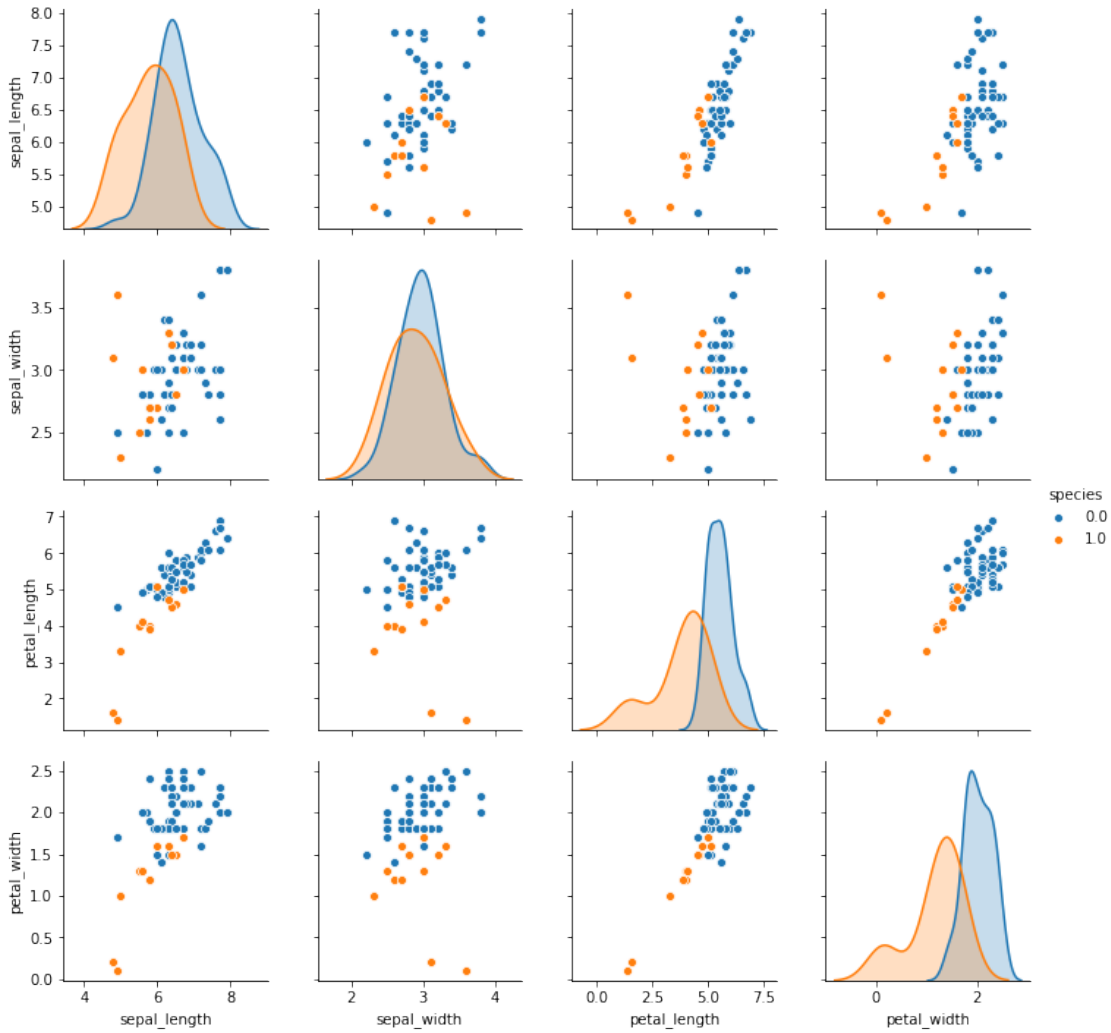
Task: Is the mean test score a good estimation of the performance we can expect on unseen data?

Answer: Depending on the number of iterations through the randomized grid search the accuracy will be similar to a full grid search that finds the optimal parameters.

