Stats 101C Homework 4

Damien Ha

In [1]:
```
import pandas as pd
from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score
```

In [2]:
```
df = pd.read_csv('smoke_detection_iot.csv')
```

In [3]:
```
df.head()
```

Out[3]:

	Unnamed: 0	UTC	Temperature[C]	Humidity[%]	TVOC[ppb]	eCO2[ppm]	Raw H2	Raw Ethanol
0	0	1654733331	20.000	57.36	0	400	12306	18520
1	1	1654733332	20.015	56.67	0	400	12345	18651
2	2	1654733333	20.029	55.96	0	400	12374	18764
3	3	1654733334	20.044	55.28	0	400	12390	18849
4	4	1654733335	20.059	54.69	0	400	12403	18921

#1

In [4]:
```
# Split the dataset into training (70%) and testing (30%) datasets
X = df.iloc[:, :-1].values
y = df['Fire Alarm'].values

# Scale the features
X = StandardScaler().fit_transform(X)

X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.3, random_state=42)

# Fit logistic regression on the training dataset
logreg = LogisticRegression()
```

```
logreg.fit(X_train, y_train)

# Evaluate logistic regression on the testing dataset
y_pred_logreg = logreg.predict(X_test)
error_log = 1 - accuracy_score(y_test, y_pred_logreg)

# Fit 5-NN on the training dataset
knn = KNeighborsClassifier(n_neighbors=5)
knn.fit(X_train, y_train)

# Evaluate 5-NN on the testing dataset
y_pred_knn = knn.predict(X_test)
error_5_nn = 1 - accuracy_score(y_test, y_pred_knn)
```

```
/Users/dha/opt/anaconda3/lib/python3.9/site-packages/sklearn/linear_model/_logis
tic.py:458: ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regressio
n
  n_iter_i = _check_optimize_result(
```

#2

In [5]:

```
# Apply 10-fold CV on the training dataset for logistic regression
and 5-NN
cv_scores_logreg = cross_val_score(logreg, X_train, y_train, cv=10,
scoring='accuracy')
cv_scores_knn = cross_val_score(knn, X_train, y_train, cv=10,
scoring='accuracy')

# Report validation errors
validation_error_logreg = 1 - cv_scores_logreg.mean()
validation_error_knn = 1 - cv_scores_knn.mean()
```

```
/Users/dha/opt/anaconda3/lib/python3.9/site-packages/sklearn/linear_model/_logis
tic.py:458: ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
```

```
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regressio
n
  n_iter_i = _check_optimize_result(
/Users/dha/opt/anaconda3/lib/python3.9/site-packages/sklearn/linear_model/_logis
tic.py:458: ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regressio
n
  n_iter_i = _check_optimize_result(
/Users/dha/opt/anaconda3/lib/python3.9/site-packages/sklearn/linear_model/_logis
tic.py:458: ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regressio
n
  n_iter_i = _check_optimize_result(
/Users/dha/opt/anaconda3/lib/python3.9/site-packages/sklearn/linear_model/_logis
tic.py:458: ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regressio
n
  n_iter_i = _check_optimize_result(
/Users/dha/opt/anaconda3/lib/python3.9/site-packages/sklearn/linear_model/_logis
tic.py:458: ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regressio
n
  n_iter_i = _check_optimize_result(
/Users/dha/opt/anaconda3/lib/python3.9/site-packages/sklearn/linear_model/_logis
tic.py:458: ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regressio
```

```
n
  n_iter_i = _check_optimize_result(
/Users/dha/opt/anaconda3/lib/python3.9/site-packages/sklearn/linear_model/_logis
tic.py:458: ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regressio
n
  n_iter_i = _check_optimize_result(
/Users/dha/opt/anaconda3/lib/python3.9/site-packages/sklearn/linear_model/_logis
tic.py:458: ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regressio
n
  n_iter_i = _check_optimize_result(
/Users/dha/opt/anaconda3/lib/python3.9/site-packages/sklearn/linear_model/_logis
tic.py:458: ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regressio
n
  n_iter_i = _check_optimize_result(
/Users/dha/opt/anaconda3/lib/python3.9/site-packages/sklearn/linear_model/_logis
tic.py:458: ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regressio
n
  n_iter_i = _check_optimize_result(
```

#3

In [6]:

```
# Compare validation error with "True Testing Error"
print("True Testing Error for Logistic Regression:
{:.4f}".format(error_log))
print("True Testing Error for 5-NN: {:.4f}".format(error_5_nn))
print("\nValidation Error for Logistic Regression:
```

```
{:.4f}".format(validation_error_logreg))
print("Validation Error for 5-NN:
{:.4f}".format(validation_error_knn))

# Draw conclusions based on observations
if validation_error_logreg < error_log:
    print("\nLogistic Regression performs better in cross-validation
than on the testing dataset.")
else:
    print("\nLogistic Regression performs better on the testing
dataset than in cross-validation.")

if validation_error_knn < error_5_nn:
    print("5-NN performs better in cross-validation than on the
testing dataset.")
else:
    print("5-NN performs better on the testing dataset than in cross-
validation.")
```

```
True Testing Error for Logistic Regression: 0.0128
True Testing Error for 5-NN: 0.0001

Validation Error for Logistic Regression: 0.0140
Validation Error for 5-NN: 0.0003

Logistic Regression performs better on the testing dataset than in cross-validat
ion.
5-NN performs better on the testing dataset than in cross-validation.
```