20250425_01

April 25, 2025

[1]:
```python
# What is Cross-Validation?
import pandas as pd
from sklearn.datasets import load_breast_cancer
from sklearn.model_selection import cross_val_score
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier

# Loading dataset
data = load_breast_cancer()
X = pd.DataFrame(data.data, columns = data.feature_names)
y = pd.Series(data.target)
```

[13]:
```python
# Creating models
logreg = LogisticRegression(solver = 'liblinear')
tree = DecisionTreeClassifier(random_state = 42)
```

[15]:
```python
# Comparing models
logreg_score = cross_val_score(logreg, X, y, cv = 5)
print('Logistic Regression cross-validation scores:', logreg_score)
print('Mean accuracy:', logreg_score.mean())

tree_score = cross_val_score(tree, X, y, cv = 5)
print('Decision Tree cross-validation scores:', tree_score)
print('Mean accuracy:', tree_score.mean())
```

```
Logistic Regression cross-validation scores: [0.92982456 0.93859649 0.97368421
0.94736842 0.96460177]
Mean accuracy: 0.9508150908244062
Decision Tree cross-validation scores: [0.9122807  0.90350877 0.92982456
0.95614035 0.88495575]
Mean accuracy: 0.9173420276354604
```

[23]:
```python
# Making pipelines
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler

# Standizing for Logistic Regression, but no need for Decision Tree
```

```
logreg_pipeline = Pipeline([('scaler', StandardScaler()),('logreg',␣
 ↪LogisticRegression(solver = 'liblinear'))])
tree_pipeline = Pipeline([('tree', DecisionTreeClassifier(random_state = 42))])

# Cross-validation
logreg_pipeline_scores = cross_val_score(logreg_pipeline, X, y, cv = 5)
print("Logistic Regression Pipeline cross-validation Scores:",␣
 ↪logreg_pipeline_scores)
print("Mean Accuracy:", logreg_pipeline_scores.mean())

tree_pipeline_scores = cross_val_score(tree_pipeline, X, y, cv = 5)
print("Decision Tree Pipeline cross-validation Scores:", tree_pipeline_scores)
print("Mean Accuracy:", tree_pipeline_scores.mean())
```

```
Logistic Regression Pipeline cross-validation Scores: [0.98245614 0.97368421
0.97368421 0.97368421 0.99115044]
Mean Accuracy: 0.9789318428815402
Decision Tree Pipeline cross-validation Scores: [0.9122807  0.90350877
0.92982456 0.95614035 0.88495575]
Mean Accuracy: 0.9173420276354604
```

0.0.1 Conclusion:

- Logistic Regression achieved higher and more stable accuracy (mean 0.951) compared to Decision Tree (mean 0.917).
- After introducing a `Pipeline` with `StandardScaler`, Logistic Regression performance **further improved**, confirming its sensitivity to feature scaling.
- Decision Tree was also wrapped in a `Pipeline` for consistency, though scaling is not needed for tree-based models.