Open in Colab

Ungraded lab: Permutation Feature Importance

Welcome, during this ungraded lab you are going to be perform Permutation Feature Importance on the wine dataset using scikit-learn. In particular you will:

1. Train a Random Forest classifier on the data.
2. Compute the feature importance score by permutating each feature.
3. Re-train the model with only the top features.
4. Check other classifiers for comparison.

Let's get started!

Inspect and pre-process the data

Begin by upgrading scikit-learn to the latest version:

In [1]:

```
!pip install -U scikit-learn
```

```
Requirement already satisfied: scikit-learn in /usr/local/lib/python3.7/dist-packages (0.22.2.post1)
Collecting scikit-learn
  Downloading scikit_learn-0.24.2-cp37-cp37m-manylinux2010_x86_64.whl (22.3 MB)
     |████████████████████████████████| 22.3 MB 1.2 MB/s
Collecting threadpoolctl>=2.0.0
  Downloading threadpoolctl-2.2.0-py3-none-any.whl (12 kB)
Requirement already satisfied: joblib>=0.11 in /usr/local/lib/python3.7/dist-packages (from scikit-learn)
(1.0.1)
Requirement already satisfied: scipy>=0.19.1 in /usr/local/lib/python3.7/dist-packages (from scikit-lear
n) (1.4.1)
Requirement already satisfied: numpy>=1.13.3 in /usr/local/lib/python3.7/dist-packages (from scikit-lear
n) (1.19.5)
Installing collected packages: threadpoolctl, scikit-learn
  Attempting uninstall: scikit-learn
    Found existing installation: scikit-learn 0.22.2.post1
    Uninstalling scikit-learn-0.22.2.post1:
      Successfully uninstalled scikit-learn-0.22.2.post1
Successfully installed scikit-learn-0.24.2 threadpoolctl-2.2.0
```

Now import the required dependencies and load the dataset:

In [2]:

```
import numpy as np
from sklearn.datasets import load_wine

# as_frame param requires scikit-learn >= 0.23
data = load_wine(as_frame=True)

# Print first rows of the data
data.frame.head()
```

Out[2]:

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonflavanoid_phenols	proanthocyanins	color_inten
0	14.23	1.71	2.43	15.6	127.0	2.80	3.06	0.28	2.29	5
1	13.20	1.78	2.14	11.2	100.0	2.65	2.76	0.26	1.28	4
2	13.16	2.36	2.67	18.6	101.0	2.80	3.24	0.30	2.81	5
3	14.37	1.95	2.50	16.8	113.0	3.85	3.49	0.24	2.18	7
4	13.24	2.59	2.87	21.0	118.0	2.80	2.69	0.39	1.82	4

This dataset is made up of 13 numerical features and there are 3 different classes of wine.

Now perform the train/test split and normalize the data using `StandardScaler`:

In [3]:

```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler

# Train / Test split
X_train, X_test, y_train, y_test = train_test_split(data.data, data.target, random_state=42)

# Instantiate StandardScaler
scaler = StandardScaler()

# Fit it to the train data
scaler.fit(X_train)

# Use it to transform the train and test data
X_train = scaler.transform(X_train)

# Notice that the scaler is trained on the train data to avoid data leakage from the test set
X_test = scaler.transform(X_test)
```

Train the classifier

Now you will fit a Random Forest classifier with 10 estimators and compute the mean accuracy achieved:

In [4]:

```
from sklearn.ensemble import RandomForestClassifier

# Fit the classifier
rf_clf = RandomForestClassifier(n_estimators=10, random_state=42).fit(X_train, y_train)

# Print the mean accuracy achieved by the classifier on the test set
rf_clf.score(X_test, y_test)
```

Out[4]:

```
0.9111111111111111
```

This model achieved a mean accuracy of 91%. Pretty good for a model without fine tunning.

Permutation Feature Importance

To perform the model inspection technique known as Permutation Feature Importance you will use scikit-learn's built-in function `permutation_importance`.

You will create a function that given a classifier, features and labels computes the feature importance for every feature:

In [5]:

```
from sklearn.inspection import permutation_importance

def feature_importance(clf, X, y, top_limit=None):

  # Retrieve the Bunch object after 50 repeats
  # n_repeats is the number of times that each feature was permuted to compute the final score
  bunch = permutation_importance(clf, X, y,
                                 n_repeats=50, random_state=42)

  # Average feature importance
  imp_means = bunch.importances_mean

  # List that contains the index of each feature in descending order of importance
  ordered_imp_means_args = np.argsort(imp_means)[::-1]

  # If no limit print all features
  if top_limit is None:
    top_limit = len(ordered_imp_means_args)

  # Print relevant information
  for i, _ in zip(ordered_imp_means_args, range(top_limit)):
    name = data.feature_names[i]
    imp_score = imp_means[i]
    imp_std = bunch.importances_std[i]
    print(f"Feature {name} with index {i} has an average importance score of {imp_score:.3f} +/- {imp_st
```

The importance score is computed in a way that higher values represent better predictive power. To know exactly how it is computed check out this link.