1.5 Evaluation Measures

To make this more interesting, the iris dataset was slightly modified to better resemble real-world datasets. The following modified Iris dataset is smaller, there are only two classes present and the prevalence for the two classes is imbalanced.

[28]:
```
data = load_iris(imbalanced_binary=True)
_ = sns.pairplot(data, hue='species', vars=data.columns[:-1])
```



[29]:
```
data
```

[29]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	4.8	3.1	1.6	0.2	1.0
1	4.9	3.6	1.4	0.1	1.0
2	5.8	2.6	4.0	1.2	1.0
3	6.7	3.0	5.0	1.7	1.0
4	6.0	2.7	5.1	1.6	1.0
5	5.5	2.5	4.0	1.3	1.0
6	5.6	3.0	4.1	1.3	1.0
7	5.8	2.7	3.9	1.2	1.0
8	6.5	2.8	4.6	1.5	1.0
9	6.4	3.2	4.5	1.5	1.0
10	5.0	2.3	3.3	1.0	1.0
11	6.3	3.3	4.7	1.6	1.0

12	6.3	2.7	4.9	1.8	0.0
13	7.6	3.0	6.6	2.1	0.0
14	5.7	2.5	5.0	2.0	0.0
15	6.5	3.2	5.1	2.0	0.0
16	6.5	3.0	5.2	2.0	0.0
17	6.7	3.3	5.7	2.5	0.0
18	6.3	2.5	5.0	1.9	0.0
19	6.5	3.0	5.8	2.2	0.0
20	7.7	2.6	6.9	2.3	0.0
21	7.1	3.0	5.9	2.1	0.0
22	5.9	3.0	5.1	1.8	0.0
23	6.2	3.4	5.4	2.3	0.0
24	7.2	3.2	6.0	1.8	0.0
25	5.8	2.7	5.1	1.9	0.0
26	6.8	3.2	5.9	2.3	0.0
27	6.9	3.1	5.4	2.1	0.0
28	6.4	3.1	5.5	1.8	0.0
29	5.8	2.8	5.1	2.4	0.0
..	…	…	…	…	…
32	4.9	2.5	4.5	1.7	0.0
33	6.3	3.3	6.0	2.5	0.0
34	6.7	3.1	5.6	2.4	0.0
35	6.3	2.8	5.1	1.5	0.0
36	7.7	3.0	6.1	2.3	0.0
37	5.8	2.7	5.1	1.9	0.0
38	6.1	2.6	5.6	1.4	0.0
39	6.4	2.8	5.6	2.2	0.0
40	7.2	3.0	5.8	1.6	0.0
41	5.6	2.8	4.9	2.0	0.0
42	6.2	2.8	4.8	1.8	0.0
43	6.9	3.1	5.1	2.3	0.0
44	7.3	2.9	6.3	1.8	0.0
45	6.4	3.2	5.3	2.3	0.0
46	7.4	2.8	6.1	1.9	0.0
47	7.7	2.8	6.7	2.0	0.0
48	6.3	2.9	5.6	1.8	0.0
49	6.7	2.5	5.8	1.8	0.0
50	7.7	3.8	6.7	2.2	0.0
51	6.0	2.2	5.0	1.5	0.0
52	6.1	3.0	4.9	1.8	0.0
53	6.8	3.0	5.5	2.1	0.0
54	6.3	3.4	5.6	2.4	0.0
55	6.0	3.0	4.8	1.8	0.0
56	6.7	3.3	5.7	2.1	0.0
57	6.4	2.8	5.6	2.1	0.0
58	6.9	3.2	5.7	2.3	0.0
59	6.7	3.0	5.2	2.3	0.0

```
60           7.9          3.8           6.4          2.0       0.0
61           6.5          3.0           5.5          1.8       0.0

[62 rows x 5 columns]
```

Task: Solve the binary classification problem by a KNN classifier, optimize the number of neighbors by 5 fold cross validation and evaluate accuracy on the test set using the best `n_neighbors`.

