Stats 101C Homework 5

Damien Ha

In [1]:
```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
```

In [2]:
```
train_data = pd.read_csv('Training_dataset.csv')
test_data = pd.read_csv('Testing_dataset.csv')
```

In [3]:
```
# Extract features (x) and labels (y) from training data
x_train = train_data.drop('Y', axis=1)
y_train = train_data['Y']

# Extract features (x) and labels (y) from testing data
x_test = test_data.drop('Y', axis=1)
y_test = test_data['Y']
```

#1

In [4]:
```
# Fit a logistic regression model
logreg_raw = LogisticRegression()
logreg_raw.fit(x_train, y_train)

# Test the model on testing data
y_pred_raw = logreg_raw.predict(x_test)

# Testing performance
accuracy_raw = accuracy_score(y_test, y_pred_raw)
print(f'Testing Accuracy with Raw Data: {accuracy_raw:.4f}')
```

```
Testing Accuracy with Raw Data: 0.7367
```

#2

In [5]:
```
# Feature engineering
from sklearn.preprocessing import PolynomialFeatures
```

```
from sklearn.preprocessing import StandardScaler
import numpy as np

# Load the datasets
training = pd.read_csv('Training_dataset.csv')
testing = pd.read_csv('Testing_dataset.csv')

# Feature Engineering
training['X2'] = np.sin(training['X']**2)
testing['X2'] = np.sin(testing['X']**2)
training['X3'] = np.sin(training['X']**8) * 4
testing['X3'] = np.exp(testing['X']**2) * 7

# Model Training
features = ['X', 'X2', 'X3']
X_train = training[features]
y_train = training['Y']

# Logistic Regression
log_poly = LogisticRegression()
log_poly.fit(X_train, y_train)

# Model Prediction
X_test = testing[features]
poly_predict = log_poly.predict(X_test)

# Testing Accuracy
poly_accuracy = accuracy_score(testing['Y'], poly_predict)
print(f'Testing Accuracy with Polynomial Features:
{poly_accuracy:.4f}')
```

```
Testing Accuracy with Polynomial Features: 0.8433
```

#3

Using polynomial features helps to increase the accuracy of the logistic regression model, though using too large a number of values could potentially lead to overfitting