Now use the `feature_importance` function on the Random Forest classifier and the train set:

In [6]:

```
feature_importance(rf_clf, X_train, y_train)
```

```
Feature flavanoids with index 6 has an average importance score of 0.227 +/- 0.025

Feature proline with index 12 has an average importance score of 0.142 +/- 0.019

Feature color_intensity with index 9 has an average importance score of 0.112 +/- 0.023

Feature od280/od315_of_diluted_wines with index 11 has an average importance score of 0.007 +/- 0.005
Feature total_phenols with index 5 has an average importance score of 0.003 +/- 0.004

Feature malic_acid with index 1 has an average importance score of 0.002 +/- 0.004

Feature proanthocyanins with index 8 has an average importance score of 0.002 +/- 0.003

Feature hue with index 10 has an average importance score of 0.002 +/- 0.003

Feature nonflavanoid_phenols with index 7 has an average importance score of 0.000 +/- 0.000

Feature magnesium with index 4 has an average importance score of 0.000 +/- 0.000

Feature alcalinity_of_ash with index 3 has an average importance score of 0.000 +/- 0.000

Feature ash with index 2 has an average importance score of 0.000 +/- 0.000

Feature alcohol with index 0 has an average importance score of 0.000 +/- 0.000
```

Looks like many of the features have a fairly low importance score. This points that the predictive power of this dataset is conmdensed in a few features.

However it is important to notice that this process was done for the training set, so this feature importance does NOT have into account if the feature might help with the generalization power of the model.

To check this, repeat the process for the test set:

In [7]:
```python
feature_importance(rf_clf, X_test, y_test)
```

```
Feature flavanoids with index 6 has an average importance score of 0.202 +/- 0.047

Feature proline with index 12 has an average importance score of 0.143 +/- 0.042

Feature color_intensity with index 9 has an average importance score of 0.112 +/- 0.043

Feature alcohol with index 0 has an average importance score of 0.024 +/- 0.017

Feature magnesium with index 4 has an average importance score of 0.021 +/- 0.015

Feature od280/od315_of_diluted_wines with index 11 has an average importance score of 0.015 +/- 0.018

Feature hue with index 10 has an average importance score of 0.013 +/- 0.018

Feature total_phenols with index 5 has an average importance score of 0.002 +/- 0.016

Feature nonflavanoid_phenols with index 7 has an average importance score of 0.000 +/- 0.000

Feature alcalinity_of_ash with index 3 has an average importance score of 0.000 +/- 0.000

Feature malic_acid with index 1 has an average importance score of -0.002 +/- 0.017

Feature ash with index 2 has an average importance score of -0.003 +/- 0.008

Feature proanthocyanins with index 8 has an average importance score of -0.021 +/- 0.020
```

Notice that the top most important features are the same for both sets. However features such as **alcohol**, which was considered not important for the training set is much more important when using the testing set. This hints that this feature will contribute to the generalization power of the model.

If a feature is deemed as important for the train set but not for the testing, this feature will probably cause the model to overfit.

Re-train the model with the most important features

Now you will re-train the Random Forest classifier with only the top 3 most important features.

In this case they are the same for both sets:

In [8]:
```python
print("On TRAIN split:\n")
feature_importance(rf_clf, X_train, y_train, top_limit=3)

print("\nOn TEST split:\n")
feature_importance(rf_clf, X_test, y_test, top_limit=3)
```

```
On TRAIN split:

Feature flavanoids with index 6 has an average importance score of 0.227 +/- 0.025

Feature proline with index 12 has an average importance score of 0.142 +/- 0.019

Feature color_intensity with index 9 has an average importance score of 0.112 +/- 0.023


On TEST split:

Feature flavanoids with index 6 has an average importance score of 0.202 +/- 0.047

Feature proline with index 12 has an average importance score of 0.143 +/- 0.042

Feature color_intensity with index 9 has an average importance score of 0.112 +/- 0.043
```

In [9]:

```python
# Preserve only the top 3 features
X_train_top_features = X_train[:,[6, 9, 12]]
X_test_top_features = X_test[:,[6, 9, 12]]

# Re-train with only these features
rf_clf_top = RandomForestClassifier(n_estimators=10, random_state=42).fit(X_train_top_features, y_train)

# Compute mean accuracy achieved
rf_clf_top.score(X_test_top_features, y_test)
```

Out[9]:

```
0.9333333333333333
```

Notice that by using only the 3 most important features the model achieved a mean accuracy even higher than the one using all 13 features.