[47]:
```
X_train, X_test, y_train, y_test = train_test_split(data[['sepal_length',␣
↪'sepal_width', 'petal_length', 'petal_width']].as_matrix(),␣
↪data[['species']].values.flatten(), test_size=0.5, random_state=1)


k_fold = KFold(n_splits=5)

pipe = Pipeline([('scaler', StandardScaler()), ('knn', KNN())])
parameters = {'knn__n_neighbors':[1, 2, 3, 4, 5, 6, 7, 8, 9, 10]}

gcv = GridSearchCV(pipe, parameters, cv = 5)
gcv.fit(X_train, y_train)
y_pred = gcv.predict(X_test)
acc = accuracy_score(y_test, y_pred)
print(acc)
show_grid_search_results(gcv)[:]
```

```
0.9032258064516129

/opt/conda/lib/python3.7/site-packages/ipykernel_launcher.py:1: FutureWarning:
Method .as_matrix will be removed in a future version. Use .values instead.
  """Entry point for launching an IPython kernel.
/opt/conda/lib/python3.7/site-packages/sklearn/model_selection/_search.py:813:
DeprecationWarning: The default of the `iid` parameter will change from True to
False in version 0.22 and will be removed in 0.24. This will change numeric
results when test-set sizes are unequal.
  DeprecationWarning)
```

[47]:

	mean_fit_time	std_fit_time	mean_score_time	std_score_time	\
2	0.000758	0.000045	0.000899	0.000032	
3	0.001043	0.000241	0.001234	0.000158	
4	0.000740	0.000026	0.000899	0.000074	
0	0.000828	0.000026	0.000967	0.000079	
1	0.000875	0.000118	0.001031	0.000154	
5	0.000759	0.000097	0.000863	0.000024	
6	0.000695	0.000006	0.000877	0.000070	
7	0.000711	0.000014	0.000922	0.000135	
8	0.000757	0.000072	0.000925	0.000059	
9	0.000715	0.000024	0.000882	0.000049	

```
  param_knn__n_neighbors                   params  split0_test_score  \
2                      3   {'knn__n_neighbors': 3}           0.857143
3                      4   {'knn__n_neighbors': 4}           0.857143
4                      5   {'knn__n_neighbors': 5}           0.857143
0                      1   {'knn__n_neighbors': 1}           0.714286
1                      2   {'knn__n_neighbors': 2}           0.857143
5                      6   {'knn__n_neighbors': 6}           0.857143
6                      7   {'knn__n_neighbors': 7}           0.857143
7                      8   {'knn__n_neighbors': 8}           0.857143
8                      9   {'knn__n_neighbors': 9}           0.857143
9                     10  {'knn__n_neighbors': 10}           0.714286

   split1_test_score  split2_test_score  split3_test_score  split4_test_score  \
2           1.000000           1.000000           0.666667                0.8
3           1.000000           1.000000           0.666667                0.8
4           1.000000           1.000000           0.666667                0.8
0           1.000000           1.000000           0.666667                0.8
1           1.000000           0.833333           0.666667                0.8
5           1.000000           0.833333           0.666667                0.8
6           1.000000           0.833333           0.666667                0.8
7           0.857143           0.833333           0.666667                0.8
8           0.857143           0.833333           0.666667                0.8
9           0.857143           0.833333           0.666667                0.8

   mean_test_score  std_test_score  rank_test_score
2         0.870968        0.126158                1
3         0.870968        0.126158                1
4         0.870968        0.126158                1
0         0.838710        0.142743                4
1         0.838710        0.109211                4
5         0.838710        0.109211                4
6         0.838710        0.109211                4
7         0.806452        0.071337                8
8         0.806452        0.071337                8
9         0.774194        0.073391               10
```

[57]:
```
model_KNN = KNN(n_neighbors=3)
model_KNN.fit(X_train, y_train)
y_pred_KNN = model_KNN.predict(X_test)
accuracy_score(y_pred_KNN, y_test)
```

[57]:
```
0.967741935483871
```

Not too bad. You probably achieved an accuracy near 90%. After this first result we should take a more detailed look at the performance using a confusion matrix, i.e. find out the number of true positives, true, negatives, false positives and false negatives.

