KNN

Zahoor Tanoli (PhD)

CUI Attock

KNN

- KNN is a non-parametric classification method, which means that no parameters of the population distribution are estimated.
- KNN is a supervised machine learning algorithm, which means that we need data with known class/group.
- KNN is a type of lazy learning algorithm because it does not create a model in comparison with most other classification methods. Instead, it predicts directly based on the training data.
- The algorithm can be continuously updated with new training data.

Example data

Infection	CRP (mg/L)	PCT (µg/L)
Viral	40	36
Viral	30	30
Viral	32	45
Viral	37	55
Viral	40	47
Viral	52	40
Bacterial	48	59
Bacterial	60	67
Bacterial	68	50
Bacterial	40	65
Bacterial	55	55
Bacterial	58	68

The blood concentration of the c-reactive protein (CRP) and procalcitonin (PCT) have been measured on 12 patients that have entered a hospital.

Example data

Infection	CRP (mg/L)	PCT (µg/L)
Viral	40	36
Viral	30	30
Viral	32	45
Viral	37	55
Viral	40	47
Viral	52	40
Bacterial	48	59
Bacterial	60	67
Bacterial	68	50
Bacterial	40	65
Bacterial	55	55
Bacterial	58	68

Once the patients had entered the hospital, the presence of bacteria and viruses were analyzed. However, it usually takes several hours or days to determine if a patient has a viral infection or a bacterial infection.

Example data

Infection	CRP (mg/L)	PCT (µg/L)
Viral	40	36
Viral	30	30
Viral	32	45
Viral	37	55
Viral	40	47
Viral	52	40
Bacterial	48	59
Bacterial	60	67
Bacterial	68	50
Bacterial	40	65
Bacterial	55	55
Bacterial	58	68

After two days at the hospital, these six patients were found to be infected by a virus,

Example data

Infection	CRP (mg/L)	PCT (µg/L)
Viral	40	36
Viral	30	30
Viral	32	45
Viral	37	55
Viral	40	47
Viral	52	40
Bacterial	48	59
Bacterial	60	67
Bacterial	68	50
Bacterial	40	65
Bacterial	55	55
Bacterial	58	68

whereas these six patients were confirmed to have a bacterial infection. Since antibiotics are only effective on bacteria, only these patients were treated with antibiotics.

Example data

Infection	CRP (mg/L)	PCT (µg/L)
Viral	40	36
Viral	30	30
Viral	32	45
Viral	37	55
Viral	40	47
Viral	52	40
Bacterial	48	59
Bacterial	60	67
Bacterial	68	50
Bacterial	40	65
Bacterial	55	55
Bacterial	58	68

The problem is that we have to wait about two days to confirm the type of pathogen. We therefore need to wait two days to know if antibiotic treatment is appropriate or not.

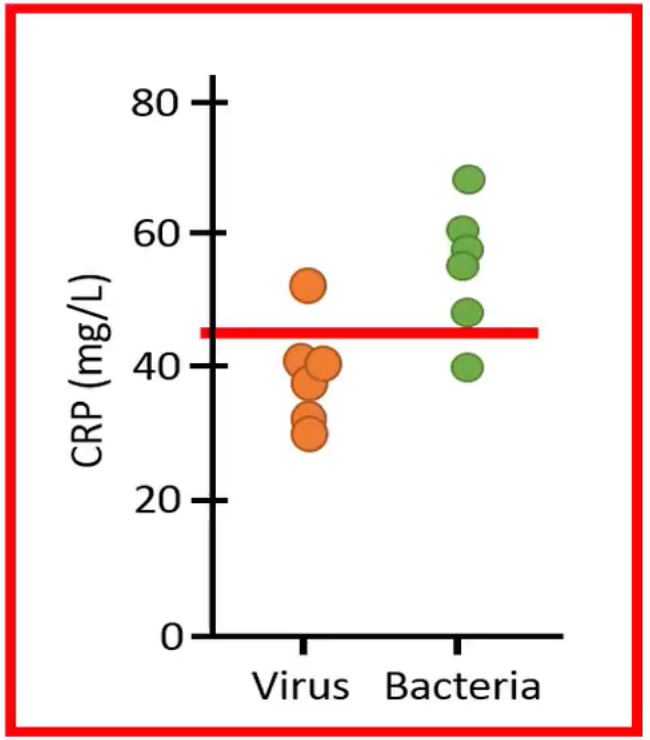
Example data

Infection	CRP (mg/L)	PCT (µg/L)
Viral	40	36
Viral	30	30
Viral	32	45
Viral	37	55
Viral	40	47
Viral	52	40
Bacterial	48	59
Bacterial	60	67
Bacterial	68	50
Bacterial	40	65
Bacterial	55	55
Bacterial	58	68

It would therefore be great if we could use the CRP or the PCT concentration to tell if a patient has a bacterial or viral infection because the measurements of these variables can be done within just an hour.

Example data

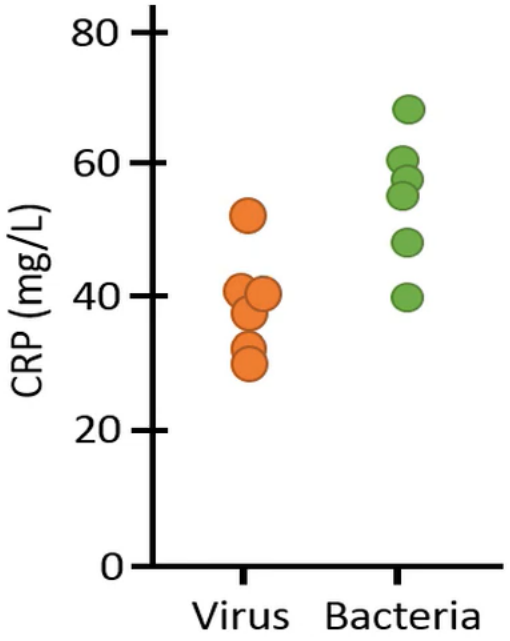
Infection	CRP (mg/L)	PCT (µg/L)
Viral	40	36
Viral	30	30
Viral	32	45
Viral	37	55
Viral	40	47
Viral	52	40
Bacterial	48	59
Bacterial	60	67
Bacterial	68	50
Bacterial	40	65
Bacterial	55	55
Bacterial	58	68

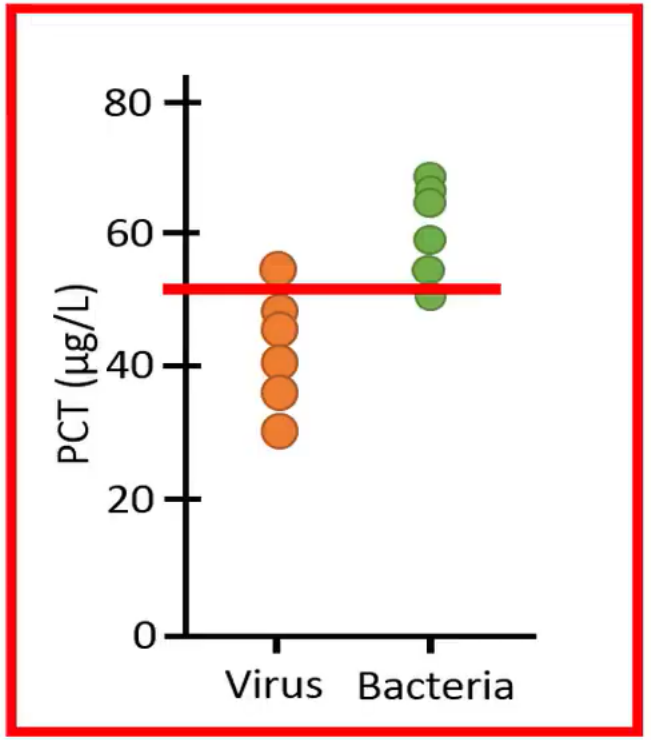


If we plot the CRP concentration of the 12 patients, we see that no simple cutoff line can be used to clearly separate the ones with a bacterial infection from the ones with a viral infection.

Example data

Infection	CRP (mg/L)	PCT (µg/L)
Viral	40	36
Viral	30	30
Viral	32	45
Viral	37	55
Viral	40	47
Viral	52	40
Bacterial	48	59
Bacterial	60	67
Bacterial	68	50
Bacterial	40	65
Bacterial	55	55
Bacterial	58	68

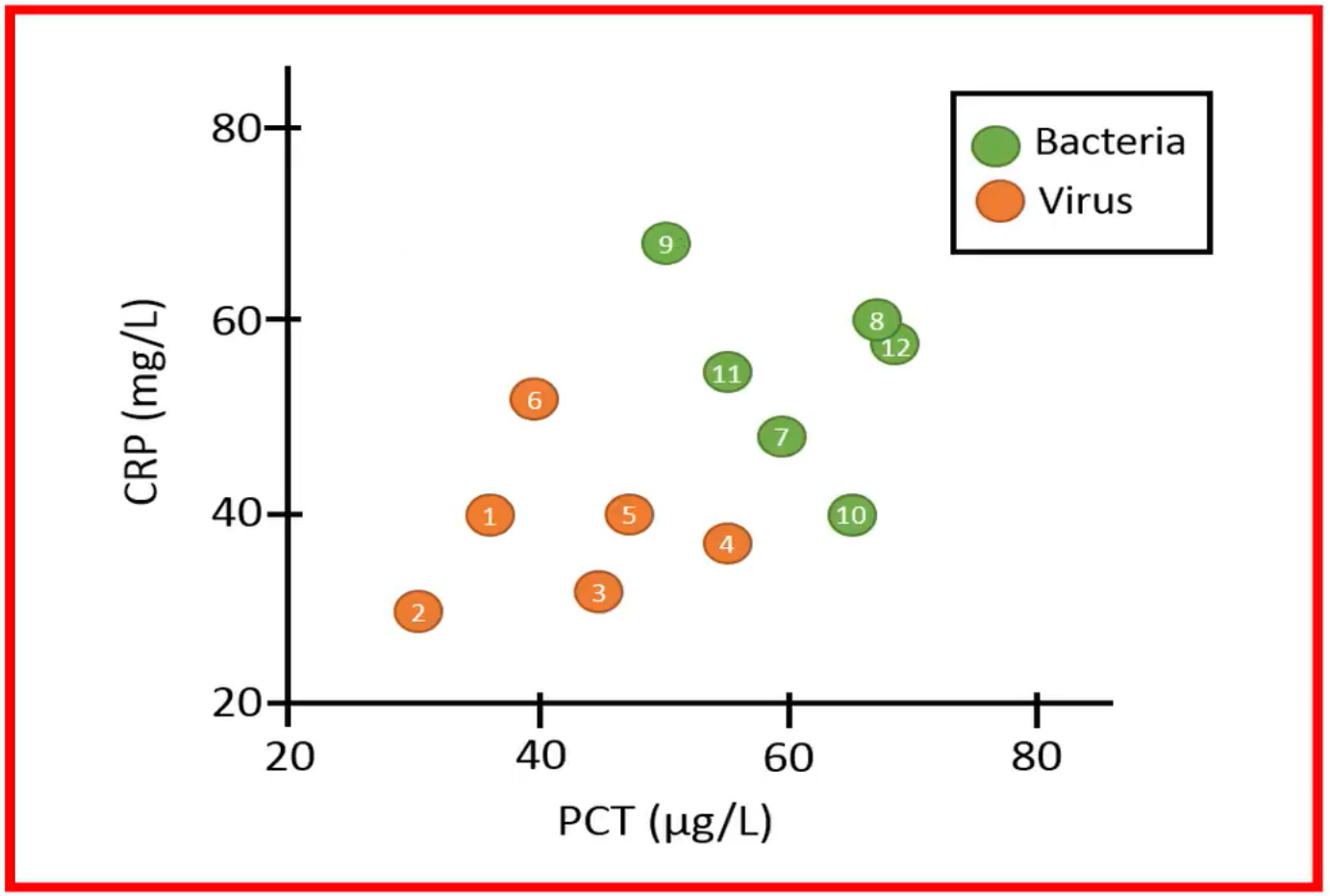




The same is also true for the PCT level, because we cannot completely separate the patients with a bacterial infection from the patients with a viral infection by using only the PCT concentration.

Example data

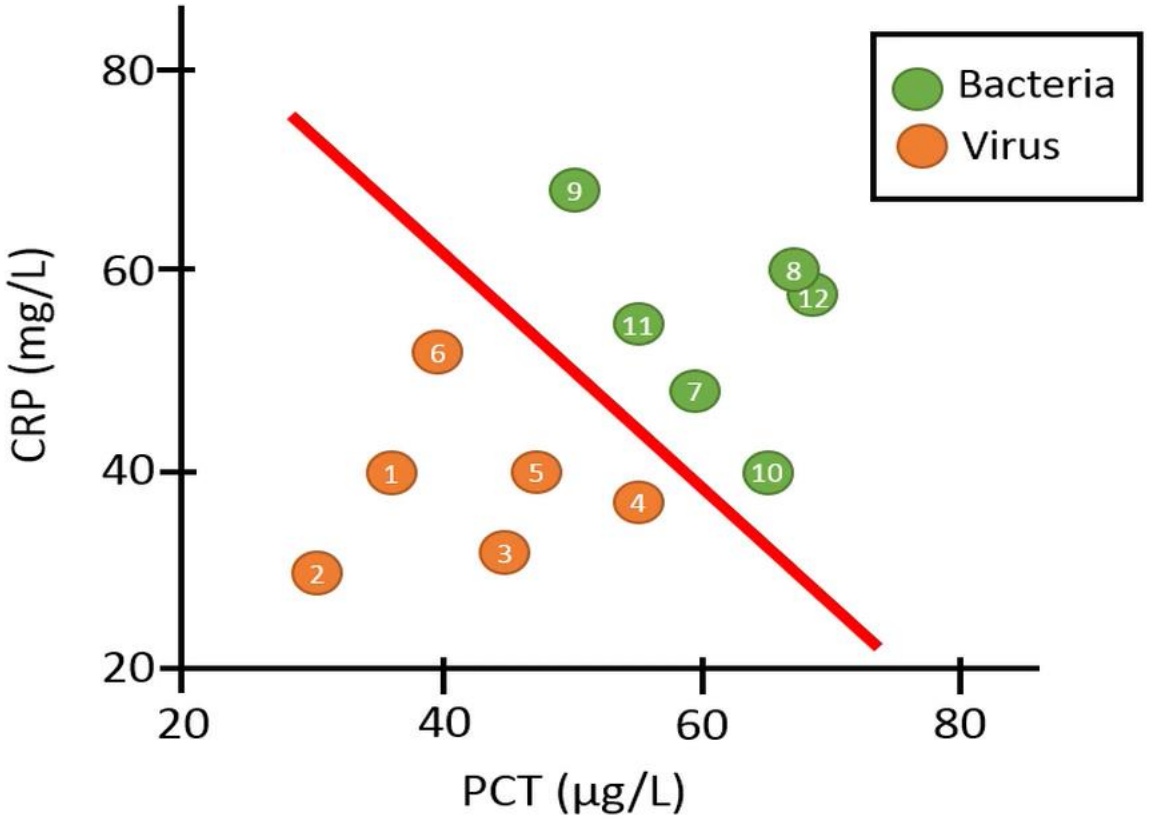
Infection	CRP (mg/L)	PCT (µg/L)
Viral	40	36
Viral	30	30
Viral	32	45
Viral	37	55
Viral	40	47
Viral	52	40
Bacterial	48	59
Bacterial	60	67
Bacterial	68	50
Bacterial	40	65
Bacterial	55	55
Bacterial	58	68



However, if we plot the CRP and the PCT concentration in the same plot,

Example data

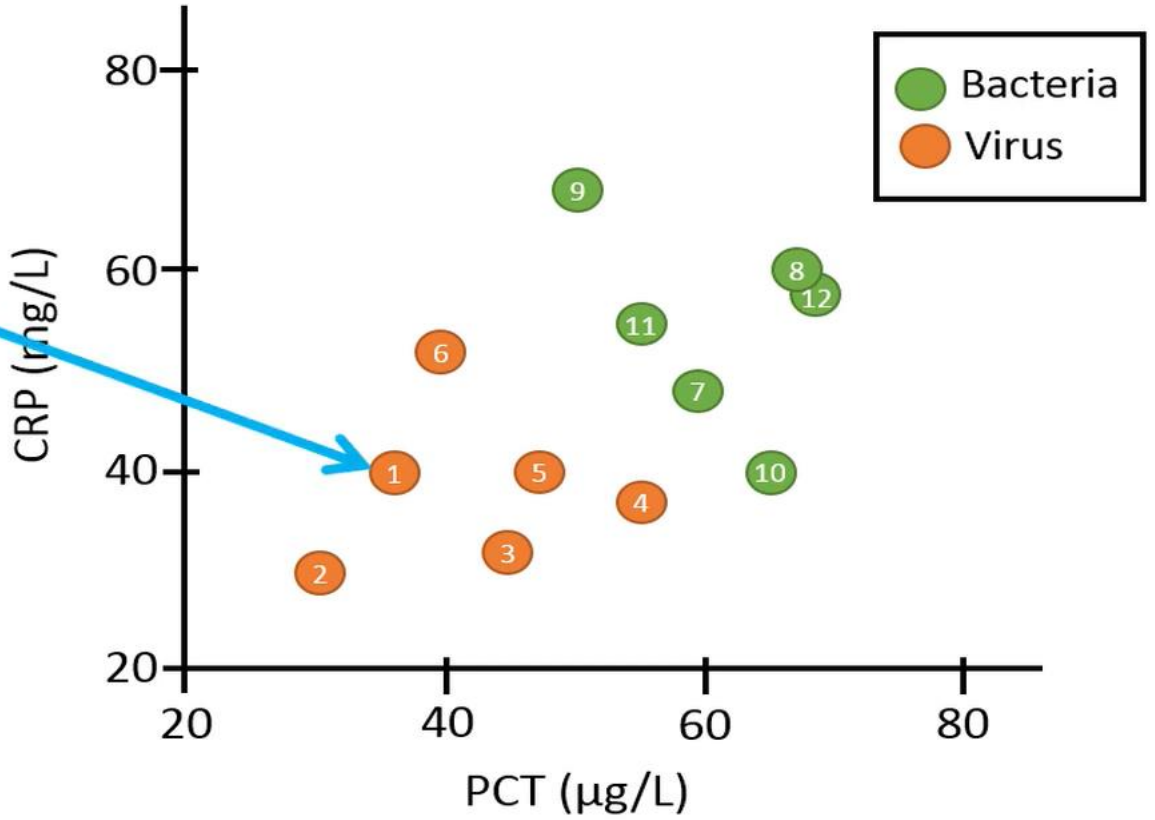
Infection	CRP (mg/L)	PCT (µg/L)
Viral	40	36
Viral	30	30
Viral	32	45
Viral	37	55
Viral	40	47
Viral	52	40
Bacterial	48	59
Bacterial	60	67
Bacterial	68	50
Bacterial	40	65
Bacterial	55	55
Bacterial	58	68



we can see that the following line can separate the two groups completely. This indicates that if make use of both variables simultaneously we can make a better prediction.