Remember that the **alcohol** feature was deemed not important in the train split but you had the hypotheses that it had important information for the generalization of the model.

Add this feature and see how the model performs:

In [10]:

```python
# Preserve only the top 3 features
X_train_top_features = X_train[:,[0, 6, 9, 12]]
X_test_top_features = X_test[:,[0, 6, 9, 12]]

# Re-train with only these features
rf_clf_top = RandomForestClassifier(n_estimators=10, random_state=42).fit(X_train_top_features, y_train)

# Compute mean accuracy achieved
rf_clf_top.score(X_test_top_features, y_test)
```

Out[10]:

```
1.0
```

Wow! By adding this additional feature you know get a mean accuracy of 100%! Quite remarkable! Looks like this feature did in fact provided some important information that helped the model do a better job at generalizing.

Try out other classifiers

The process of Permutation Feature Importance is also dependant on the classifier you are using. Since different classifiers follow different rules for classification it is natural to assume they will consider different features to be important or unimportant.

To test this, try out other classifiers:

In [11]:

```python
from sklearn.svm import SVC
from sklearn.linear_model import Lasso, Ridge
from sklearn.tree import DecisionTreeClassifier

# Select 4 new classifiers
clfs = {"Laso": Lasso(alpha=0.05),
        "Ridge": Ridge(),
        "Decision Tree": DecisionTreeClassifier(),
        "Support Vector": SVC()}


# Compute feature importance on the test set given a classifier
def fit_compute_importance(clf):
  clf.fit(X_train, y_train)
  print(f"📏 Mean accuracy score on the test set: {clf.score(X_test, y_test)*100:.2f}%\n")
  print("🔝 Top 4 features when using the test set:\n")
  feature_importance(clf, X_test, y_test, top_limit=4)


# Print results
for name, clf in clfs.items():
  print("====="*20)
  print(f"➡️ {name} classifier\n")
  fit_compute_importance(clf)
```

```
====================================================================================================
➡️ Laso classifier

📏 Mean accuracy score on the test set: 86.80%

🔝 Top 4 features when using the test set:

Feature flavanoids with index 6 has an average importance score of 0.323 +/- 0.055

Feature proline with index 12 has an average importance score of 0.203 +/- 0.035

Feature od280/od315_of_diluted_wines with index 11 has an average importance score of 0.146 +/- 0.030

Feature alcalinity_of_ash with index 3 has an average importance score of 0.038 +/- 0.014

====================================================================================================
➡️ Ridge classifier

📏 Mean accuracy score on the test set: 88.71%

🔝 Top 4 features when using the test set:

Feature flavanoids with index 6 has an average importance score of 0.445 +/- 0.071

Feature proline with index 12 has an average importance score of 0.210 +/- 0.035

Feature color_intensity with index 9 has an average importance score of 0.119 +/- 0.029

Feature od280/od315_of_diluted_wines with index 11 has an average importance score of 0.111 +/- 0.026

====================================================================================================
➡️ Decision Tree classifier

📏 Mean accuracy score on the test set: 95.56%

🔝 Top 4 features when using the test set:

Feature proline with index 12 has an average importance score of 0.300 +/- 0.056

Feature flavanoids with index 6 has an average importance score of 0.297 +/- 0.061

Feature color_intensity with index 9 has an average importance score of 0.131 +/- 0.037

Feature malic_acid with index 1 has an average importance score of 0.008 +/- 0.012

====================================================================================================
➡️ Support Vector classifier

📏 Mean accuracy score on the test set: 97.78%

🔝 Top 4 features when using the test set:

Feature proline with index 12 has an average importance score of 0.069 +/- 0.031

Feature flavanoids with index 6 has an average importance score of 0.061 +/- 0.023

Feature alcohol with index 0 has an average importance score of 0.044 +/- 0.023

Feature ash with index 2 has an average importance score of 0.032 +/- 0.018
```

Looks like **flavanoids** and **proline** are very important across all classifiers. However there is variability from one classifier to the others on what features are considered the most important ones.

Congratulations on finishing this ungraded lab! Now you should have a clearer understanding of what Permutation Feature Importance is, why it is useful and how to implement this technique using scikit-learn.

Keep it up!