[58]:
```
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report

print(confusion_matrix(y_test, y_pred_test))
print(classification_report(y_test, y_pred_test))
```

```
[[26  2]
 [ 1  2]]
              precision    recall  f1-score   support

         0.0       0.96      0.93      0.95        28
         1.0       0.50      0.67      0.57         3

    accuracy                           0.90        31
   macro avg       0.73      0.80      0.76        31
weighted avg       0.92      0.90      0.91        31
```

Apparently, we only had four positive samples in the test set, out of which one half was classified incorrectly.

Task: What do you conclude about accuracy as a metric? Whats the accuracy of just labelling all samplese as negative?

[60]:
```
y_zero = np.zeros((len(y_test)))
accuracy_score(y_zero, y_test)
```

[60]:
```
0.9032258064516129
```

** Solution:**

1.5.1 ROC Curves

A better way to assess classification performance for imbalanced problems lies in looking at the *true positive rate* and *false positive rate* (also known as sensitivity and specifity).

These two can even be plotted against each other, which results in a receiver operating curve (ROC). The higher the area under the ROC curve (AUC) the better the classifier.

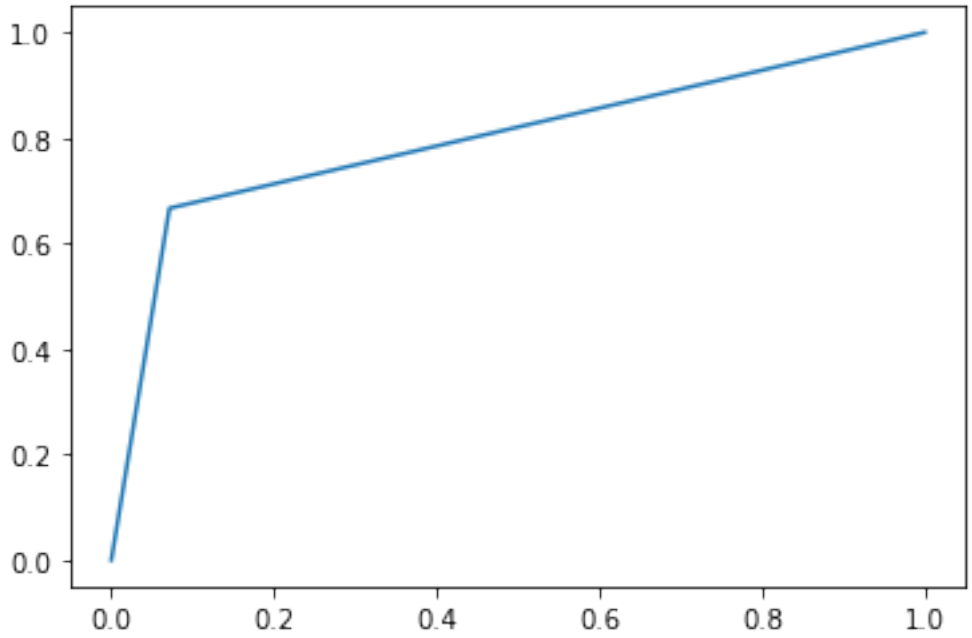
This is especially useful in medical image processing, a context in which we usually have data from a population out of which a small fraction has a particular disease that need to be detected.

Task: Whats are the *true positive rate*, *false positive rate* and *AUC* of our classifier? Plot the ROC curve!

