Self-Check #12

1) Nearest neighbors to point 6 and y_hat:

Distance from Point 6 to Point 1:

$$\text{distance} = \sqrt{(0.23-0.39)^2+(0.81-0.63)^2} = \sqrt{(-0.16)^2+(0.18)^2} = \sqrt{0.0256+0.0324} = \sqrt{0.058} \approx 0.24$$

Distance from Point 6 to Point 2:

$$\text{distance} = \sqrt{(0.42-0.39)^2+(0.78-0.63)^2} = \sqrt{(0.03)^2+(0.15)^2} = \sqrt{0.0009+0.0225} = \sqrt{0.0234} \approx 0.153$$

Distance from Point 6 to Point 3:

$$\text{distance} = \sqrt{(0.64-0.39)^2+(0.23-0.63)^2} = \sqrt{(0.25)^2+(-0.4)^2} = \sqrt{0.0625+0.16} = \sqrt{0.2225} \approx 0.47$$

Distance from Point 6 to Point 4:

$$\text{distance} = \sqrt{(0.87-0.39)^2+(0.19-0.63)^2} = \sqrt{(0.48)^2+(-0.44)^2} = \sqrt{0.2304+0.1936} = \sqrt{0.424} \approx 0.65$$

Distance from Point 6 to Point 5:

$$\text{distance} = \sqrt{(0.76-0.39)^2+(0.43-0.63)^2} = \sqrt{(0.37)^2+(-0.2)^2} = \sqrt{0.1369+0.04} = \sqrt{0.1769} \approx 0.42$$

ordering the distances the 3 nearest neighbors are Points 2, 1, and 5.

Step 5: Predict the value of y for Point 6

To predict the value of y for Point 6, we can use the average of the y-values of the 3 nearest neighbors.

$$\hat{y}_6 = \frac{0.33+0.18+0.32}{3} = \frac{0.83}{3} \approx 0.277$$

The 3 nearest neighbors to Point 6 are Points 2, 1, and 5.

The predicted value of y for Point 6 is approximately 0.277.

2) As it was stated in lecture, you do not need to take the square root to compare distances because it is already apples to apples after being squared.

3) Calculating Metrics

Let's analyze the confusion matrix provided and calculate the various performance metrics.

Confusion Matrix:

	Actual Positive	Actual Negative
Predicted Positive	329	35
Predicted Negative	87	357

From this matrix:

- True Positives (TP) = 329 (Predicted Positive and actually Positive)
- False Positives (FP) = 35 (Predicted Positive and actually Negative)
- False Negatives (FN) = 87 (Predicted Negative and actually Positive)
- True Negatives (TN) = 357 (Predicted Negative and actually Negative)

Now, the metrics.

1. Accuracy

Accuracy is the proportion of correctly predicted instances (both positive and negative) to the total instances.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Accuracy} = \frac{329 + 357}{329 + 357 + 35 + 87} = \frac{686}{808} \approx 0.849$$

So, Accuracy ≈ 0.849 or 84.9%.

2. Error

Error is the proportion of incorrectly predicted instances (both false positives and false negatives) to the total instances.

$$\text{Error} = \frac{FP + FN}{TP + TN + FP + FN}$$

$$\text{Error} = \frac{35 + 87}{329 + 357 + 35 + 87} = \frac{122}{808} \approx 0.151$$

So, Error ≈ 0.151 or 15.1%.

3. Precision

Precision (also called Positive Predictive Value) is the proportion of predicted positives that are actually positive.

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Precision} = \frac{329}{329 + 35} = \frac{329}{364} \approx 0.905$$

So, Precision ≈ 0.905 or 90.5%.

4. Recall

Recall (also called Sensitivity or True Positive Rate) is the proportion of actual positives that are correctly identified.

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Recall} = \frac{329}{329 + 87} = \frac{329}{416} \approx 0.791$$

So, Recall ≈ 0.791 or 79.1%.

4) MSE

The Mean Squared Error (MSE) is calculated using the formula:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Where n is the number of data points.

The sum of these squared errors is:

$$0.2116 + 0.1521 + 0.0196 + 0.1521 + 0.0361 = 0.5715$$

Now, divide by the number of data points $n = 5$:

$$\text{MSE} = \frac{0.5715}{5} = 0.1143$$

What is the mean of y?

$$\text{mean}(y) = \frac{3.78 + 4.82 + 2.83 + 2.76 + 3.48}{5} = \frac{17.67}{5} = 3.534$$

If we used the mean of y as the prediction for each y_i, the predicted values $\hat{y}_i$ would all be equal to the mean of y, which is 3.534.

$$(y_1 - \hat{y})^2 = (3.78 - 3.534)^2 = (0.246)^2 = 0.0605$$

$$(y_2 - \hat{y})^2 = (4.82 - 3.534)^2 = (1.286)^2 = 1.6518$$

$$(y_3 - \hat{y})^2 = (2.83 - 3.534)^2 = (-0.704)^2 = 0.4950$$

$$(y_4 - \hat{y})^2 = (2.76 - 3.534)^2 = (-0.774)^2 = 0.5998$$

$$(y_5 - \hat{y})^2 = (3.48 - 3.534)^2 = (-0.054)^2 = 0.0029$$

Summing the squared errors:

$$0.0605 + 1.6518 + 0.4950 + 0.5998 + 0.0029 = 2.8099$$

Finally, divide by the number of data points $n = 5$:

$$\text{MSE} = \frac{2.8099}{5} = 0.56198$$

So, the MSE if we used the mean of y as the predictor is approximately 0.562.

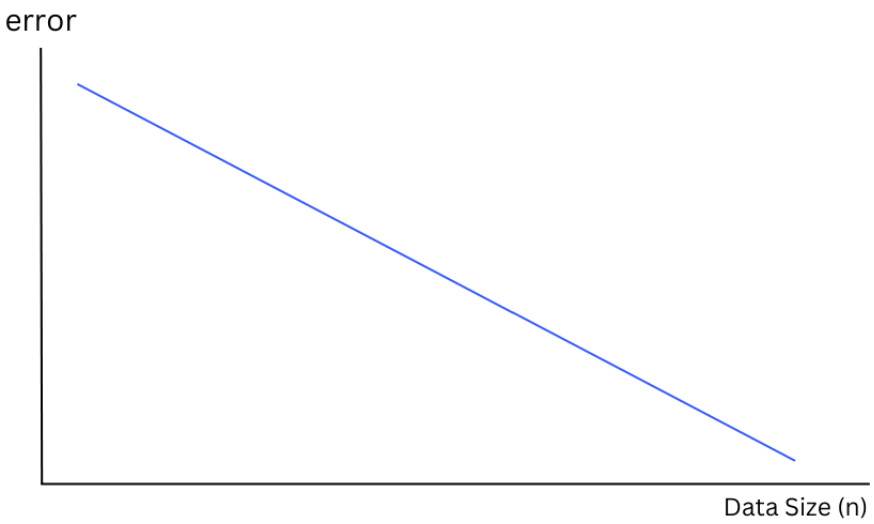
The variance of a population is given by the formula:

$$\text{Variance} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu)^2$$

Where μ is the mean of y (in this case, 3.534), and n is the number of data points.

This is very similar to the calculation of the MSE using the mean of y as the predictor. The only difference is that the variance typically uses all data points in the population, while MSE evaluates a model's predictions. If you use the mean of y as the predicted value for all observations, the MSE becomes equivalent to the population variance.

A learning curve where collecting more data will help looks like:



A learning curve where collecting more data will not help is: