Exercise Sheet 4

1. Exercise: Overfitting and Underfitting

(a) Generate a dataset with a quadratic relationship between x and y using the function $y = ax^2 + bx + c + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 4.5)$. The parameters are $a = 0.25, b = -5$ and $c = 0.2$. The dataset should be composed of in total $M = 20$ samples, which are generated by randomly sampling x in the range of $[0, 30]$.

(b) Fit the data by a linear, a quadratic, and a high-degree polynomial regression model (10th degree) using

```
from sklearn.preprocessing import PolynomialFeatures
from sklearn.linear_model import LinearRegression
from sklearn.pipeline import make_pipeline

def PolynomialRegression(degree=2, **kwargs):
        return make_pipeline(PolynomialFeatures(degree),
                             LinearRegression(**kwargs))
```

The data can be fit by using

```
model = PolynomialRegression(degree).fit(X, y)
```

and

```
polynominal = model.predict(xfit)
```

where `xfit` is a one-dimensional column vector that spans from 0 to 30, consisting of 5000 evenly spaced points.

(c) Plot all the models along with the data in a single plot with multiple axes. Ensure that the plot contains axis labels and subtitles for clarity.

(d) Compute and display the Mean Squared Error (MSE) of each model.

(e) Explain which model is overfitting and which is underfitting.

(f) Explain what problems arise from overfitting and underfitting.

2. Exercise: Model Selection and Evaluation

(a) How can Receiver Operating Characteristic (ROC) curves be used to compare the performance of different classifiers?

(b) Why is a ROC curve that falls along the diagonal line considered not helpful?

(c) How does a perfect classifier look on the ROC curve?

(d) Two predictive models, designated as model f_1 and model f_2, are used for the purpose of predicting whether various examinations would result in a fail ($y = 0$) or pass ($y = 1$) for a student. The predictions generated by these two classifiers are denoted by $\hat{\boldsymbol{y}}_1$ and $\hat{\boldsymbol{y}}_2$ respectively. Concurrently, the real outcomes or actual results, indicating if the exams were passed or not, are represented by the ground truth vector $\boldsymbol{y}$. The values assigned to these arrays are as follows:

$$\hat{\boldsymbol{y}}_1 = [0.1, 0.8, 0.62, 0.3, 0.45, 0.2, 0.2, 0.9, 0.6, 0.2, 0.1, 0.8]^{\mathrm{T}} \tag{1}$$

$$\hat{\boldsymbol{y}}_2 = [0.2, 0.9, 0.45, 0.9, 0.1, 0.2, 0.55, 0.85, 0.15, 0.1, 0.3, 0.7]^{\mathrm{T}} \tag{2}$$

$$\boldsymbol{y} = [0, 1, 1, 1, 0, 1, 0, 1, 0, 0, 0, 1]^{\mathrm{T}} \tag{3}$$

This data provides the basis for comparing and evaluating the performance of the predictive models.

(e) Generate the ROC curve for both predictive models on a single plot that includes multiple axes. Ensure that the plot contains axis labels and subtitles for clarity. Additionally, compute the Area Under the Curve (AUC) for each model.

(f) Consider the condition where a prediction $\hat{y}$ that is greater than or equal to 0.5 is classified as a pass. For each model, generate a confusion matrix and calculate the Sensitivity, Specificity and F-Score.

(g) Comparatively analyze the results of (e)-(f) for both models and determine which model is better based on these metrics. Explain your answer.