Example data

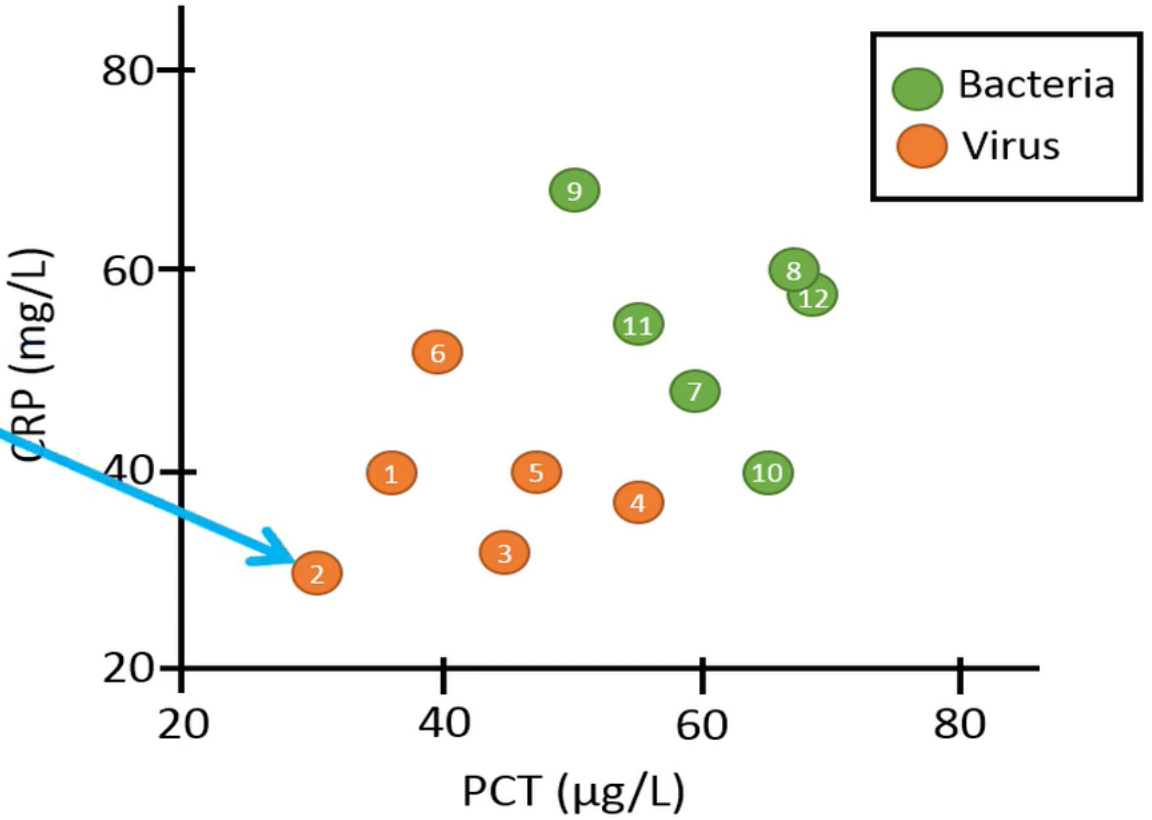
Infection	CRP (mg/L)	PCT (µg/L)
Viral	40	36
Viral	30	30
Viral	32	45
Viral	37	55
Viral	40	47
Viral	52	40
Bacterial	48	59
Bacterial	60	67
Bacterial	68	50
Bacterial	40	65
Bacterial	55	55
Bacterial	58	68



This data point represents the CRP and PCT concentration of patient number 1,

Example data

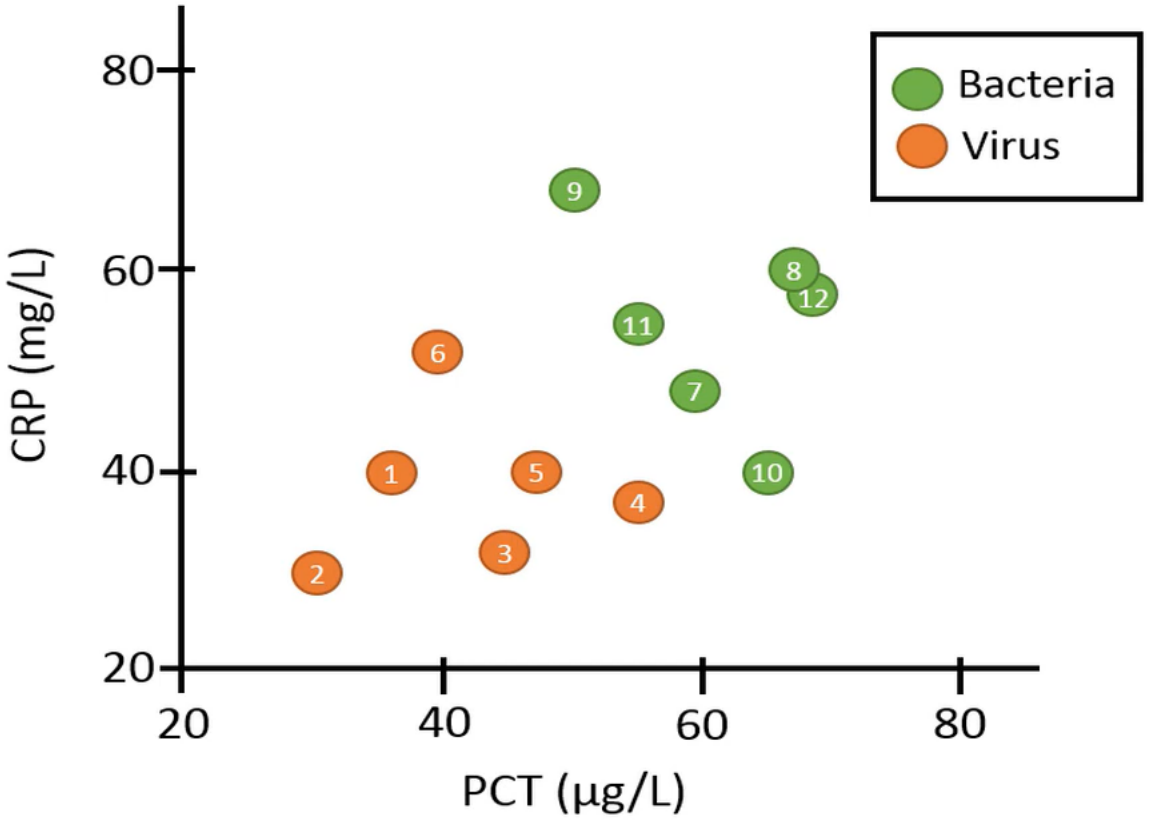
Infection	CRP (mg/L)	PCT (µg/L)
Viral	40	36
Viral	30	30
Viral	32	45
Viral	37	55
Viral	40	47
Viral	52	40
Bacterial	48	59
Bacterial	60	67
Bacterial	68	50
Bacterial	40	65
Bacterial	55	55
Bacterial	58	68



whereas this data point represents the CRP and PCT concentration of patient number 2 and so forth.

KNN

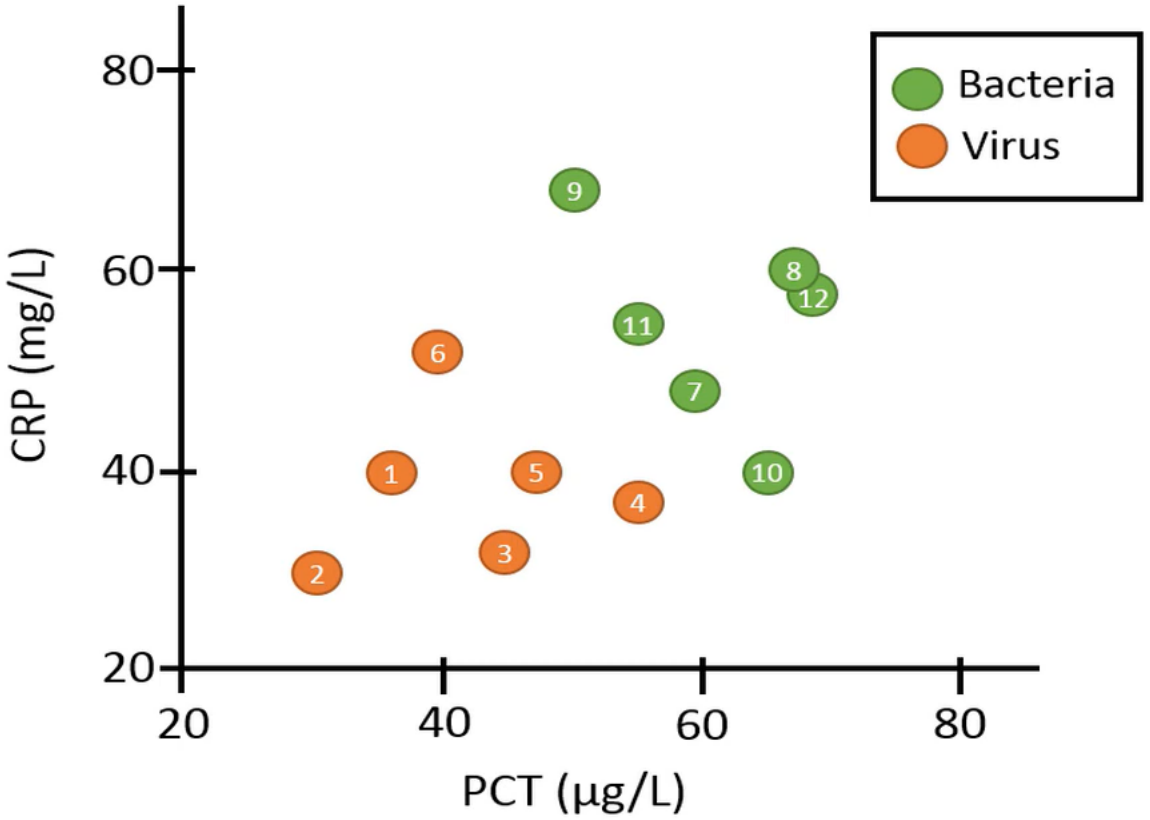
Infection	CRP (mg/L)	PCT (µg/L)
Viral	40	36
Viral	30	30
Viral	32	45
Viral	37	55
Viral	40	47
Viral	52	40
Bacterial	48	59
Bacterial	60	67
Bacterial	68	50
Bacterial	40	65
Bacterial	55	55
Bacterial	58	68



The KNN algorithm makes use of data with a known class or group when it makes predictions.

KNN

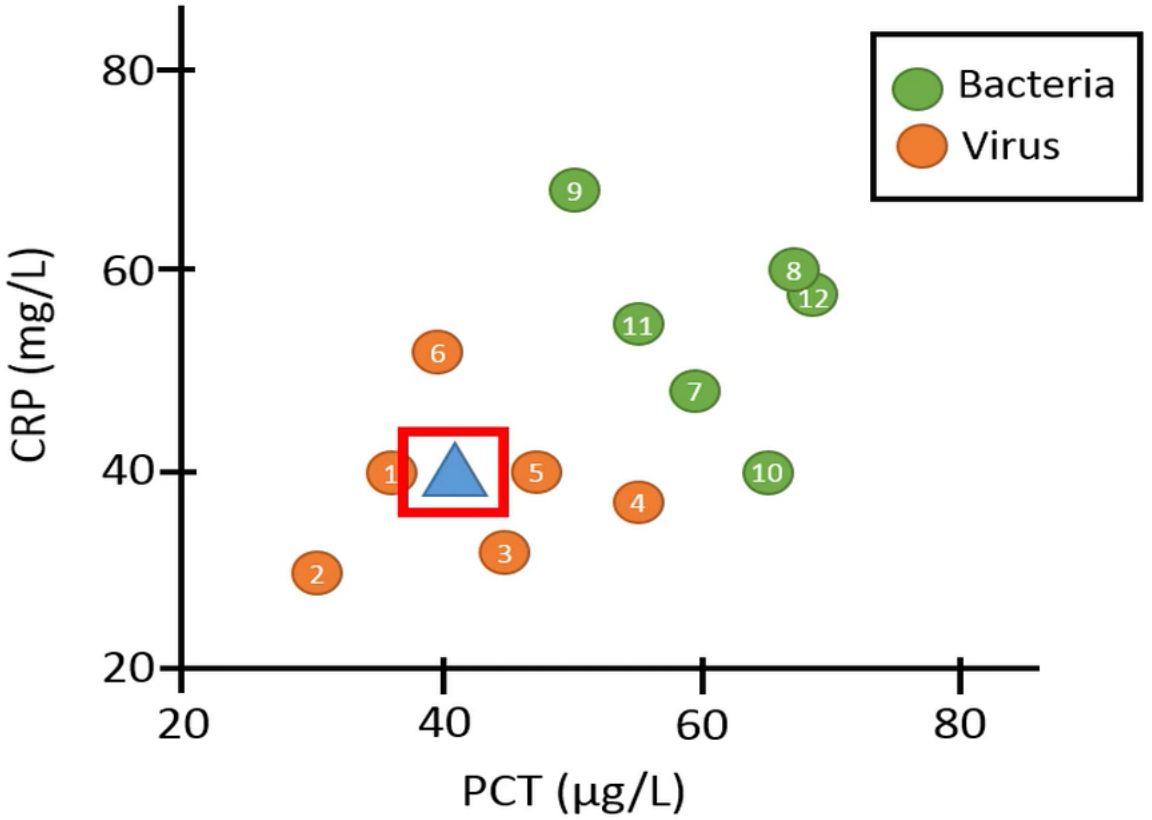
Infection	CRP (mg/L)	PCT (µg/L)
Viral	40	36
Viral	30	30
Viral	32	45
Viral	37	55
Viral	40	47
Viral	52	40
Bacterial	48	59
Bacterial	60	67
Bacterial	68	50
Bacterial	40	65
Bacterial	55	55
Bacterial	58	68



and we also know if they had a bacterial or viral infection.

KNN

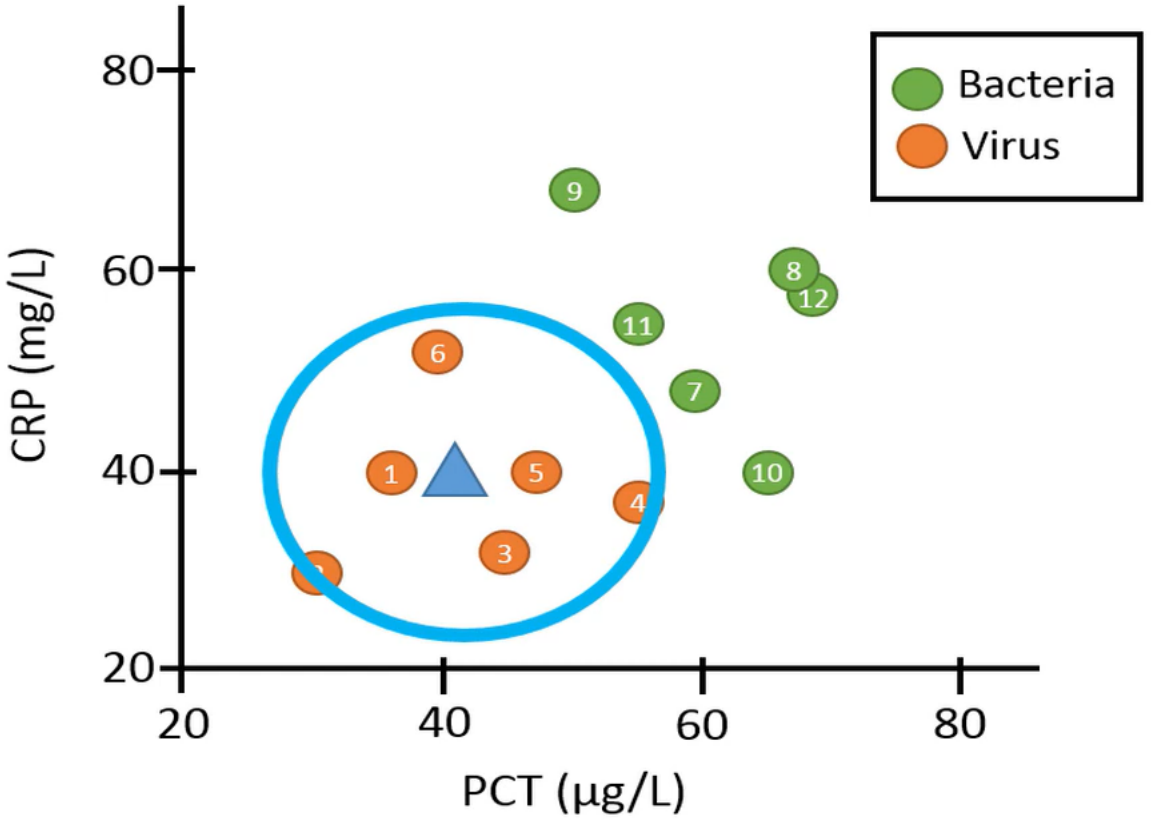
Infection	CRP (mg/L)	PCT (µg/L)
Viral	40	36
Viral	30	30
Viral	32	45
Viral	37	55
Viral	40	47
Viral	52	40
Bacterial	48	59
Bacterial	60	67
Bacterial	68	50
Bacterial	40	65
Bacterial	55	55
Bacterial	58	68



For example, let's say that we have a new patient that enters the hospital with a CRP concentration of 40, and a PCT concentration of 41,

KNN

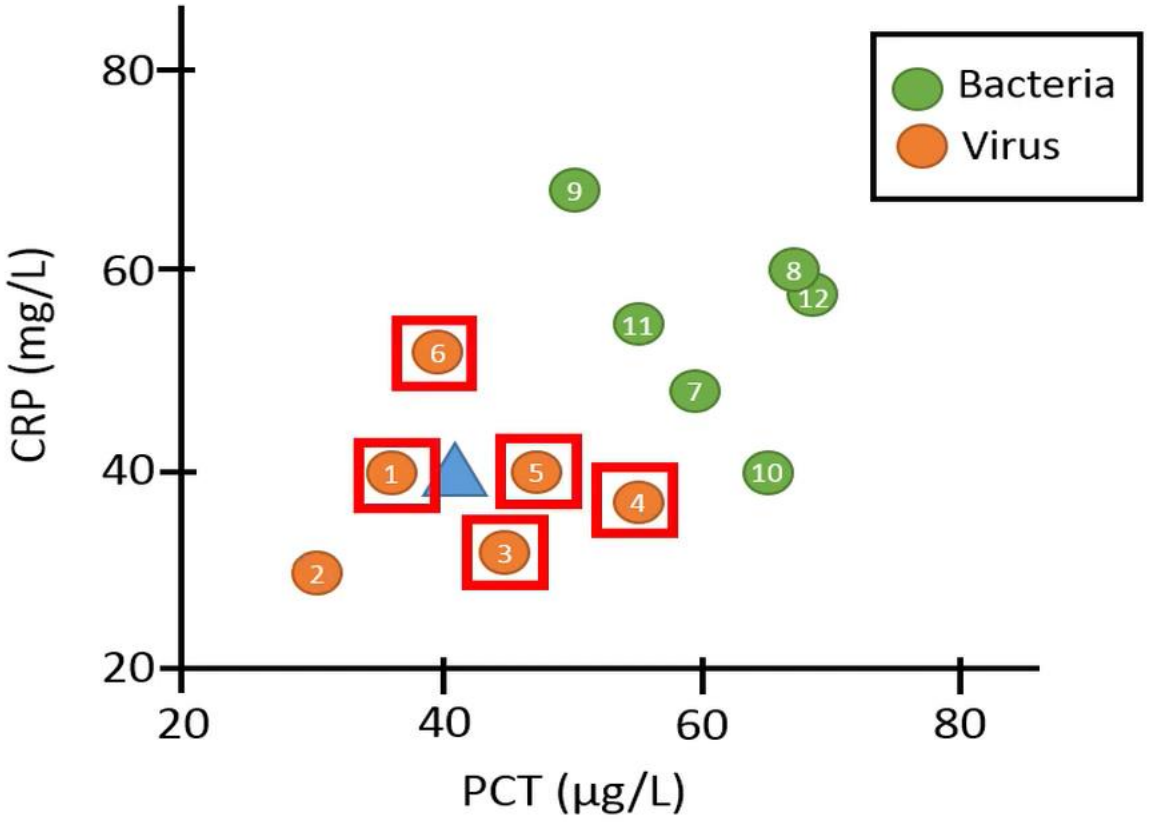
Infection	CRP (mg/L)	PCT (µg/L)
Viral	40	36
Viral	30	30
Viral	32	45
Viral	37	55
Viral	40	47
Viral	52	40
Bacterial	48	59
Bacterial	60	67
Bacterial	68	50
Bacterial	40	65
Bacterial	55	55
Bacterial	58	68



KNN then determines the class of the new observation based on the majority class of its *k* closest neighbors. For example, if *k* is set to five, the five closest neighbors will be evaluated.

KNN

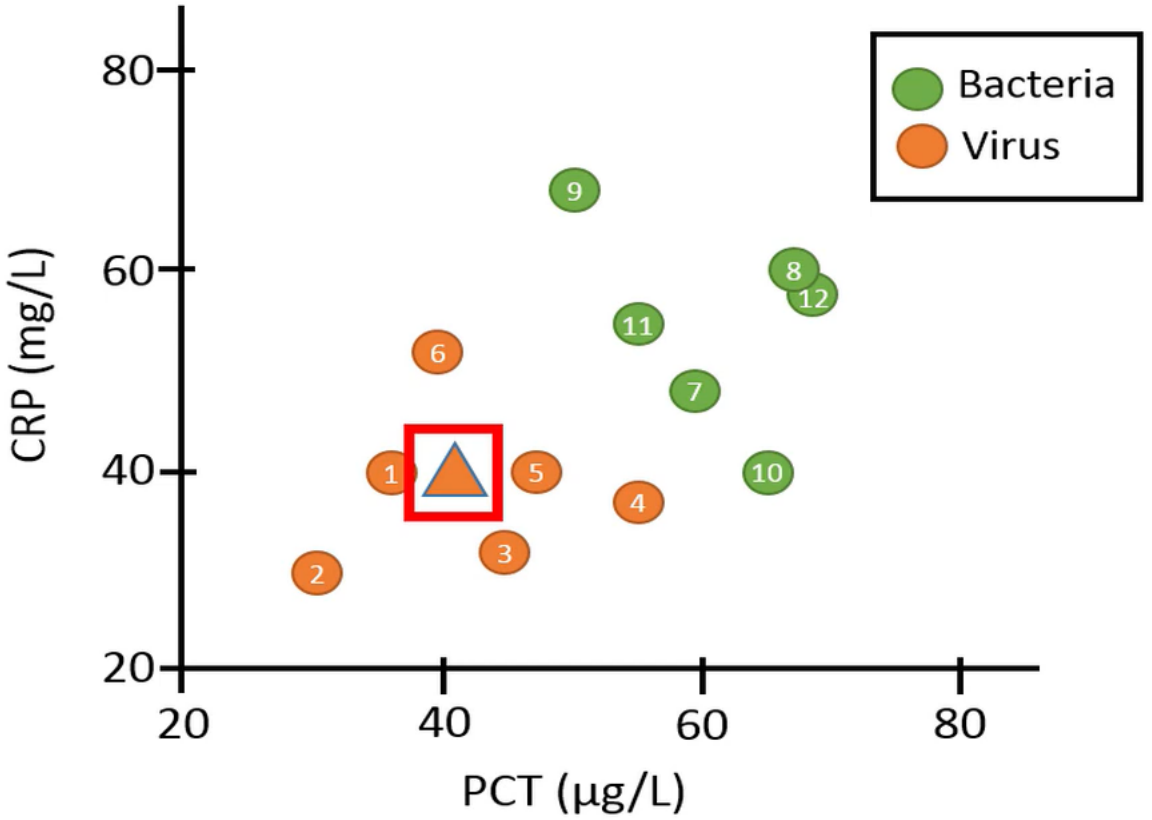
Infection	CRP (mg/L)	PCT (µg/L)
Viral	40	36
Viral	30	30
Viral	32	45
Viral	37	55
Viral	40	47
Viral	52	40
Bacterial	48	59
Bacterial	60	67
Bacterial	68	50
Bacterial	40	65
Bacterial	55	55
Bacterial	58	68



In this example, the five closest neighbors are patients which are known to have a viral infection. Since the majority of the five closest neighbors is of class “virus”,

KNN

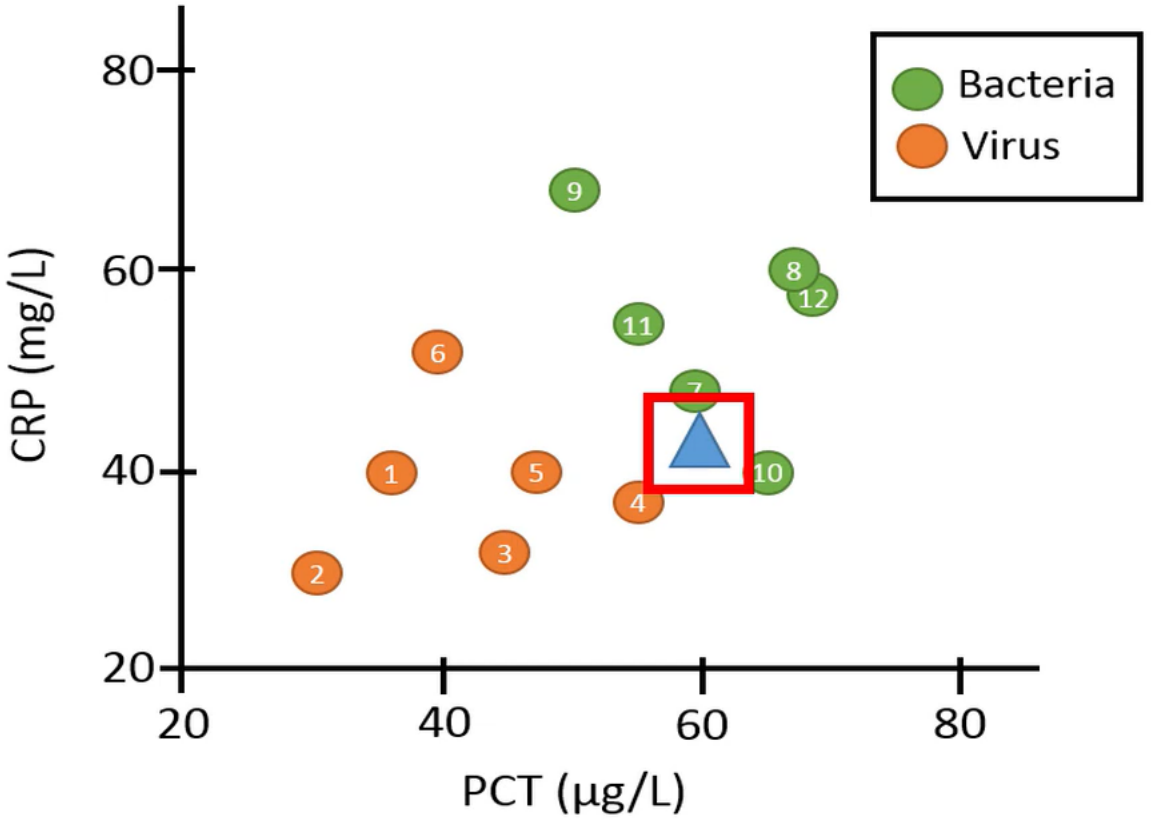
Infection	CRP (mg/L)	PCT (µg/L)
Viral	40	36
Viral	30	30
Viral	32	45
Viral	37	55
Viral	40	47
Viral	52	40
Bacterial	48	59
Bacterial	60	67
Bacterial	68	50
Bacterial	40	65
Bacterial	55	55
Bacterial	58	68



the patient with an unknown infection will be classified as having a viral infection.

KNN

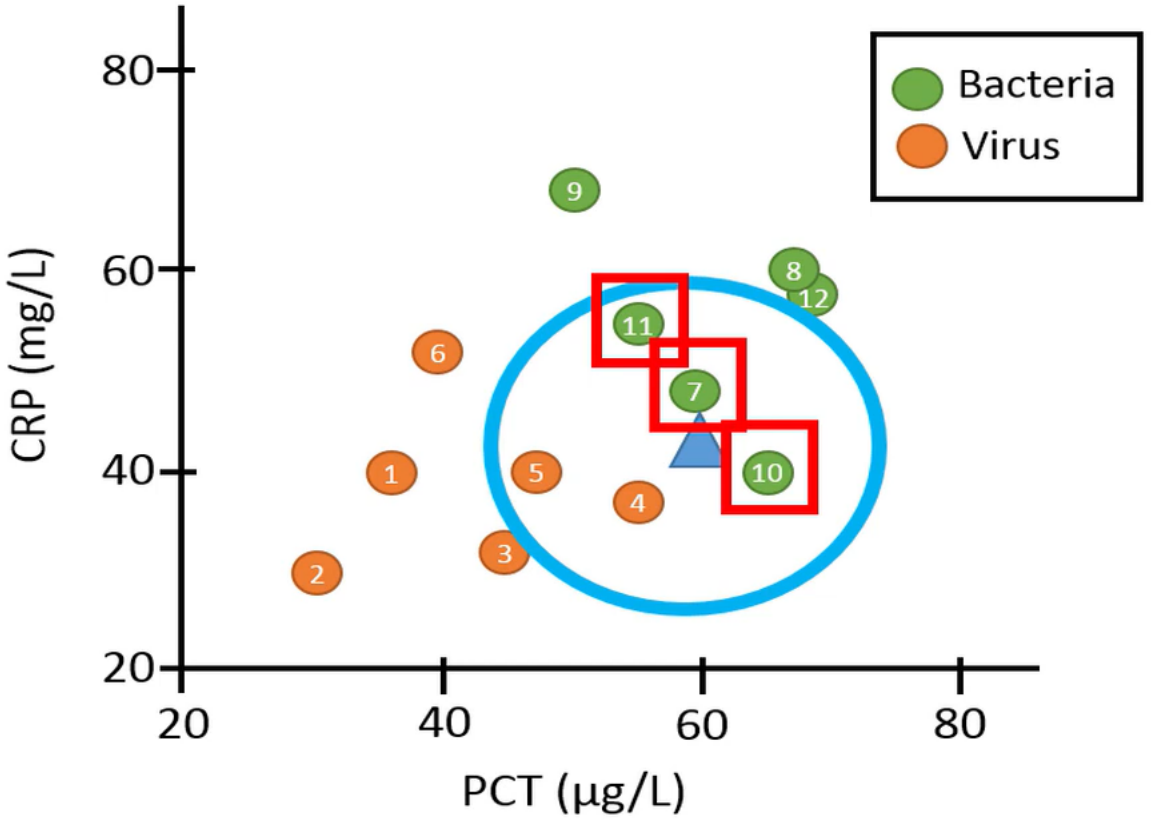
Infection	CRP (mg/L)	PCT (µg/L)
Viral	40	36
Viral	30	30
Viral	32	45
Viral	37	55
Viral	40	47
Viral	52	40
Bacterial	48	59
Bacterial	60	67
Bacterial	68	50
Bacterial	40	65
Bacterial	55	55
Bacterial	58	68



Now, suppose that the patient instead has a PCT concentration of 60 and a CRP concentration of 42.

