

Week 6 Quiz

Perrin Anto - paj2117

```
In [1]: import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.datasets import load_breast_cancer
from sklearn.dummy import DummyClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import cross_val_score

# to supress warnings about a change in the LogisticRegression solver
import warnings
warnings.simplefilter(action='ignore', category=FutureWarning)
```

```
In [2]: # Load the sample breast_cancer dataset from Scikit-Learn
#   returning just the X features matrix and y label vector.
#   The target here is a binary classification task.
#   For more information, see https://scikit-learn.org/stable/datasets/index.html#breast-cancer-dataset
X,y = load_breast_cancer(return_X_y=True)
```

```
In [3]: # Split X and y into X_train,X_test,y_train,y_test
#   using train_test_split, stratify using y.
X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=.1,stratify=y,random_state=123)
```

```
In [8]: # Get a baseline, mean 5-fold cross-validation accuracy score
#   for a DummyClassifier with default parameter settings
#   using X_train,y_train.
scores = cross_val_score(DummyClassifier(),X_train,y_train,cv=5)
print(f'mean cv accuracy: {np.mean(scores):0.2f}')
```

mean cv accuracy: 0.57

```
In [10]: # Get a mean, 5-fold cross-validation accuracy score
#   for a LogisticRegression model with default parameter settings
#   using X_train,y_train.
scores = cross_val_score(LogisticRegression(),X_train,y_train,cv=5)
print(f'mean cv accuracy: {np.mean(scores):0.2f}')
```

mean cv accuracy: 0.95

```
In [13]: # Retrain a LogisticRegression model with default parameters on the full
training set.
lr = LogisticRegression().fit(X_train,y_train)
```

```
In [15]: # Evaluate generalization accuracy of the trained LogisticRegression model on the test set.  
acc = lr.score(X_test, y_test)  
print(f'test-set accuracy: {acc:0.2f}')
```

```
test-set accuracy: 0.96
```

Question: Does our LogisticRegression model seem to be overfitting, underfitting or performing well and why?

Performing well: given the model was trained on training data and had a similar accuracy score on the training and testing sets, it shows that the model is not overfit or underfit.