**Worksheet: Lecture 3 Classification**

**Training & test datasets:**

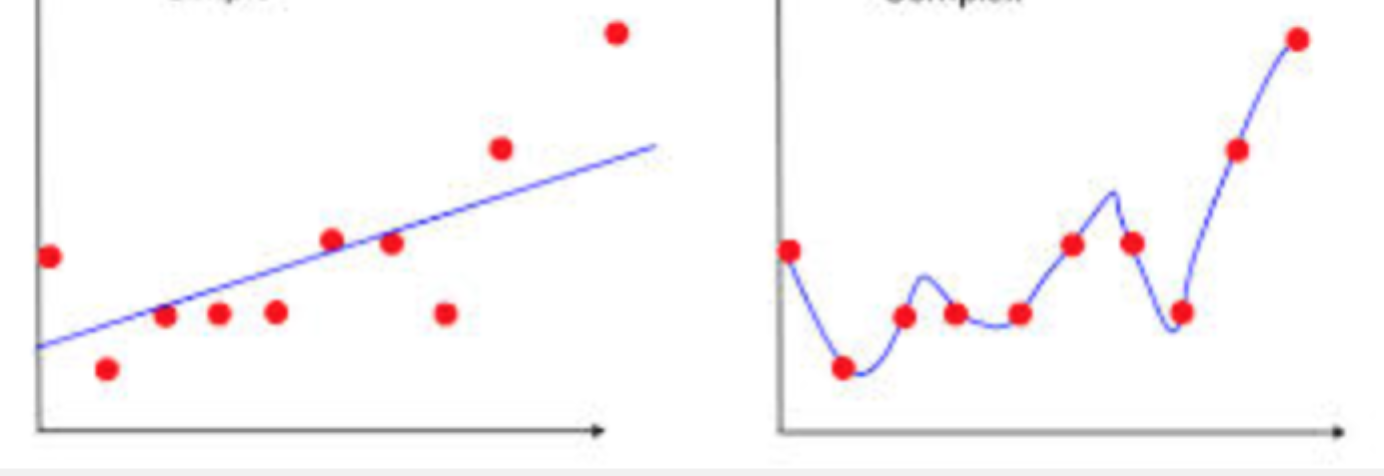
1. What is a training dataset?
2. What is a test dataset?
3. Would a training datset have more rows than a test dataset?

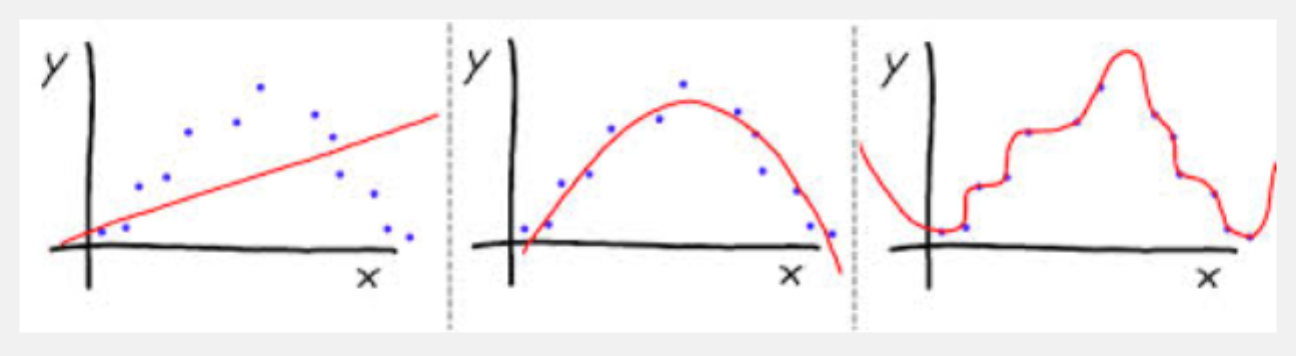
**Overfitting:**

1. What is a **generalisation error?**
2. Which of the following images illustrates over fitting, and which illustrates under-fitting?

**b.**

**a.**





**e.**

**d.**

**c.**

1. Why is over-fitting a problem?
2. If k is set to 10, and cross validation returns the following 10 accuracies, how would you calculate overall model accuracy?

What model is returned when using Cross Validation?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 67% | 69% | 72% | 71% | 74% |
| 69% | 70% | 67% | 72% | 71% |

1. For each of the following confusion matrices, calculate overall accuracy, and the precision and recall for each class.

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **Actual Class Label** | |
| **Good risk** | **Bad Risk** |
| **Predicted class label** | **Good risk** | 60 | 10 |
| **Bad risk** | 40 | 80 |

**Total number of rows: 190**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **Actual Class Label** | |
| **Poisonous** | **Edible** |
| **Predicted class label** | **Poisonous** | 80 | 40 |
| **Edible** | 20 | 40 |

**Total number of rows: 180**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | **Actual Class Label** | | |
| **Fail** | **Pass** | **Honours** |
| **Predicted class label** | **Fail** | 95 | 15 | 0 |
| **Pass** | 5 | 50 | 30 |
| **Honours** | 0 | 15 | 30 |

**Total number of rows: 240**