KNN

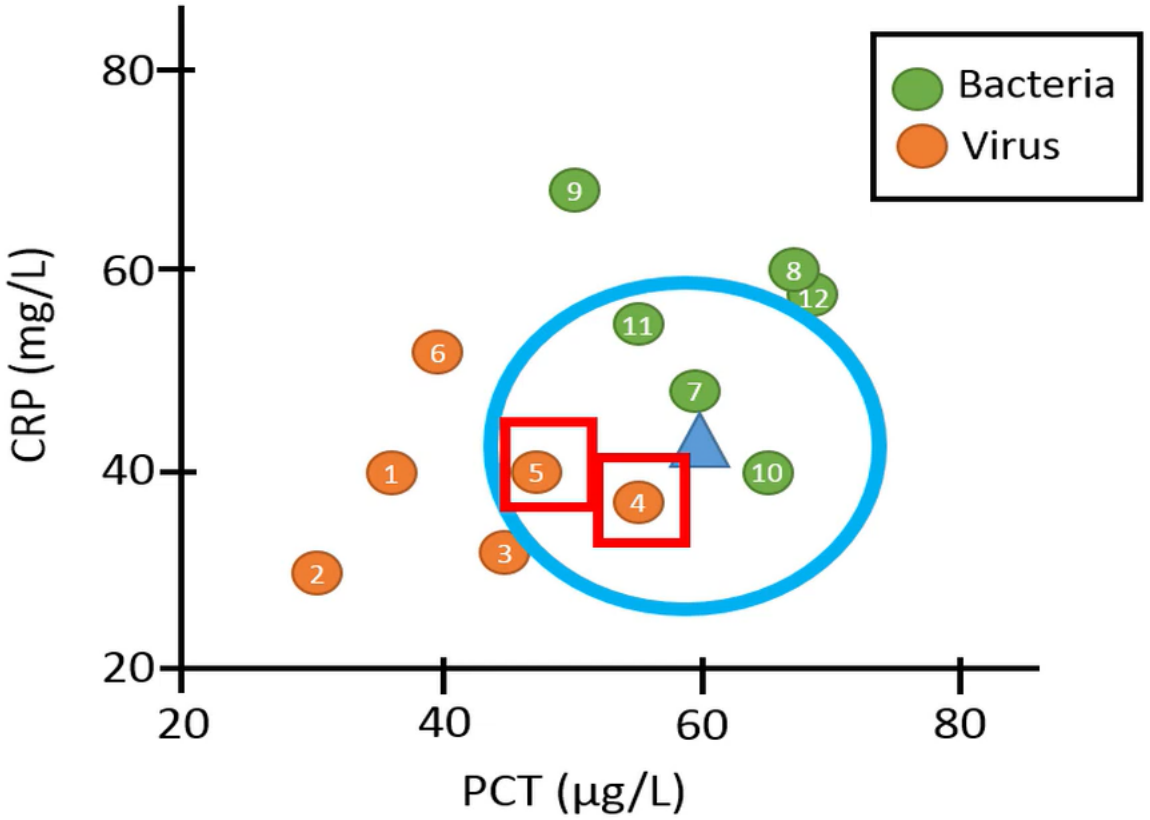
Infection	CRP (mg/L)	PCT (µg/L)
Viral	40	36
Viral	30	30
Viral	32	45
Viral	37	55
Viral	40	47
Viral	52	40
Bacterial	48	59
Bacterial	60	67
Bacterial	68	50
Bacterial	40	65
Bacterial	55	55
Bacterial	58	68



Since three out of the five closest neighbors are of class “bacteria”,

KNN

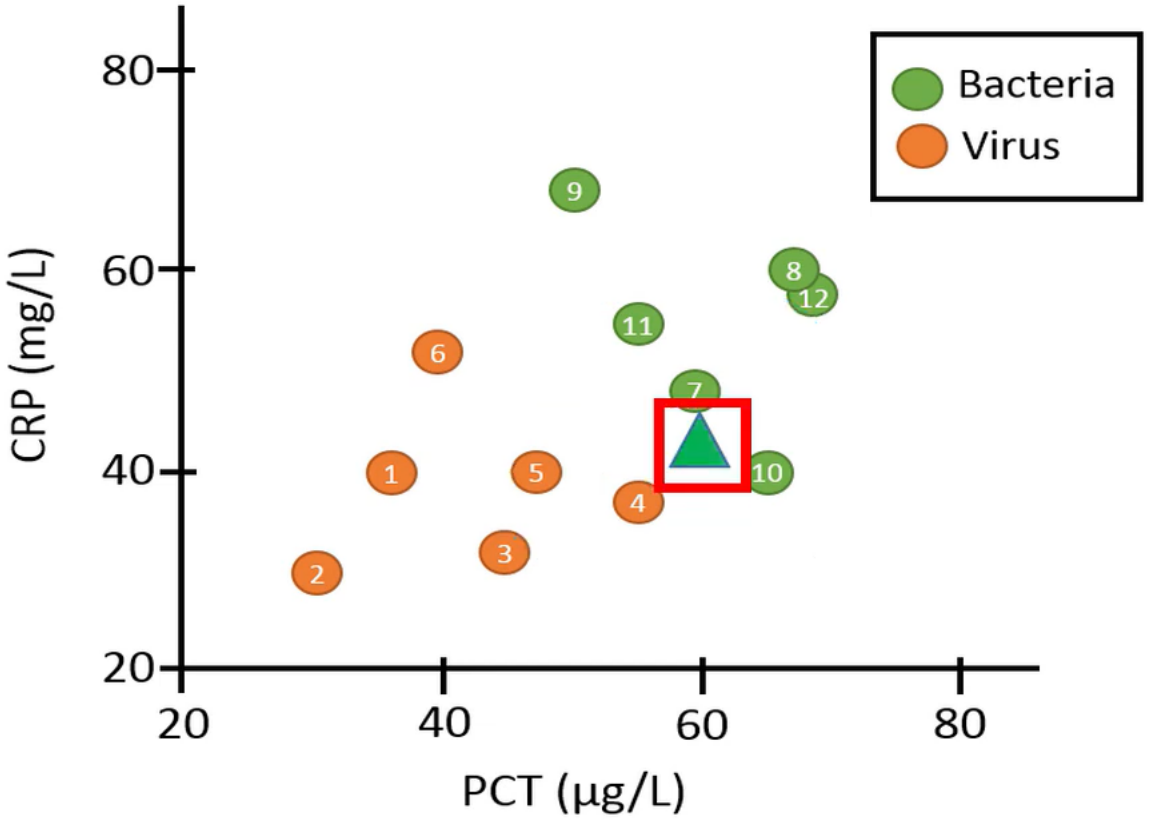
Infection	CRP (mg/L)	PCT (µg/L)
Viral	40	36
Viral	30	30
Viral	32	45
Viral	37	55
Viral	40	47
Viral	52	40
Bacterial	48	59
Bacterial	60	67
Bacterial	68	50
Bacterial	40	65
Bacterial	55	55
Bacterial	58	68



whereas only two are of class “virus”,

KNN

Infection	CRP (mg/L)	PCT (µg/L)
Viral	40	36
Viral	30	30
Viral	32	45
Viral	37	55
Viral	40	47
Viral	52	40
Bacterial	48	59
Bacterial	60	67
Bacterial	68	50
Bacterial	40	65
Bacterial	55	55
Bacterial	58	68



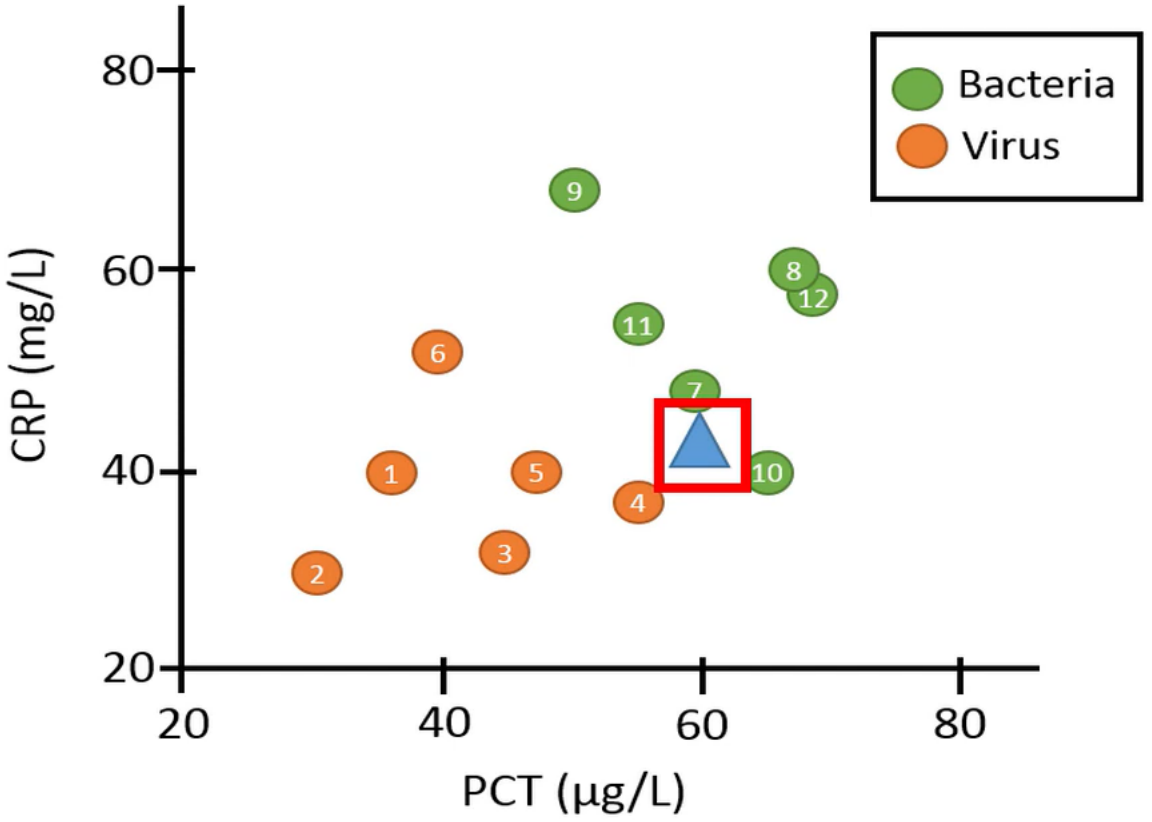
the patient will be predicted to have a bacterial infection because the majority of the neighbors are of class bacteria.

KNN

1. Determine the distance (e.g. Euclidean distance) between the new observation and all the data points in the training set.
2. Sort the distances.
3. Identify the *k* closest neighbors.
4. Determine the class of the new observation based on the group majority of the *k* nearest neighbors.

1. Determine the distance between the new observation and all the data points in the training set.

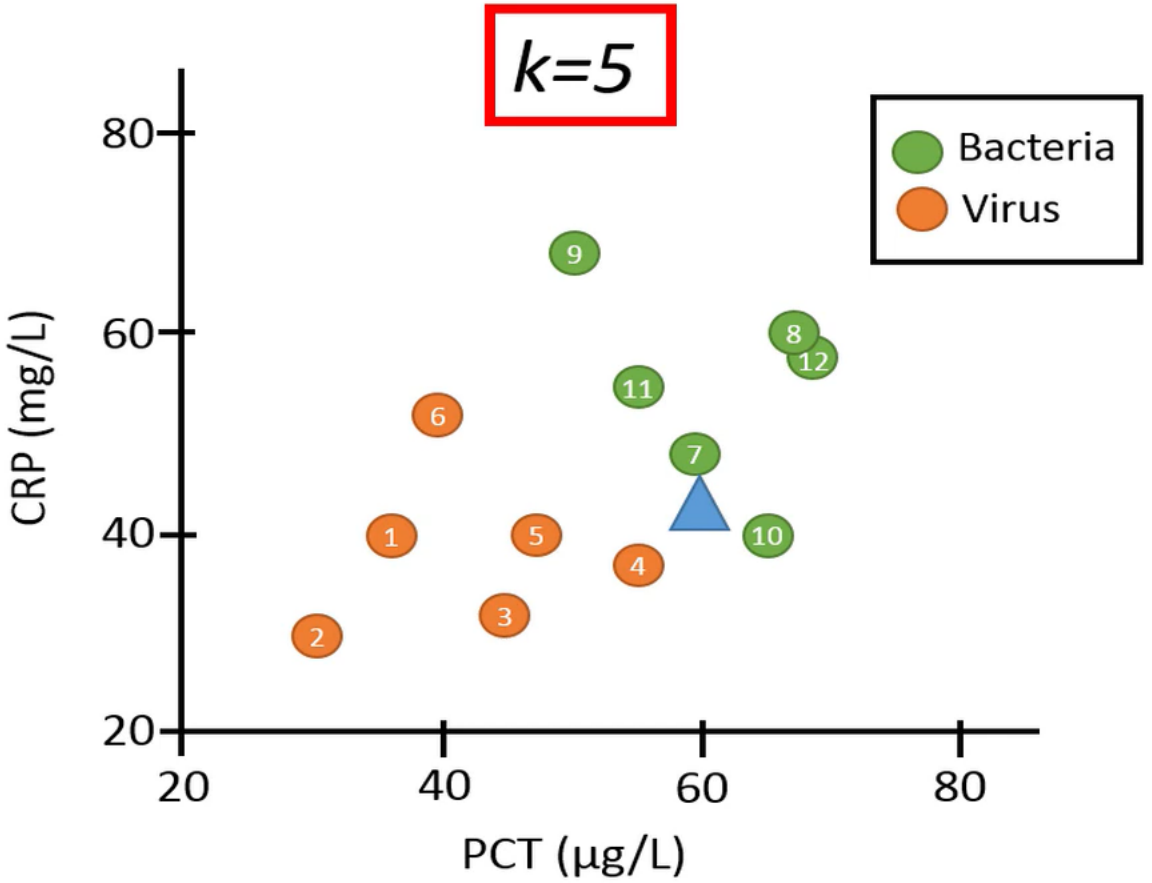
Infection	CRP (mg/L)	PCT (µg/L)	Distance
Viral	40	36	
Viral	30	30	
Viral	32	45	
Viral	37	55	
Viral	40	47	
Viral	52	40	
Bacterial	48	59	
Bacterial	60	67	
Bacterial	68	50	
Bacterial	40	65	
Bacterial	55	55	
Bacterial	58	68	



Let's follow these steps on our example data where we like to predict the class of a new observation with a PCT concentration of 60 and a CRP concentration of 42.

1. Determine the distance between the new observation and all the data points in the training set.

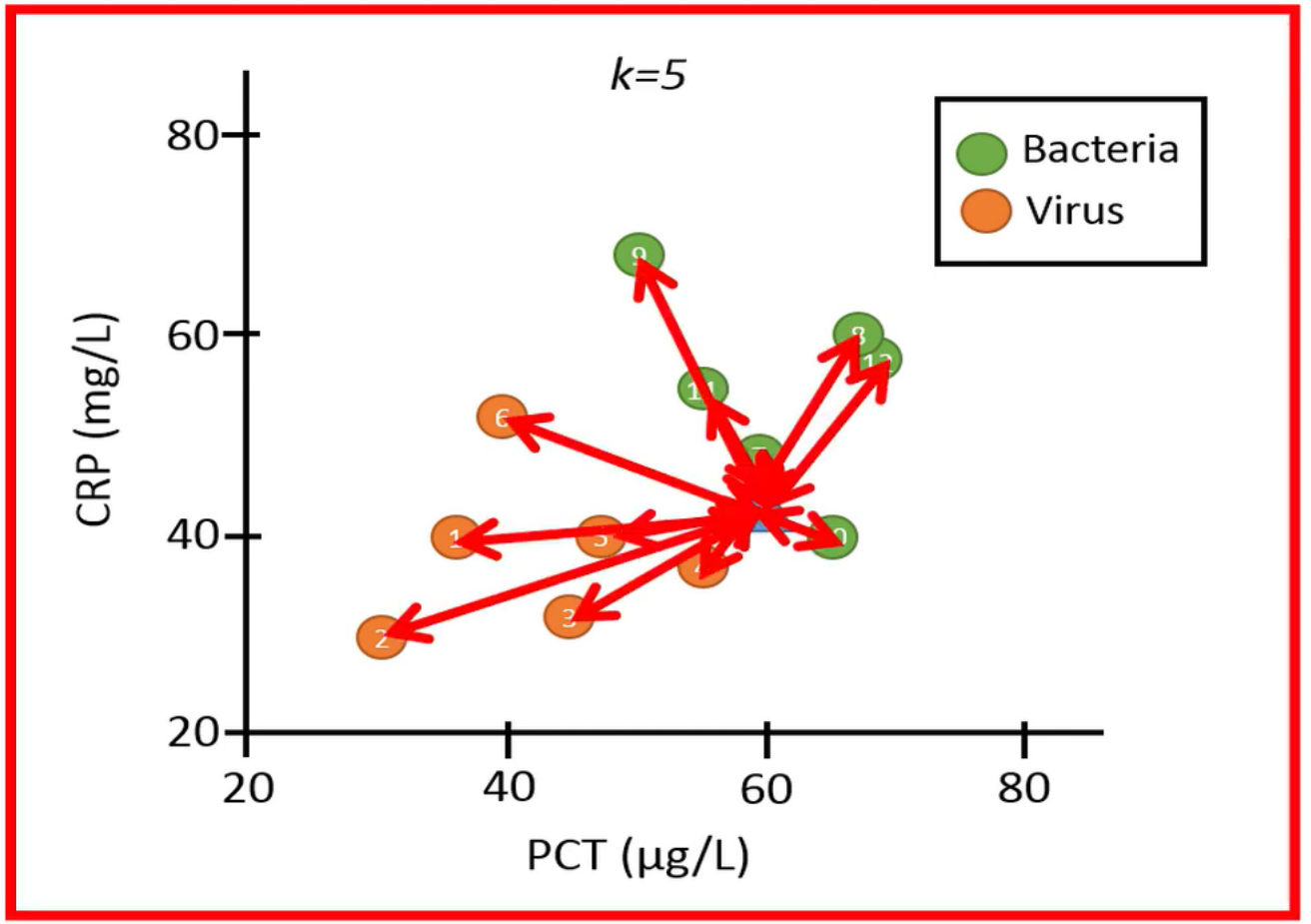
Infection	CRP (mg/L)	PCT (µg/L)	Distance
Viral	40	36	
Viral	30	30	
Viral	32	45	
Viral	37	55	
Viral	40	47	
Viral	52	40	
Bacterial	48	59	
Bacterial	60	67	
Bacterial	68	50	
Bacterial	40	65	
Bacterial	55	55	
Bacterial	58	68	



In this example, we set the value of *k* to five, which means that we check the class of the 5 closest neighbors.