[61]:
```
from sklearn.metrics import roc_auc_score, roc_curve
import matplotlib.pyplot as plt

#roc_auc_score(y_test, y_pred_test)
fpr, tpr, thresholds = roc_curve(y_test, y_pred_test)
```

```
plt.plot(fpr, tpr)
plt.show()
roc_auc_score(y_test, y_pred_test)
```



[61]: 0.7976190476190476

1.6 Yayy, high accuracy!... Just by chance?

Task: What if we were to consider only two features for classification, which ones would you choose? Would the accuracy be 'significantly' different depending on the choice of features?

[62]:
```
import pandas as pd

df = pd.DataFrame(data=X_train)
df.corr()
```

[62]:

	0	1	2	3
0	1.000000	0.169329	0.812832	0.698209
1	0.169329	1.000000	-0.128152	-0.065811
2	0.812832	-0.128152	1.000000	0.906489
3	0.698209	-0.065811	0.906489	1.000000

Answer:

Accuracy will probably only slightly vary. Correlation between features can be reduced by choosing features that are uncorrelated. Taking into consideration the data distributions in the evaluation

metrics choosing sepal width (1) and petal length (2) is probably the most sensible solutions given the data at this point.

We will use the scipy.stats module for simple statistical tests.

[88]:
```
from scipy import stats
```

1.7 Student's t-test

Let us consider two feature groups and see if we get similar accuracies or not: 1. only sepal length and petal length 2. only sepal width and petal width

[87]:
```
data = load_iris()
X = data[['sepal_length', 'sepal_width', 'petal_length', 'petal_width']].
 ↪as_matrix()
X_length = data[['petal_length', 'sepal_length']].as_matrix() # Considering␣
 ↪only lengths
X_width = data[['sepal_width', 'petal_width']].as_matrix() # Considering only␣
 ↪widths

y = data['species'].values

scores = np.zeros(100)
scores_length = np.zeros(100)
scores_width = np.zeros(100)

# Calculate accuracy when considering all the features
for i in range(100):
    X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=i,␣
 ↪test_size=0.5)

    model = KNN(n_neighbors=2)
    model.fit(X_train, y_train)

    y_pred = model.predict(X_test)
    scores[i] = accuracy_score(y_test, y_pred)
print('Mean accuracy considering all the features = {:2.3}'.format(scores.
 ↪mean()))

# Calculate accuracy when considering only lengths
for i in range(100):
    X_length_train, X_length_test, y_length_train, y_length_test =␣
 ↪train_test_split(X_length, y, random_state=i, test_size=0.5)

    model = KNN(n_neighbors=2)
    model.fit(X_length_train, y_length_train)
```

```
    y_length_pred = model.predict(X_length_test)
    scores_length[i] = accuracy_score(y_length_test, y_length_pred)
print('Mean accuracy considering petal_length and sepal_length only = {:2.3}'.
 ↪format(scores_length.mean()))

# Calculate accuracy when considering only widths
for i in range(100):
    X_width_train, X_width_test, y_width_train, y_width_test =␣
 ↪train_test_split(X_width, y, random_state=i, test_size=0.5)

    model = KNN(n_neighbors=2)
    model.fit(X_width_train, y_width_train)

    y_width_pred = model.predict(X_width_test)
    scores_width[i] = accuracy_score(y_width_test, y_width_pred)
print('Mean accuracy considering petal_width and sepal_width only = {:2.3}'.
 ↪format(scores_width.mean()))
```

```
/opt/conda/lib/python3.7/site-packages/ipykernel_launcher.py:2: FutureWarning:
Method .as_matrix will be removed in a future version. Use .values instead.

/opt/conda/lib/python3.7/site-packages/ipykernel_launcher.py:3: FutureWarning:
Method .as_matrix will be removed in a future version. Use .values instead.
  This is separate from the ipykernel package so we can avoid doing imports
until
/opt/conda/lib/python3.7/site-packages/ipykernel_launcher.py:4: FutureWarning:
Method .as_matrix will be removed in a future version. Use .values instead.
  after removing the cwd from sys.path.
```

```
Mean accuracy considering all the features = 0.942
Mean accuracy considering petal_length and sepal_length only = 0.929
Mean accuracy considering petal_width and sepal_width only = 0.933
```

