MANCHESTER
1824
The University of Manchester

UCIL20122 – AI: Robot Overlord, Replacement or Colleague?
April 2021

INDIVIDUAL PROJECT

Reflective Exercise

Instructions:

Answer the following questions based on the Jupyter Notebooks activities, and upload a pdf of this document on Blackboard. More information and deadlines on the assessment page on Blackboard.

A. What happens when you set k to 9 with the Welsh men's rugby player Leigh Halfpenny, whose height is 178cm, and weight is 85kg? (2 marks)

Answer: Leigh Halfpenny (178, 85) will be classified as a Ballet dancer.

Why is this? (2 marks)

Answer: k = 9 means we are supposed to calculate the 9 nearest points (i.e. training data) to the input data point (178, 85). Of these nine neighbours, more are labelled as ballet dancers than as rugby players, according to the k-nn theory, this new data is classified as rugby players

B. What is the maximum number of neighbours we can use before our model returns invalid results? (2 marks)

Answer: 12.

Why is this? (2 marks)

Answer: Because it should be smaller or equal to the number of training data.

C. What is the effect of using a higher number for k (2 marks), and a lower number for k (2 marks) on the model? Explain your answers using examples from our dataset.

Answer: If using a higher number of k, neighbours will include too many samples from other classes, which may leads to wrong prediction(classification). For example, when k = 9, a rugby player (example from Question A, height = 178cm, weight = 85kg) will be classified as a ballet dancer.

If using a lower number of k, the classification result is likely to be more accurate, using the same example (rugby player, weight = 85kg, height = 178cm), setting k = 1or k = 3 this data will be classified as rugby play, which is a correct classification. It is worth noting that this accuracy depends on the quality of the training sample, smaller k will model noisy, taking Audrey Abadie (from women's rugby world, weight = 62kg, height = 166cm) as an example, k = 3, she will be classified as a ballet dancer.

D. Looking at our dataset, we should avoid setting k to certain numbers – which are those? (2 marks)

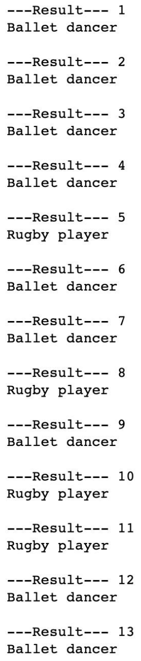
Answer: Even numbers, i.e. 2, 4, 6, 8, 10, 12.

Why is this? (2 marks)

Answer: The training data contains 6 rugby players, and 6 ballet dancers, so it is half-half. When k is even number, it is likely that half of the neighbour(s) belongs to one class, and half of the neighbour(s) belongs another class, so it is hard to perform the correct classification (in this case, according to the source code, we could know that all the testing date will be classified as rugby player).

E. What happens when you test the data for the players from the French women's rugby team in the classifier, with k set to 8? (2 marks)

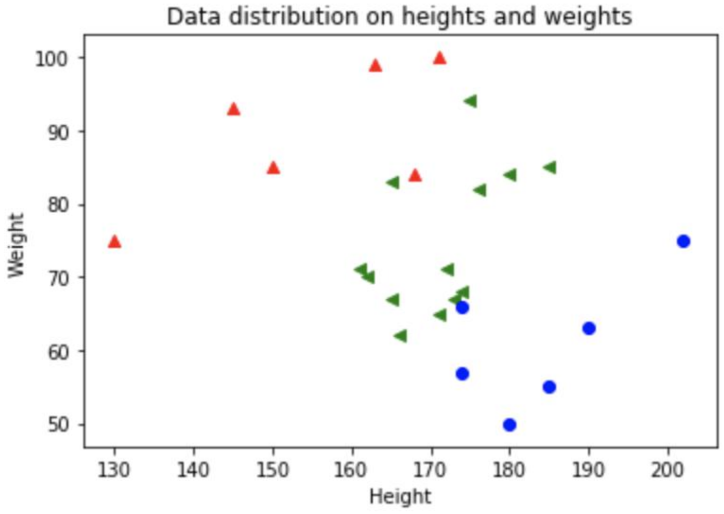
Answer: Most of the testing data will be classified as ballet dancer. To be more specific, see below, where result 1 represents the first woman in the Women's Rugby World Cup Table and so on...



```
---Result--- 1
Ballet dancer

---Result--- 2
Ballet dancer

---Result--- 3
Ballet dancer

---Result--- 4
Ballet dancer

---Result--- 5
Rugby player

---Result--- 6
Ballet dancer

---Result--- 7
Ballet dancer

---Result--- 8
Rugby player

---Result--- 9
Ballet dancer

---Result--- 10
Rugby player

---Result--- 11
Rugby player

---Result--- 12
Ballet dancer

---Result--- 13
Ballet dancer
```

Why is this? (2 marks)

Answer: (The green points represent the testing data, red points represent the training data where are labelled as rugby player, and the blue points represent the training data where are labelled as ballet dancer)



From the graph, we could see that, most of the green points (testing data) are closer to the points which labelled as ballet dancer, so when k = 8, most of the neighbours are blue points rather than red points, thus, more testing point will be classified as ballet dancer while a few of them will be classified as rugby player since they get more neighbours which are labelled as rugby player.

Moreover, all the testing data are with the sex labelled as 1(women), while the training data not only contains men(sex = 0), but also contains women(sex = 1), so the training data coverage is not comprehensive (insufficient information), in other words, the quality of the training data is not good enough, which will also cause misclassification.