1. Determine the distance between the new observation and all the data points in the training set.

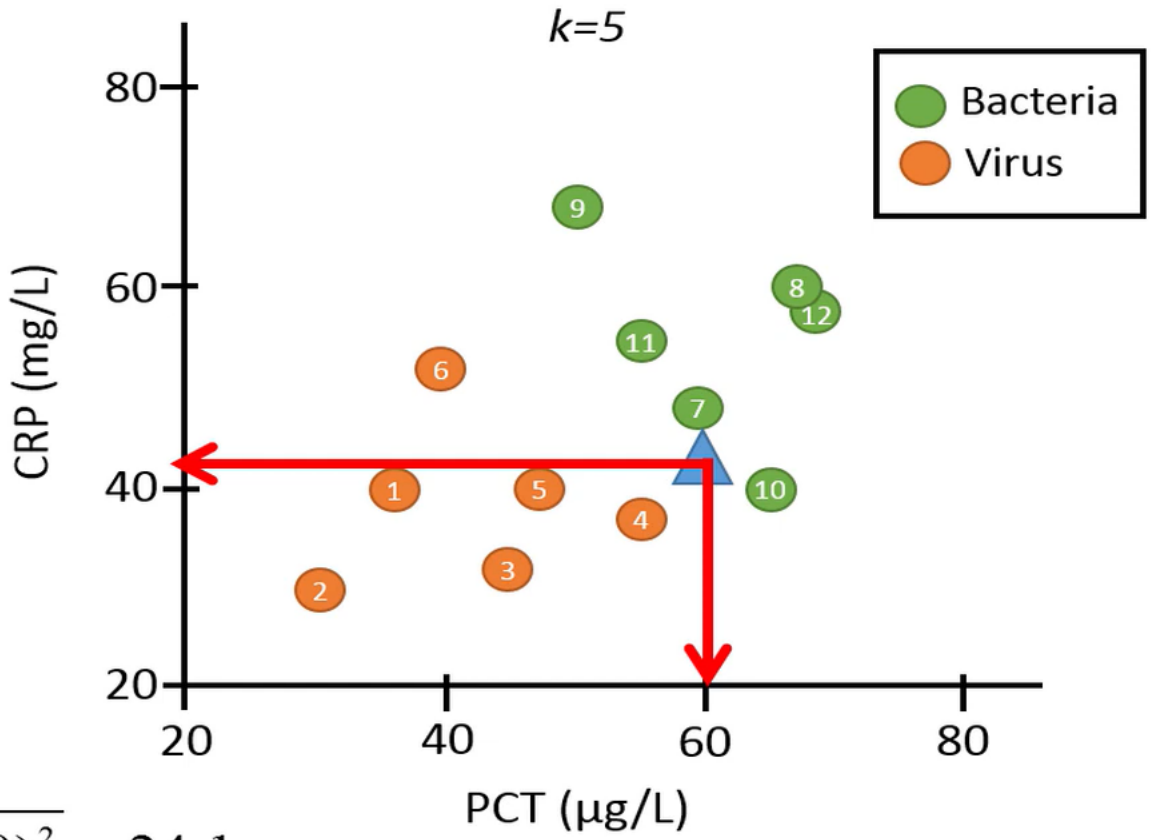
Infection	CRP (mg/L)	PCT (µg/L)	Distance
Viral	40	36	
Viral	30	30	
Viral	32	45	
Viral	37	55	
Viral	40	47	
Viral	52	40	
Bacterial	48	59	
Bacterial	60	67	
Bacterial	68	50	
Bacterial	40	65	
Bacterial	55	55	
Bacterial	58	68	



We begin by calculating the Euclidean distance to all the data points.

1. Determine the distance between the new observation and all the data points in the training set.

Infection	CRP (mg/L)	PCT (µg/L)	Distance
Viral	40	36	
Viral	30	30	
Viral	32	45	
Viral	37	55	
Viral	40	47	
Viral	52	40	
Bacterial	48	59	
Bacterial	60	67	
Bacterial	68	50	
Bacterial	40	65	
Bacterial	55	55	
Bacterial	58	68	

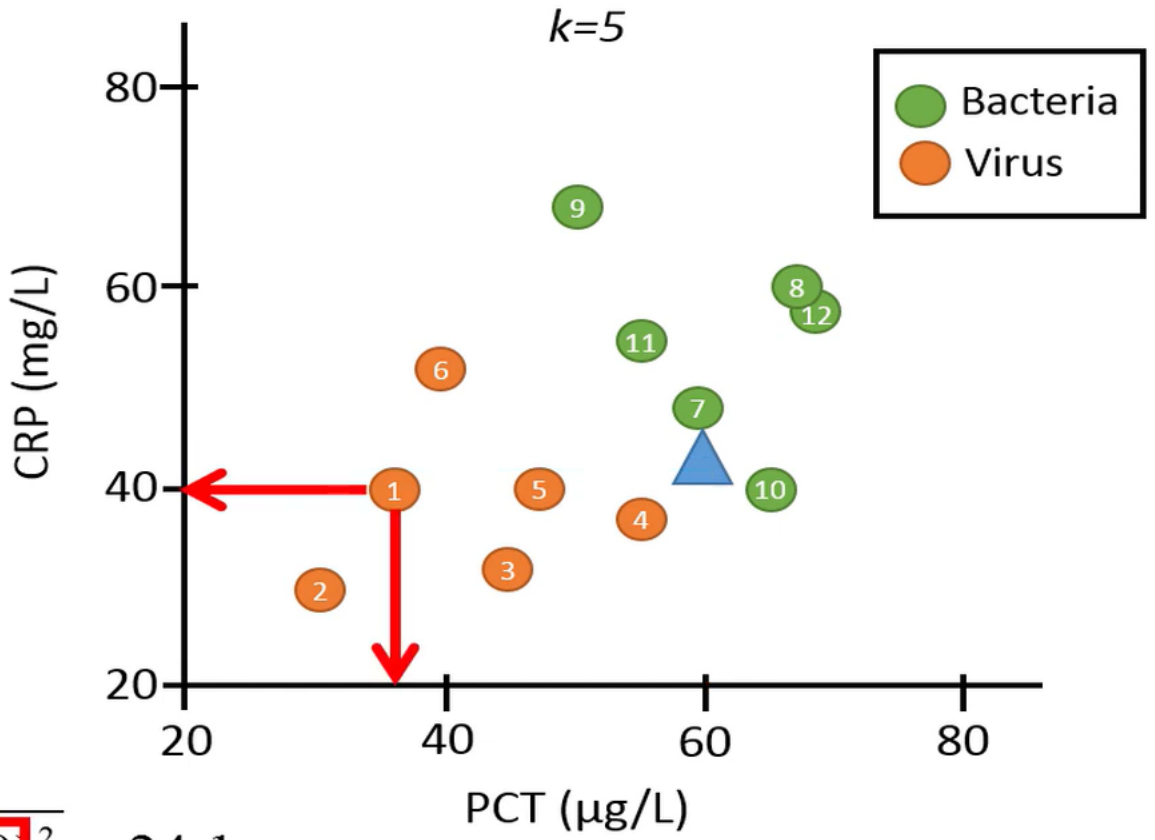


$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} = \sqrt{(60 - 36)^2 + (42 - 40)^2} = 24.1$$

We plug in the x and y coordinates of the new data point,

1. Determine the distance between the new observation and all the data points in the training set.

Infection	CRP (mg/L)	PCT (µg/L)	Distance
Viral	40	36	
Viral	30	30	
Viral	32	45	
Viral	37	55	
Viral	40	47	
Viral	52	40	
Bacterial	48	59	
Bacterial	60	67	
Bacterial	68	50	
Bacterial	40	65	
Bacterial	55	55	
Bacterial	58	68	

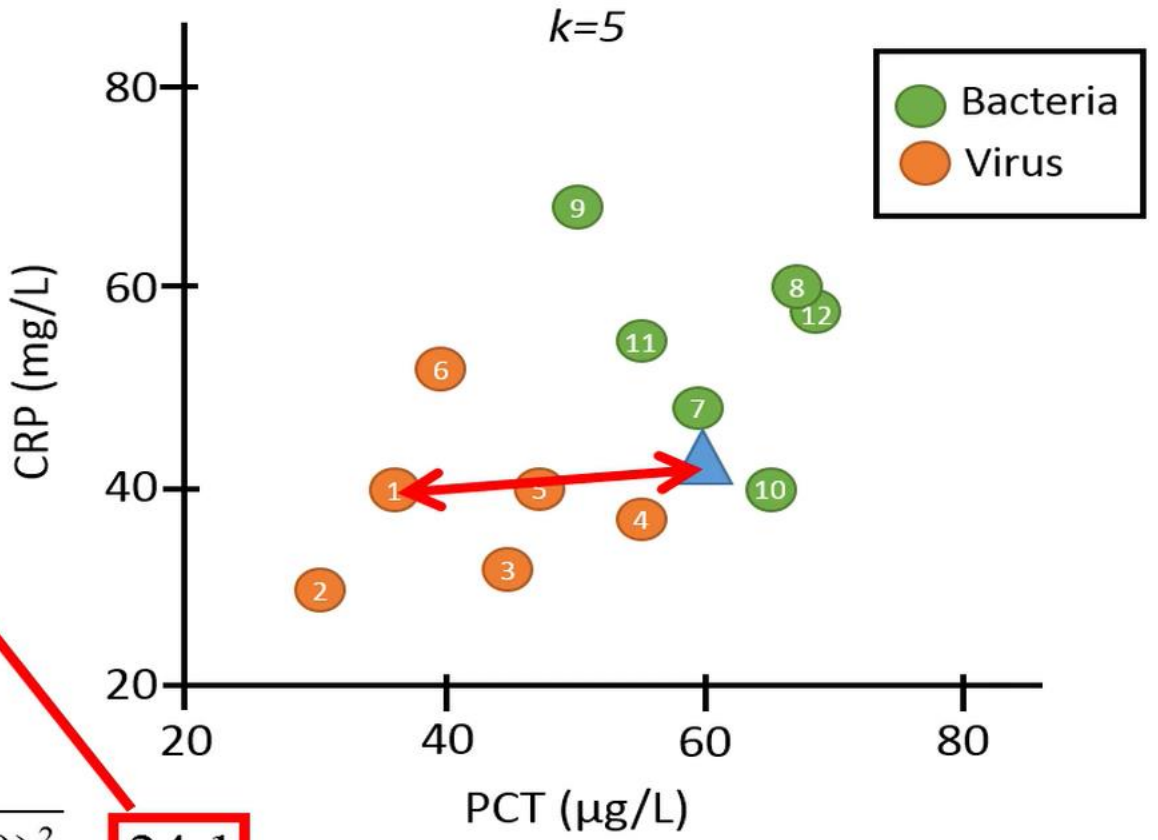


$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} = \sqrt{(60 - 36)^2 + (42 - 40)^2} = 24.1$$

and for data point number 1.

1. Determine the distance between the new observation and all the data points in the training set.

Infection	CRP (mg/L)	PCT (µg/L)	Distance
Viral	40	36	24.1
Viral	30	30	
Viral	32	45	
Viral	37	55	
Viral	40	47	
Viral	52	40	
Bacterial	48	59	
Bacterial	60	67	
Bacterial	68	50	
Bacterial	40	65	
Bacterial	55	55	
Bacterial	58	68	

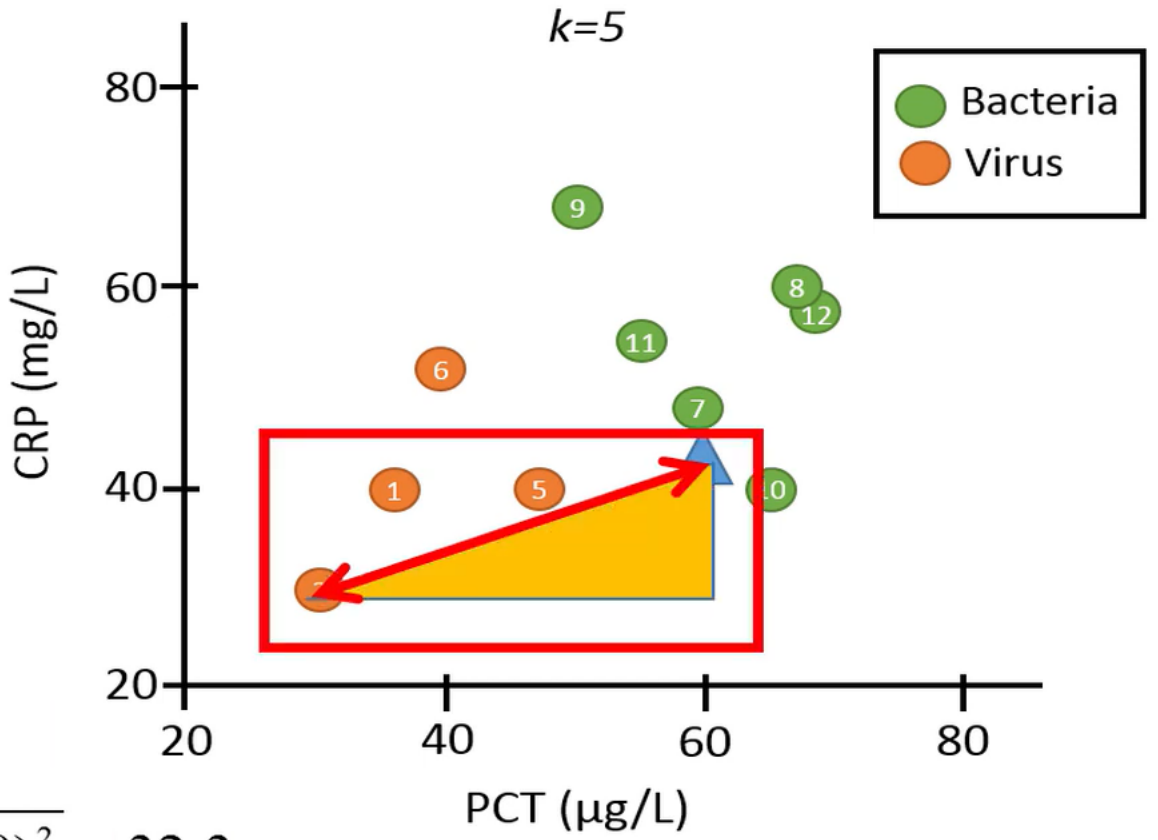


$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} = \sqrt{(60 - 36)^2 + (42 - 40)^2} = 24.1$$

We fill in the distance in the table.

1. Determine the distance between the new observation and all the data points in the training set.

Infection	CRP (mg/L)	PCT (µg/L)	Distance
Viral	40	36	24.1
Viral	30	30	32.3
Viral	32	45	
Viral	37	55	
Viral	40	47	
Viral	52	40	
Bacterial	48	59	
Bacterial	60	67	
Bacterial	68	50	
Bacterial	40	65	
Bacterial	55	55	
Bacterial	58	68	

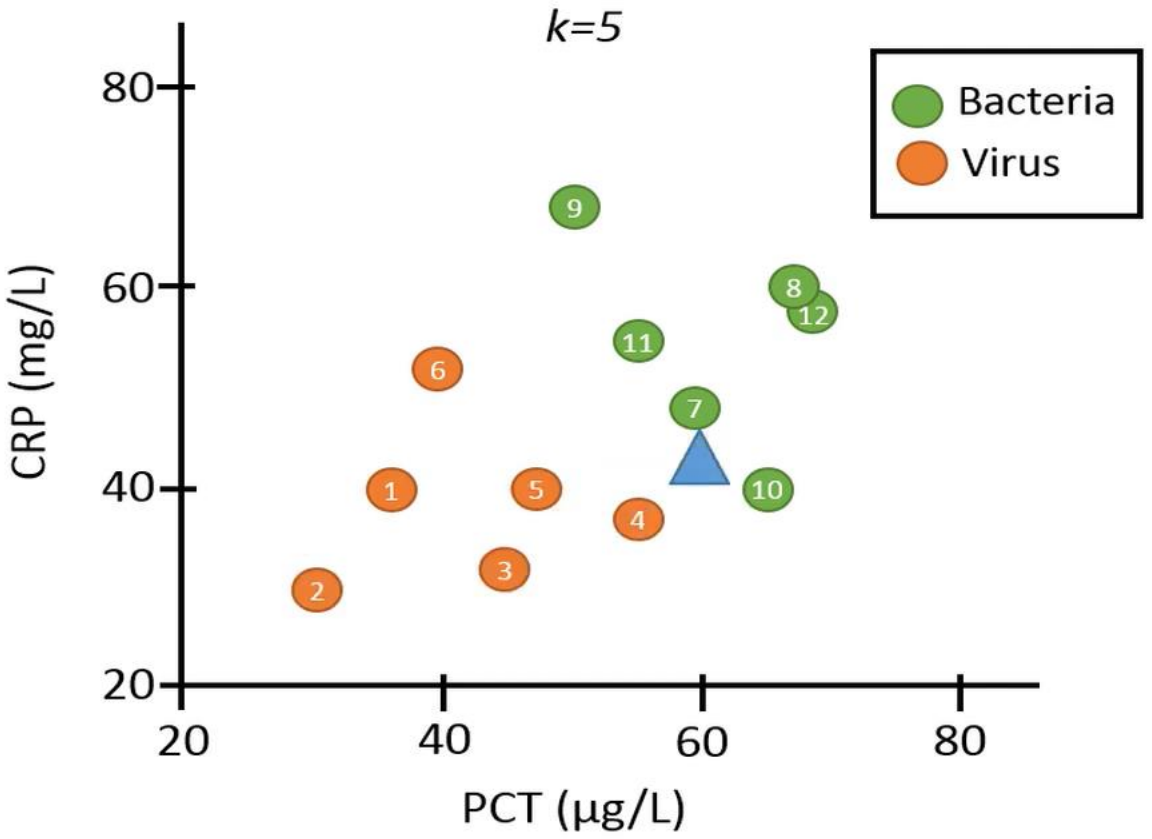


$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} = \sqrt{(60 - 30)^2 + (42 - 30)^2} = 32.3$$

Note that the Euclidean distance in two dimensions can be seen as applying the Pythagoras' theorem to a right triangle.