1.7.1 2-sample t-test: testing for difference across populations

The means of both the groups are different. To test if this is significant, we perform a 2-sample t-test with scipy.ttest_ind()

[89]:
```
stats.ttest_ind(scores_length, scores_width)
```

[89]:
```
Ttest_indResult(statistic=-1.1926689440344542, pvalue=0.23442635988265995)
```

If the observed p-value is smaller than the threshold value (typically 0.01, 0.05, 0.1), we reject the null hypothesis of equal means. In this case, however, we fail to find statistical difference.

Task: Are the groups X_petal and X_sepal significantly different from X?

[94]:
```
X_petal = data[['petal_length', 'petal_width']].as_matrix() # Considering only␣
 ↪lengths
X_sepal = data[['sepal_length', 'sepal_width']].as_matrix() # Considering only␣
 ↪widths

scores_petal = np.zeros(100)
scores_sepal = np.zeros(100)

# Calculate accuracy when considering only X_petal
for i in range(100):
    X_petal_train, X_petal_test, y_petal_train, y_petal_test =␣
 ↪train_test_split(X_petal, y, random_state=i, test_size=0.5)

    model = KNN(n_neighbors=2)
    model.fit(X_petal_train, y_petal_train)

    y_petal_pred = model.predict(X_petal_test)
    scores_petal[i] = accuracy_score(y_petal_test, y_petal_pred)
print('Mean accuracy considering X_petal only = {:2.3}'.format(scores_petal.
 ↪mean()))


# Calculate accuracy when considering only X_sepal
for i in range(100):
    X_sepal_train, X_sepal_test, y_sepal_train, y_sepal_test =␣
 ↪train_test_split(X_sepal, y, random_state=i, test_size=0.5)

    model = KNN(n_neighbors=2)
    model.fit(X_sepal_train, y_sepal_train)

    y_sepal_pred = model.predict(X_sepal_test)
    scores_sepal[i] = accuracy_score(y_sepal_test, y_sepal_pred)
print('Mean accuracy considering X_sepal only = {:2.3}'.format(scores_sepal.
 ↪mean()))
stats.ttest_ind(scores_petal, scores_sepal)
```

```
/opt/conda/lib/python3.7/site-packages/ipykernel_launcher.py:1: FutureWarning:
Method .as_matrix will be removed in a future version. Use .values instead.
  """Entry point for launching an IPython kernel.
/opt/conda/lib/python3.7/site-packages/ipykernel_launcher.py:2: FutureWarning:
Method .as_matrix will be removed in a future version. Use .values instead.


Mean accuracy considering X_petal only = 0.952
Mean accuracy considering X_sepal only = 0.722
```

[94]:
```
Ttest_indResult(statistic=47.064373672156485, pvalue=1.849910693729214e-109)
```

Conclusion

```
X_petal is significantly different from X
X_sepal is significantly different from X
```

1.7.2 Paired t-test: repeated measurements

Note that the measurements for petal and sepal are made on the same flowers! In such a case where each subject is measured twice, we can use a paired sample t-test.

[95]:
```
stats.ttest_rel(scores_length, scores_width)
```

[95]:
```
Ttest_relResult(statistic=-1.3582048803087174, pvalue=0.17748587365840177)
```

Another thing we overlooked is the fact that t-tests assume a gaussian distribution of the data. But is this true for our data? Let's investigate:

