Week 6 Quiz

Bryan Gibson - brg2130

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,stratify=y)
```

In [4]:
```
# 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: 0.58
```

In [5]:
```
# 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 [6]:
```
# Retrain a LogisticRegression model with default parameters on the full training set.
lr = LogisticRegression().fit(X_train,y_train)
```

In [7]:
```
# 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: 0.95
```

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

It seems to be performing well as both the cross-validation accuracy on the training set and accuracy on the test set are much higher than our baseline performance.