2. Sort the distances

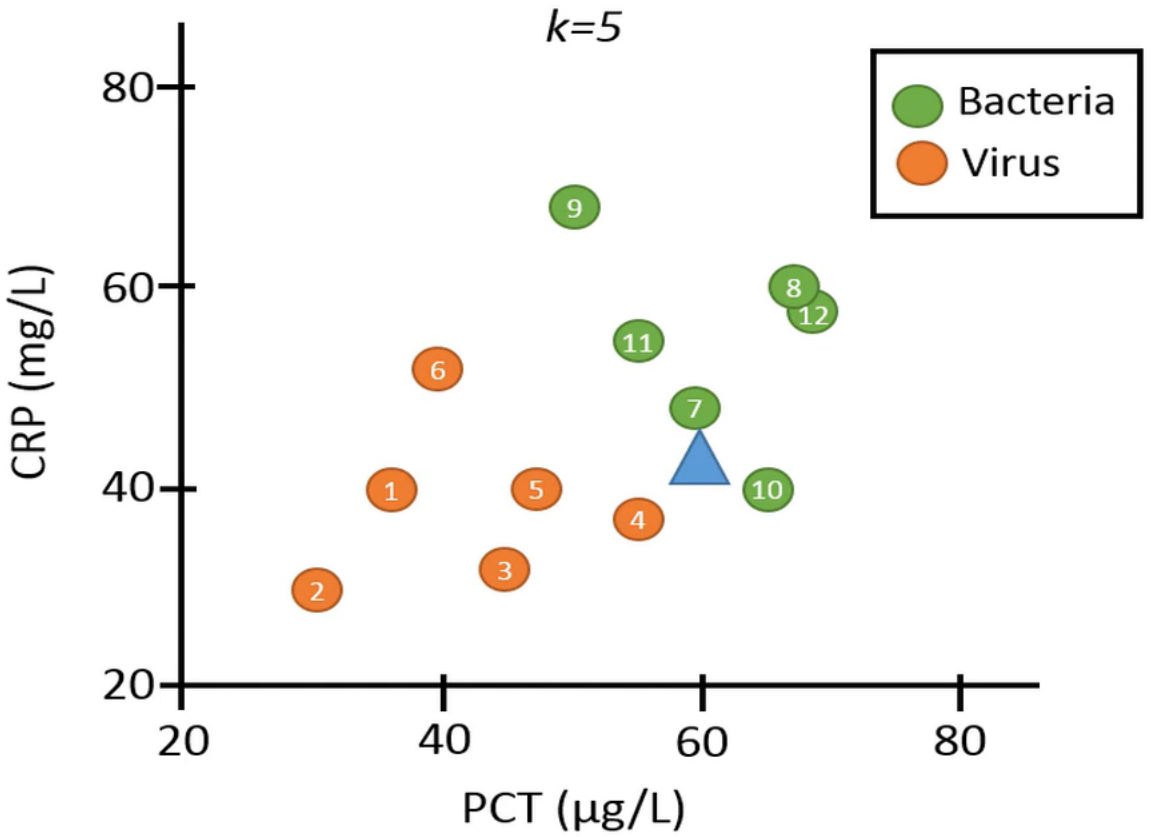
Infection	CRP (mg/L)	PCT (µg/L)	Distance
Viral	40	36	24.1
Viral	30	30	32.3
Viral	32	45	18.0
Viral	37	55	7.1
Viral	40	47	13.2
Viral	52	40	22.4
Bacterial	48	59	6.1
Bacterial	60	67	19.3
Bacterial	68	50	27.9
Bacterial	40	65	5.4
Bacterial	55	55	13.9
Bacterial	58	68	17.9



Once we have calculated the Euclidean distances between the new observation and all the data points, we will sort this table based on the distances.

2. Sort the distances

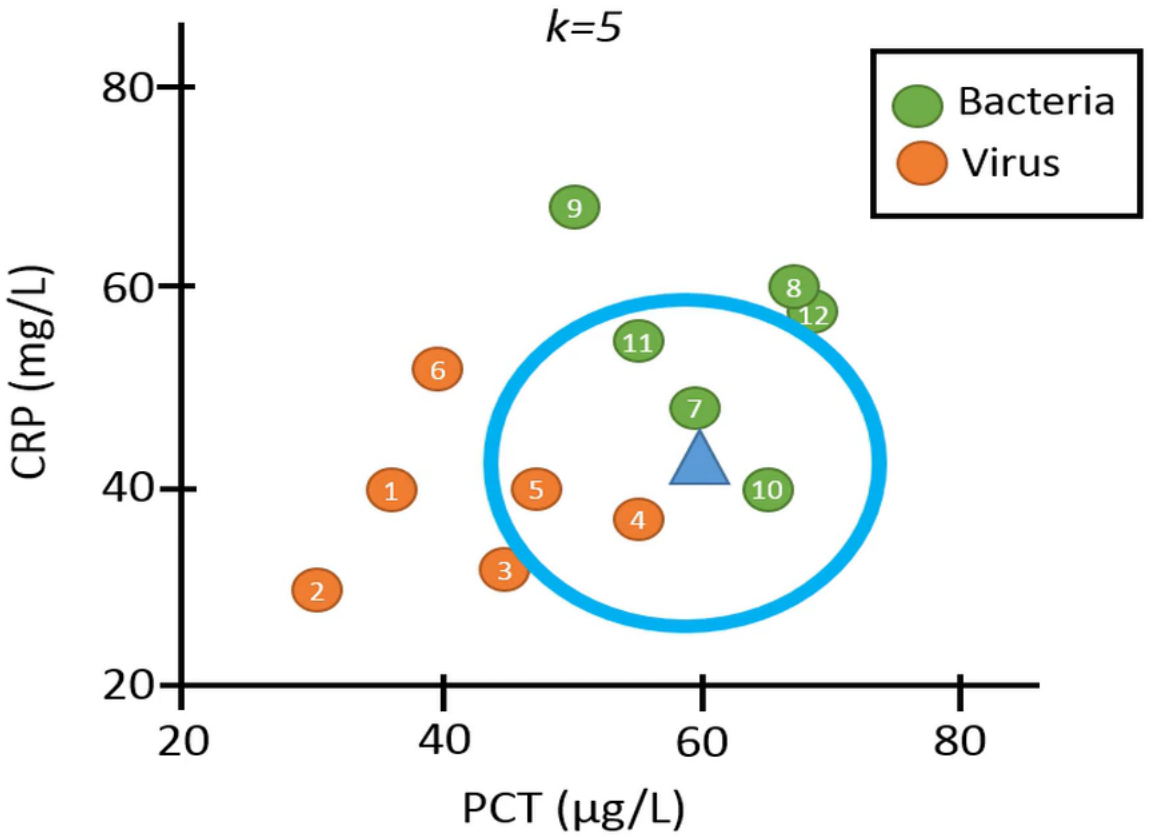
Infection	CRP (mg/L)	PCT (µg/L)	Distance
Bacterial	40	65	5.4
Bacterial	48	59	6.1
Viral	37	55	7.1
Viral	40	47	13.2
Bacterial	55	55	13.9
Bacterial	58	68	17.9
Viral	32	45	18.0
Bacterial	60	67	19.3
Viral	52	40	22.4
Viral	40	36	24.1
Bacterial	68	50	27.9
Viral	30	30	32.3



After we have sorted the patients based on the distances to the new observation,

3. Identify the *k* closest neighbors.

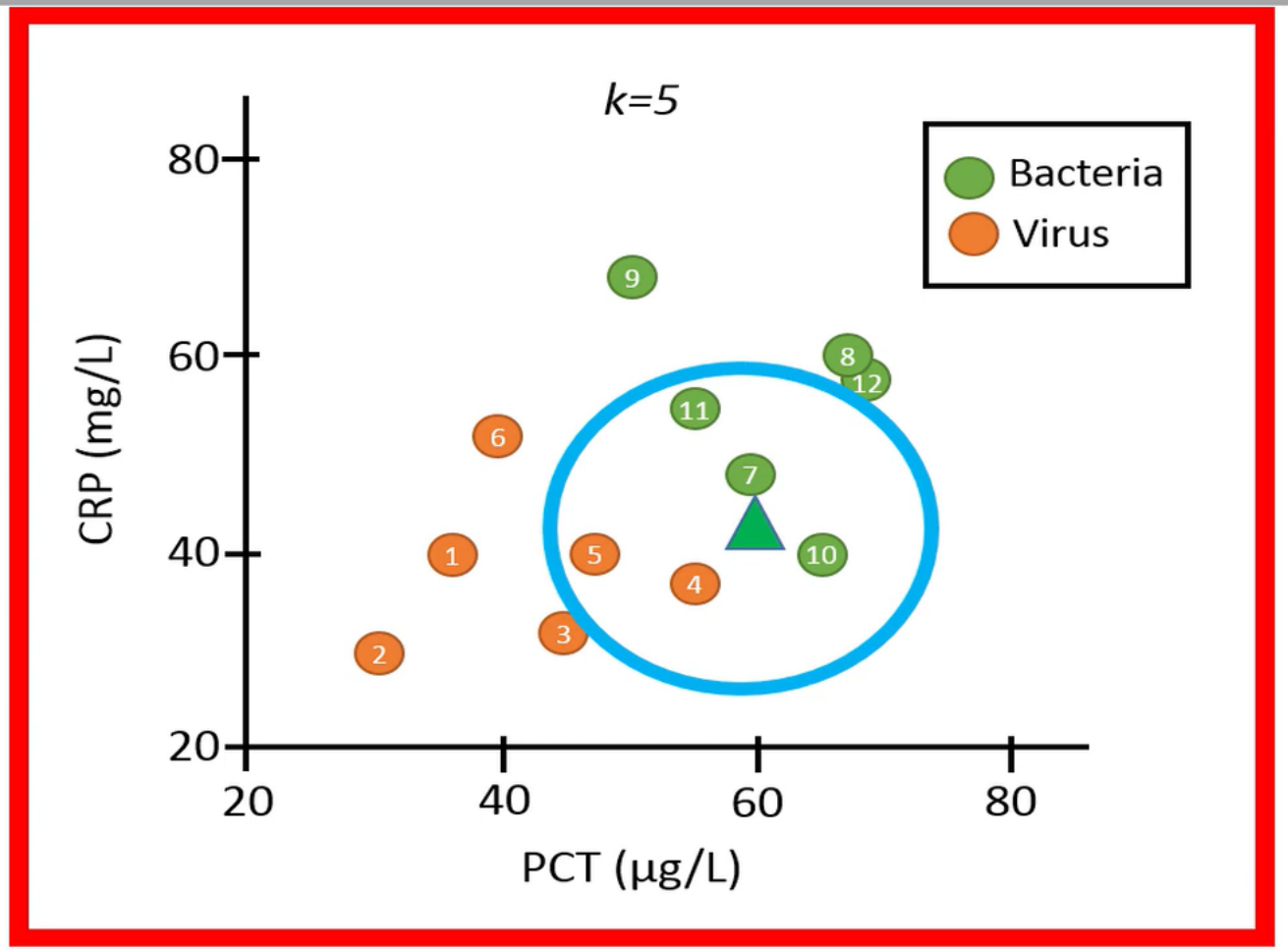
Infection	CRP (mg/L)	PCT (µg/L)	Distance
Bacterial	40	65	5.4
Bacterial	48	59	6.1
Viral	37	55	7.1
Viral	40	47	13.2
Bacterial	55	55	13.9
Bacterial	58	68	17.9
Viral	32	45	18.0
Bacterial	60	67	19.3
Viral	52	40	22.4
Viral	40	36	24.1
Bacterial	68	50	27.9
Viral	30	30	32.3



we see that three out of the five closest neighbors are of class “bacteria” whereas two are of class “virus”.

4. Determine the class of the new observation

Infection	CRP (mg/L)	PCT (µg/L)	Distance
Bacterial	40	65	5.4
Bacterial	48	59	6.1
Viral	37	55	7.1
Viral	40	47	13.2
Bacterial	55	55	13.9
Bacterial	58	68	17.9
Viral	32	45	18.0
Bacterial	60	67	19.3
Viral	52	40	22.4
Viral	40	36	24.1
Bacterial	68	50	27.9
Viral	30	30	32.3



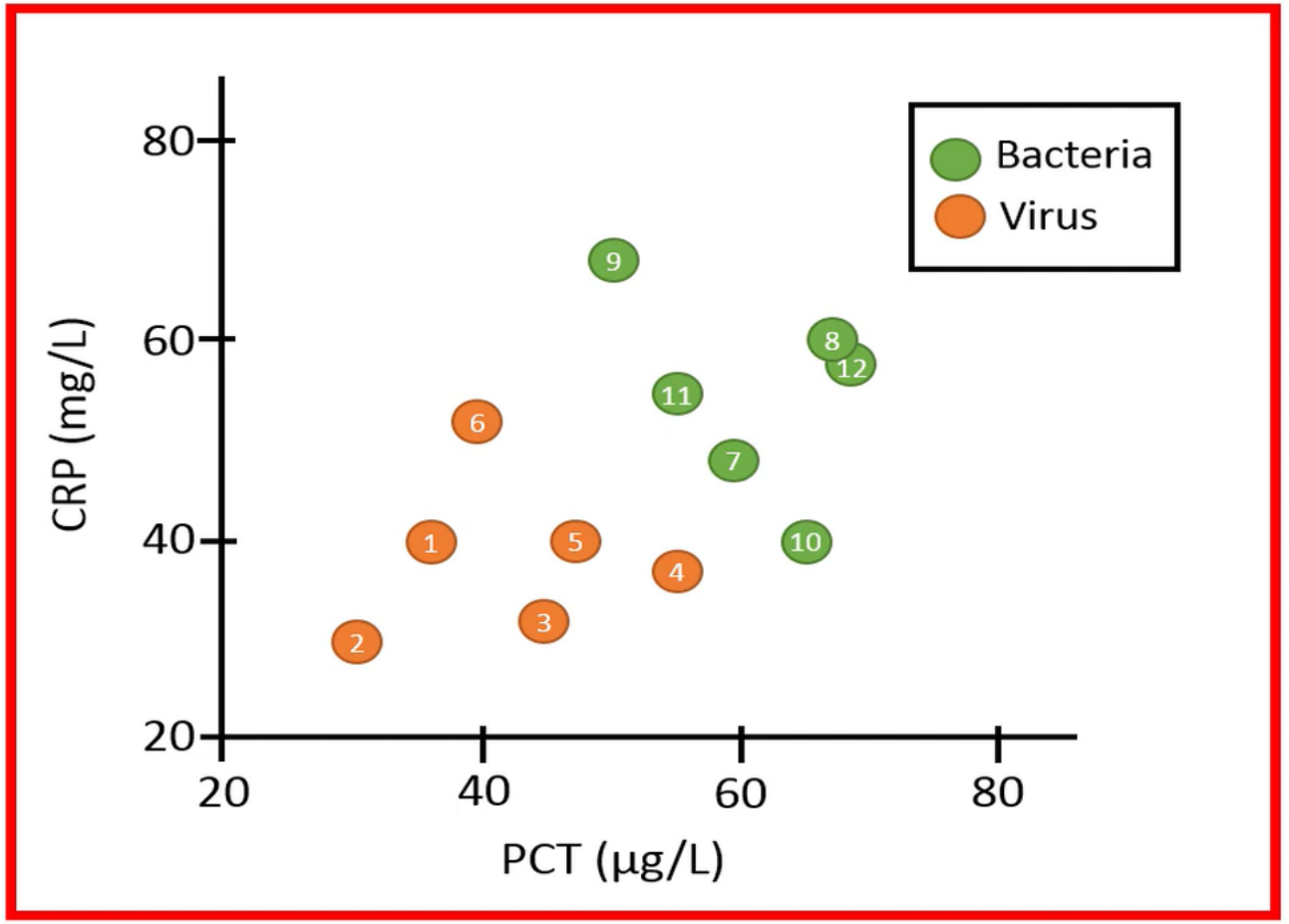
Since the majority of the *k* nearest neighbors are of class “bacteria”, we classify the new observation as “bacteria”. This means that we predict that the patient has a bacterial infection.