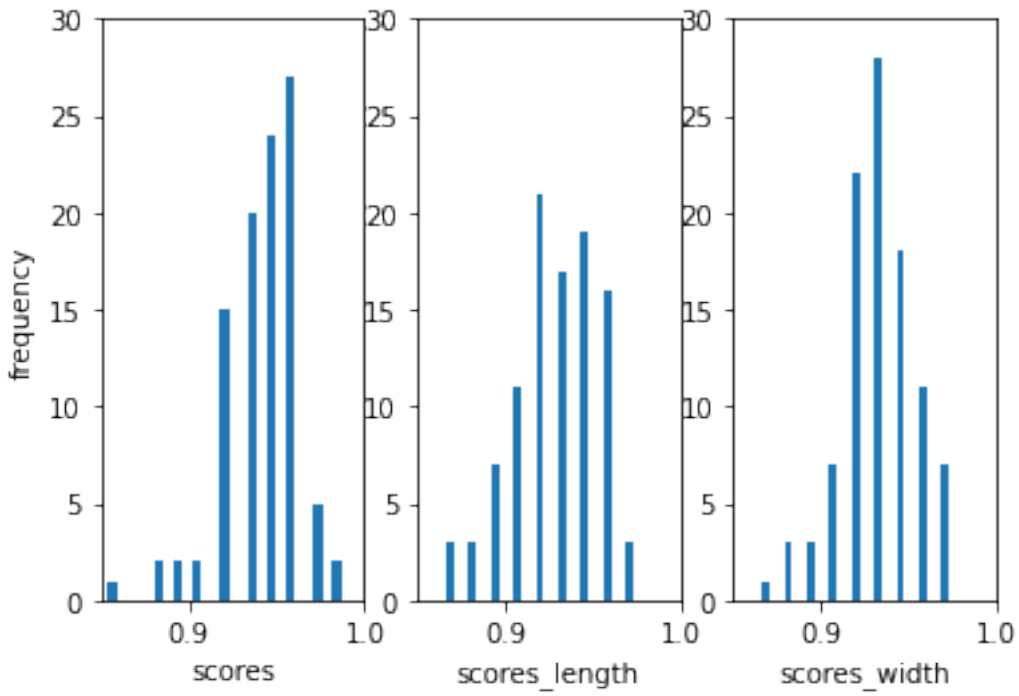
[96]:
```
# Plot the histograms to see the data distribution

bins = 25
data_x = (scores, scores_length, scores_width)
x_labels = ('scores', 'scores_length', 'scores_width')
xlims = [[0.85, 1]]*3
ylims = [[0, 30]]*3

for j, (x, xlim, ylim, x_label) in enumerate(zip(data_x, xlims, ylims,␣
 ↪x_labels)):
    ax = plt.subplot(1, 3, j+1)
    ax.hist(np.reshape(x,(-1,1)), bins = bins)
    ax.set_xlim(xlim)
    ax.set_ylim(ylim)
    ax.set_xlabel(x_label)

ax = plt.subplot(1,3,1)
ax.set_ylabel('frequency');
```

```
/opt/conda/lib/python3.7/site-packages/ipykernel_launcher.py:16:
MatplotlibDeprecationWarning: Adding an axes using the same arguments as a
previous axes currently reuses the earlier instance.  In a future version, a new
instance will always be created and returned.  Meanwhile, this warning can be
suppressed, and the future behavior ensured, by passing a unique label to each
axes instance.
  app.launch_new_instance()
```



Oops! Our data does not have a gaussian distibution! To take this into account, we must use Wilcoxon signed-rank test, with the help of scipy.stats.wilcoxon().

Task: Calculate the wilcoxon signed-rank test for the groups above. Draw a conclusion!

[100]:
```
from scipy.stats import wilcoxon
w_l, p_w = wilcoxon(scores_length, scores_width)
w, p = wilcoxon(scores_petal, scores_sepal)
print(w_l, p_w )
print(w, p)
```

```
1448.0 0.13010696808686215
0.0 3.742150539945557e-18
```

pvalue > 0.1 Hence we accept the Null Hypothesis of equal means, which concludes that we fail to find significant differences at p=0.1

Note: The corresponding test for non-paired case is the Mann-Whitney U test (scipy.stats.mannwhitneyu)

1.7.3 Feedback

That's it, we're done

If you have any suggestions on how we could improve this session, please let us know in the following cell. What did you particularly like or dislike? Did you miss any contents?