KNN

How good is the classifier?

How good is the classifier?

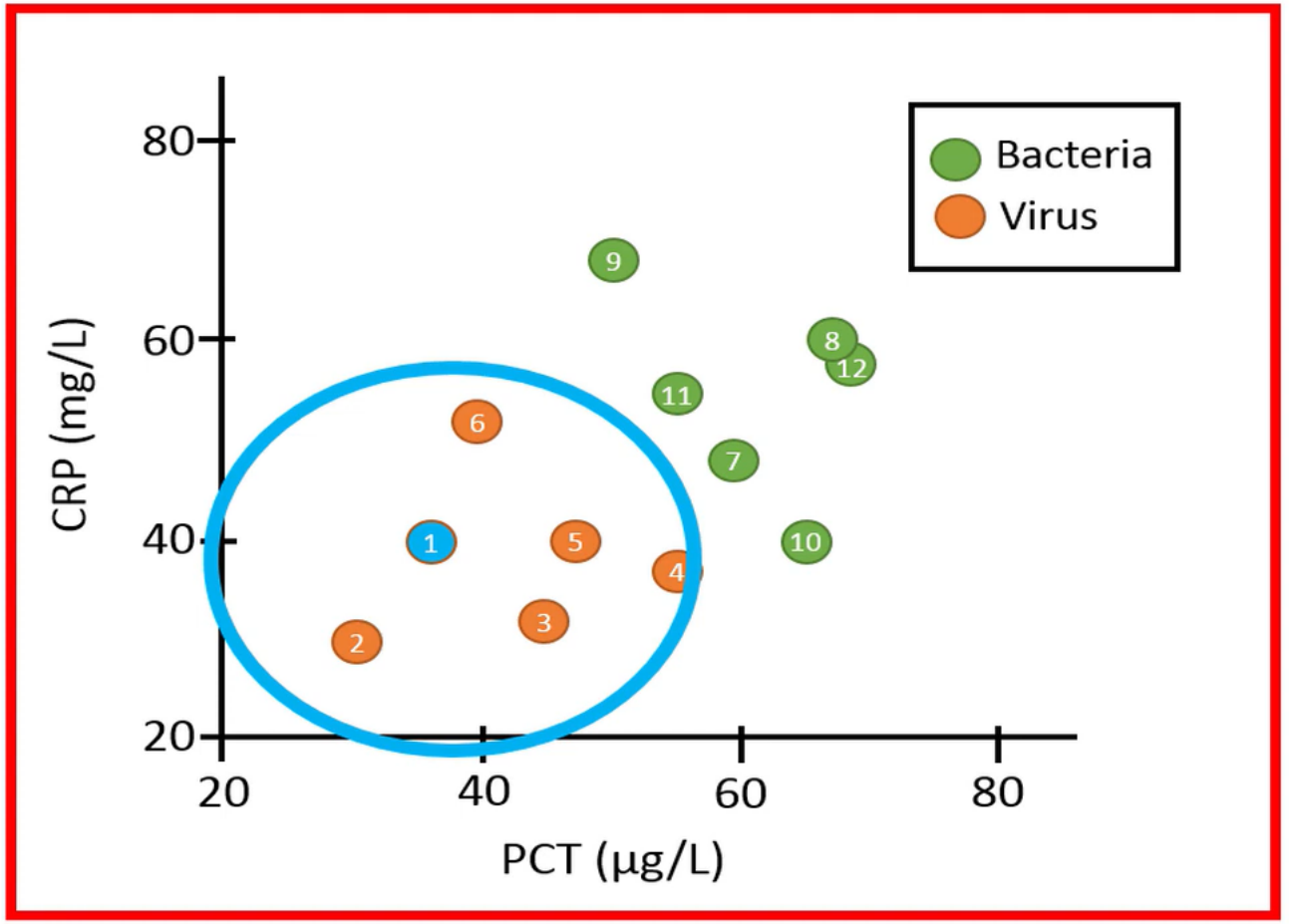
Infection	CRP (mg/L)	PCT (µg/L)	Predict
Viral	40	36	
Viral	30	30	
Viral	32	45	
Viral	37	55	
Viral	40	47	
Viral	52	40	
Bacterial	48	59	
Bacterial	60	67	
Bacterial	68	50	
Bacterial	40	65	
Bacterial	55	55	
Bacterial	58	68	



One way to evaluate how well KNN predicts the class of new cases is to use the leave-one-out cross-validation (LOOCV) method on the existing data with a known class.

How good is the classifier?

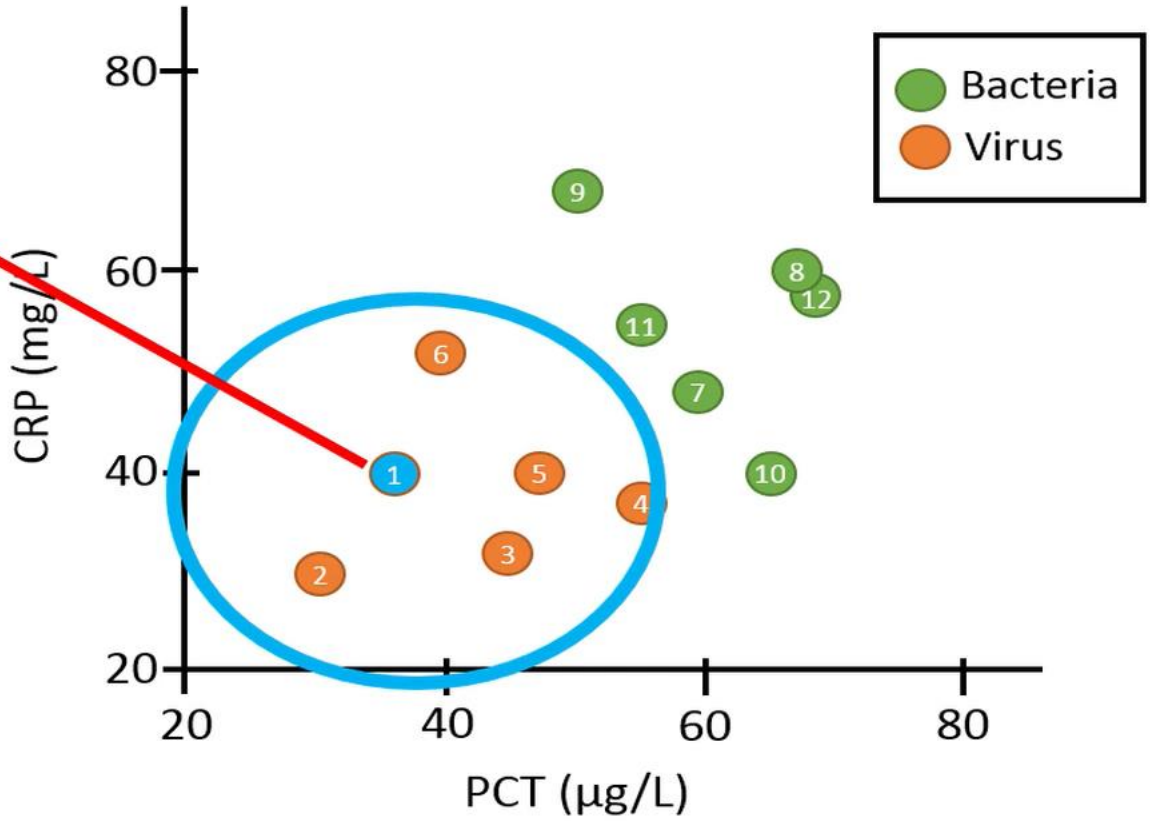
Infection	CRP (mg/L)	PCT (µg/L)	Predict
Viral	40	36	
Viral	30	30	
Viral	32	45	
Viral	37	55	
Viral	40	47	
Viral	52	40	
Bacterial	48	59	
Bacterial	60	67	
Bacterial	68	50	
Bacterial	40	65	
Bacterial	55	55	
Bacterial	58	68	



If we leave out the first patient from the training data where we pretend that we do not know that this patient has a viral infection, we can let KNN predict this for us and then see if the prediction is correct or not.

How good is the classifier?

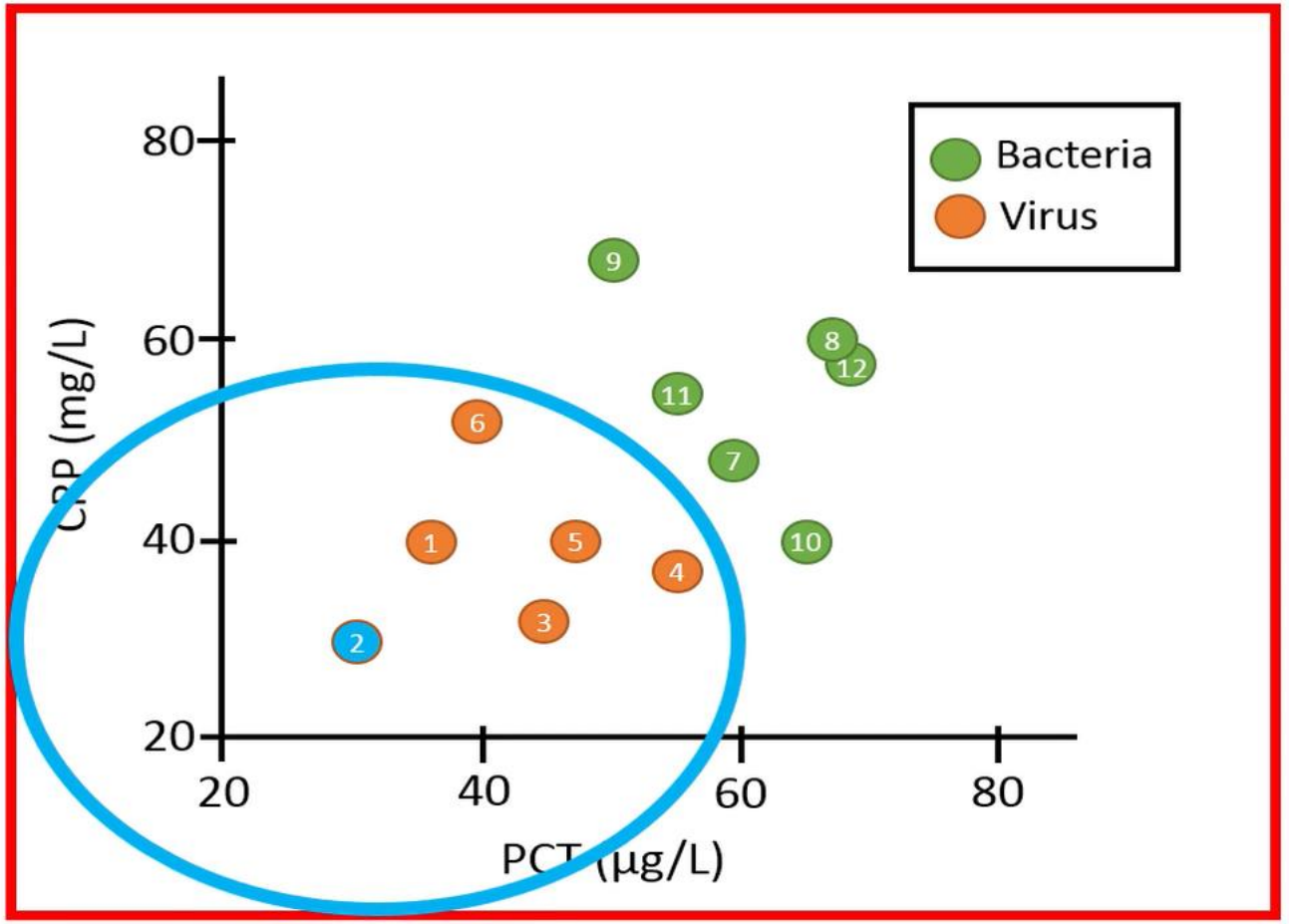
Infection	CRP (mg/L)	PCT (µg/L)	Predict
Viral	40	36	Viral
Viral	30	30	
Viral	32	45	
Viral	37	55	
Viral	40	47	
Viral	52	40	
Bacterial	48	59	
Bacterial	60	67	
Bacterial	68	50	
Bacterial	40	65	
Bacterial	55	55	
Bacterial	58	68	



Since the five closest neighbors around the data point number 1 are of class “virus”, the predicted class of this observation is “virus”.

How good is the classifier?

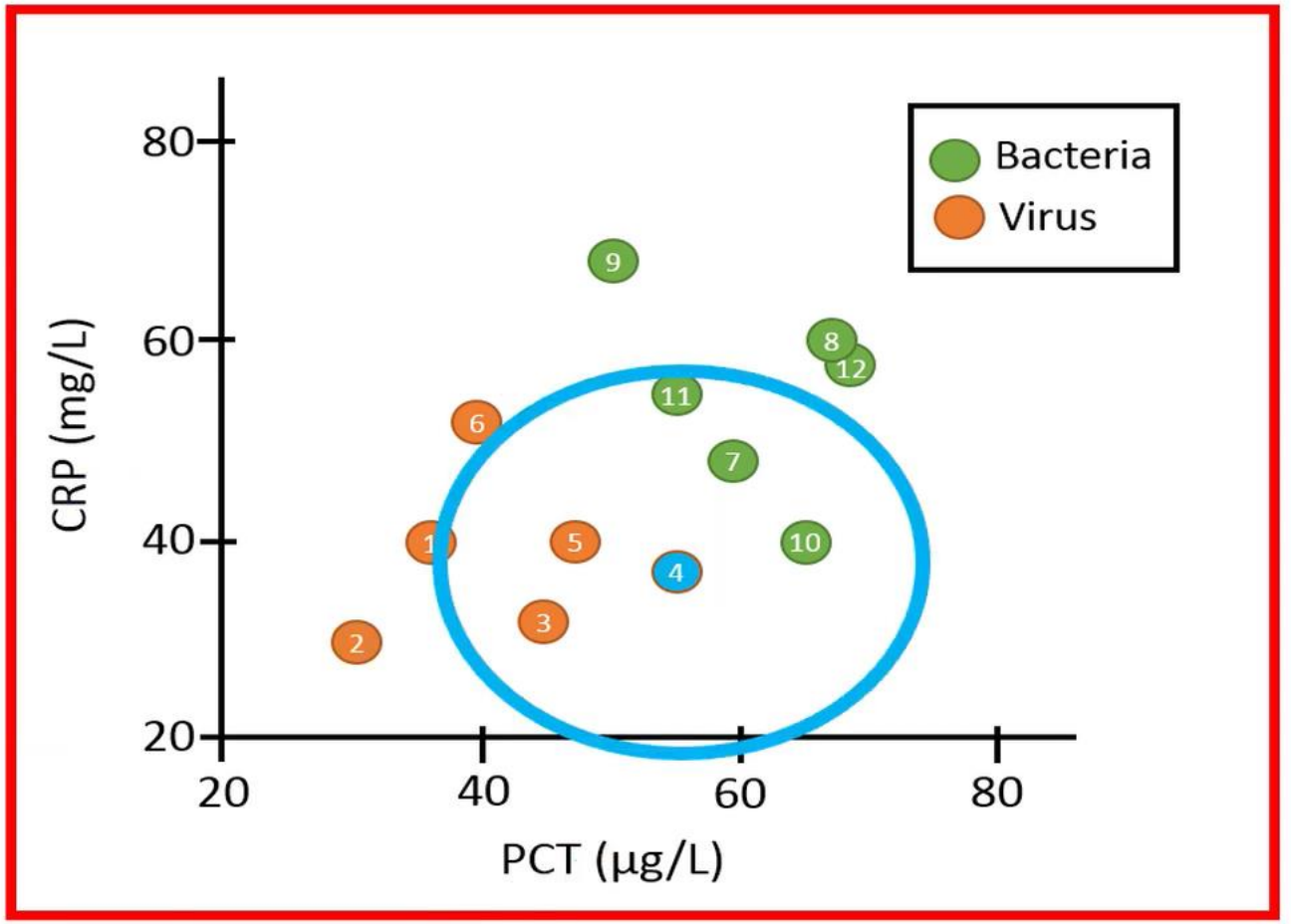
Infection	CRP (mg/L)	PCT (µg/L)	Predict
Viral	40	36	Viral
Viral	30	30	Viral
Viral	32	45	
Viral	37	55	
Viral	40	47	
Viral	52	40	
Bacterial	48	59	
Bacterial	60	67	
Bacterial	68	50	
Bacterial	40	65	
Bacterial	55	55	
Bacterial	58	68	



Similarly, the second patient is also correctly predicted to have a viral infection because all the five closest neighbors around this point are of class “virus”.

How good is the classifier?

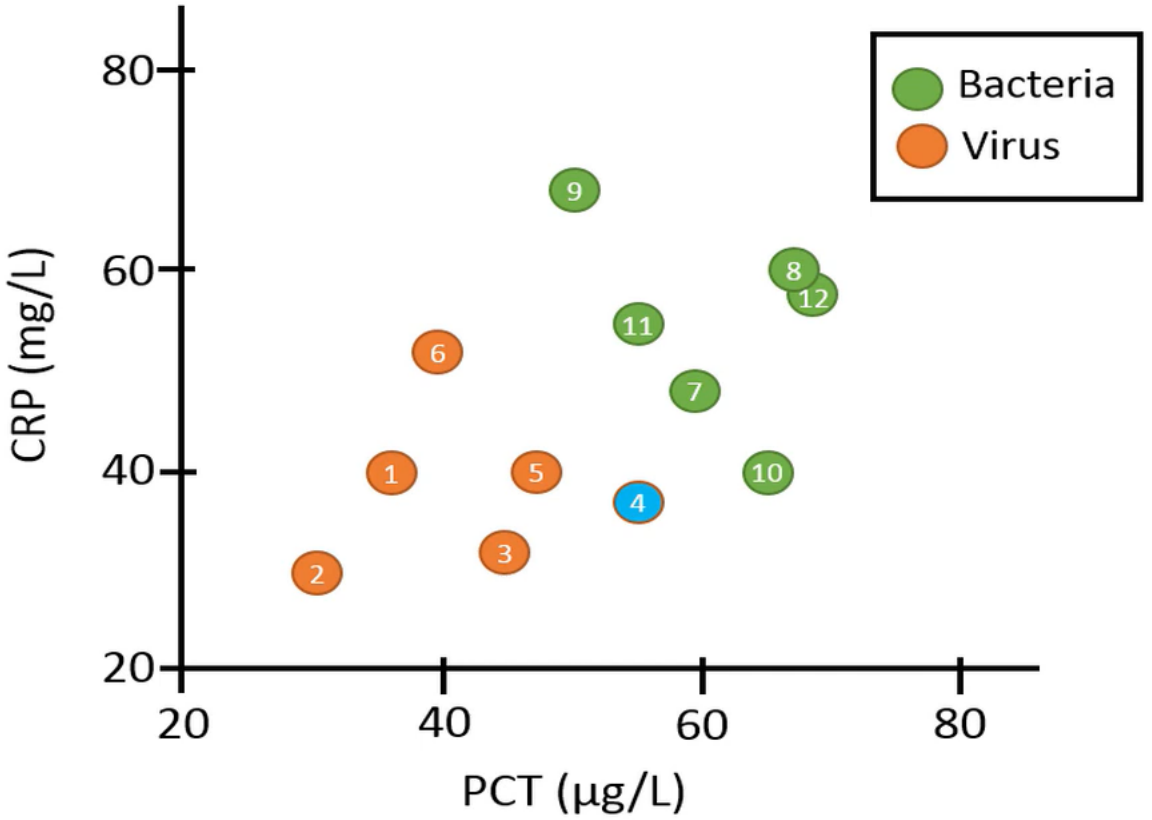
Infection	CRP (mg/L)	PCT (µg/L)	Predict
Viral	40	36	Viral
Viral	30	30	Viral
Viral	32	45	Viral
Viral	37	55	
Viral	40	47	
Viral	52	40	
Bacterial	48	59	
Bacterial	60	67	
Bacterial	68	50	
Bacterial	40	65	
Bacterial	55	55	
Bacterial	58	68	



However, when we predict the class of the fourth person, which we know had a viral infection, KNN predicts that the person has a bacterial infection because three out of the five closest neighbors are of class bacteria.

How good is the classifier?

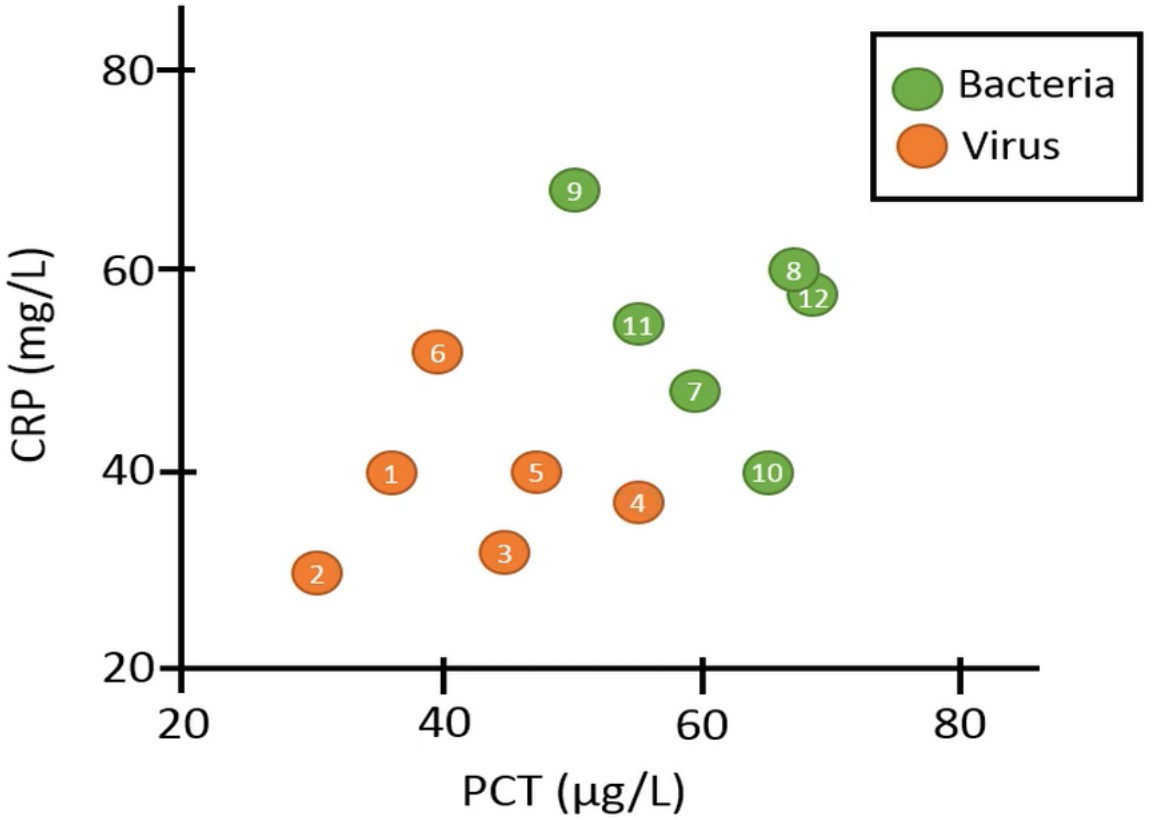
Infection	CRP (mg/L)	PCT (µg/L)	Predict
Viral	40	36	Viral
Viral	30	30	Viral
Viral	32	45	Viral
Viral	37	55	Bacterial
Viral	40	47	
Viral	52	40	
Bacterial	48	59	
Bacterial	60	67	
Bacterial	68	50	
Bacterial	40	65	
Bacterial	55	55	
Bacterial	58	68	



Since we know that person number 4 had a viral infection, we know that KNN has made the wrong prediction in this case.

How good is the classifier?

Infection	CRP (mg/L)	PCT (µg/L)	Predict
Viral	40	36	Viral
Viral	30	30	Viral
Viral	32	45	Viral
Viral	37	55	Bacterial
Viral	40	47	Viral
Viral	52	40	Bacterial
Bacterial	48	59	Bacterial
Bacterial	60	67	Bacterial
Bacterial	68	50	Bacterial
Bacterial	40	65	Bacterial
Bacterial	55	55	Bacterial
Bacterial	58	68	Bacterial

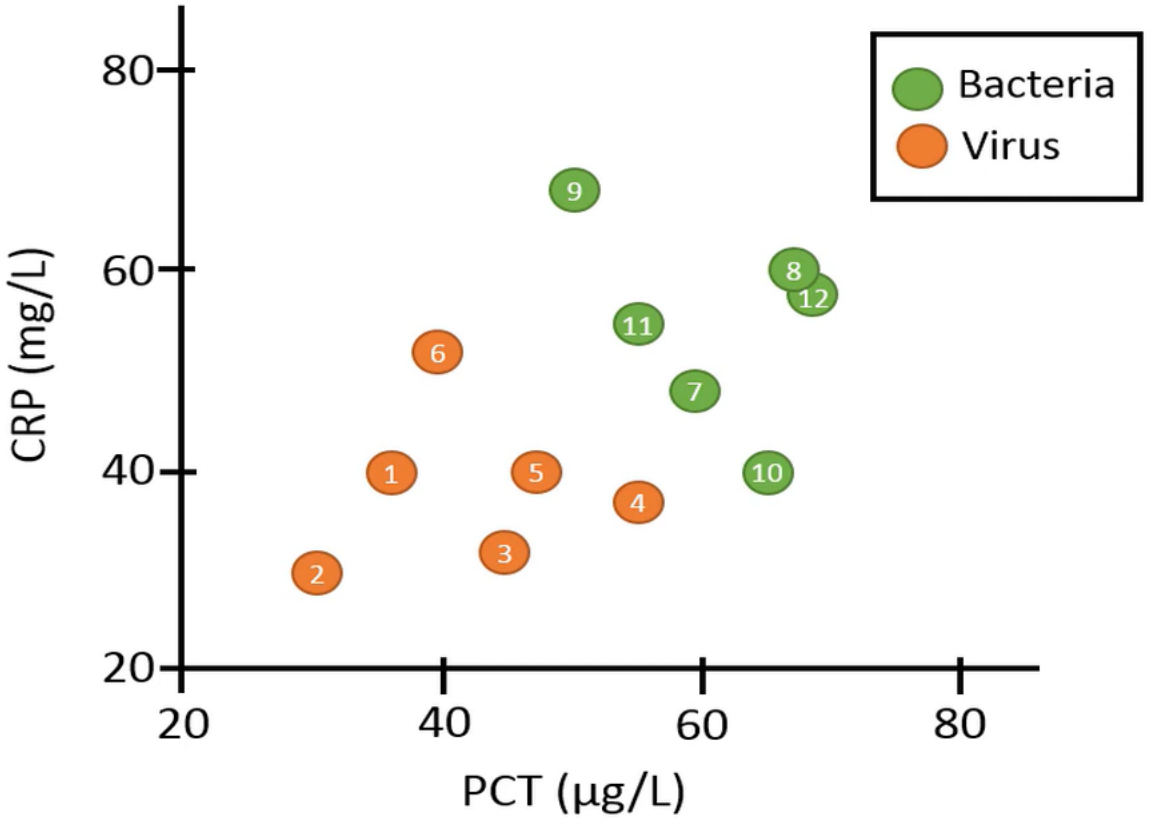


Based on the LOOCV method, we see that we make 10 correct predictions out of 12 possible.

How good is the classifier?

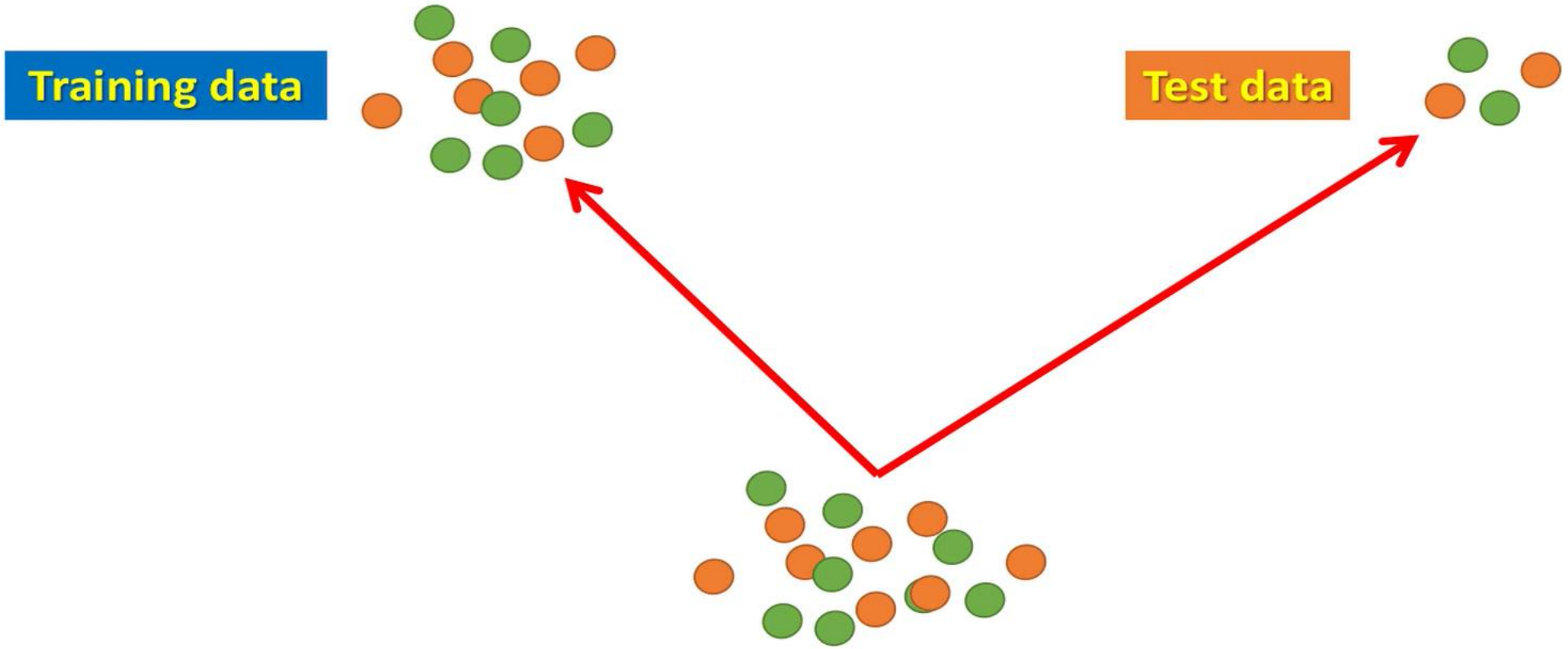
Infection	CRP (mg/L)	PCT (µg/L)	Predict
Viral	40	36	Viral
Viral	30	30	Viral
Viral	32	45	Viral
Viral	37	55	Bacterial
Viral	40	47	Viral
Viral	52	40	Bacterial
Bacterial	48	59	Bacterial
Bacterial	60	67	Bacterial
Bacterial	68	50	Bacterial
Bacterial	40	65	Bacterial
Bacterial	55	55	Bacterial
Bacterial	58	68	Bacterial

$$Accuracy = \frac{10}{12} = 0.83$$



This gives an accuracy of about 83%.

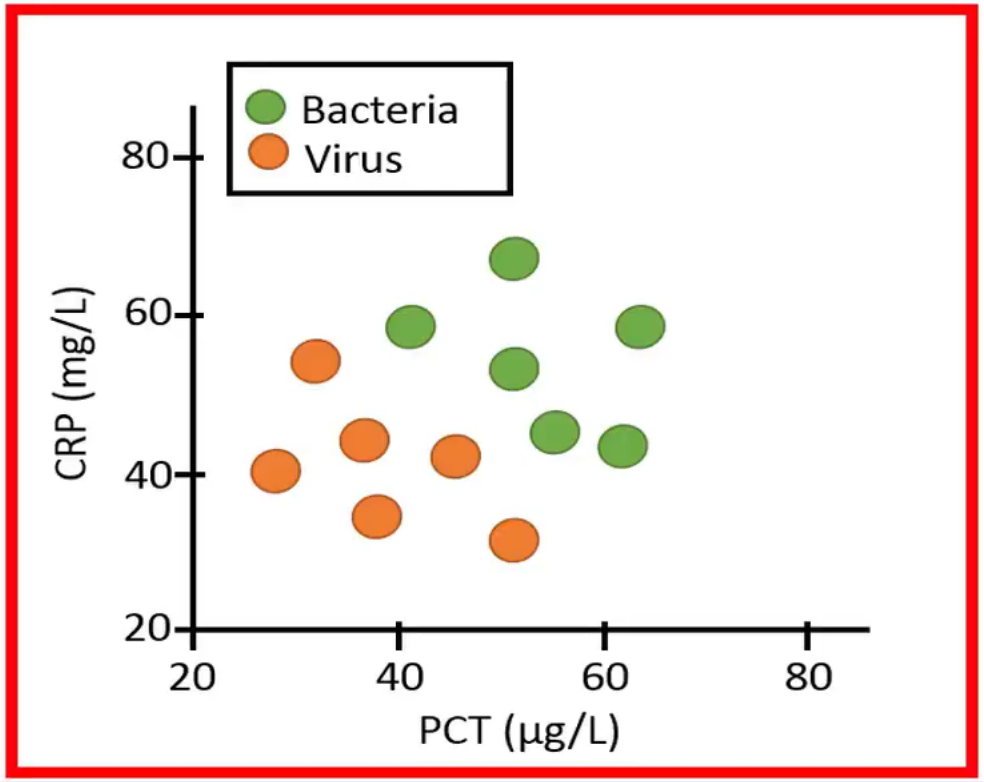
How good is the classifier?



If we have a very big data set, the LOOCV might take a long time to run. If we have plenty of data, we can instead use the hold-out method where we split the data set into a large training data set and a smaller test data set.

How good is the classifier?

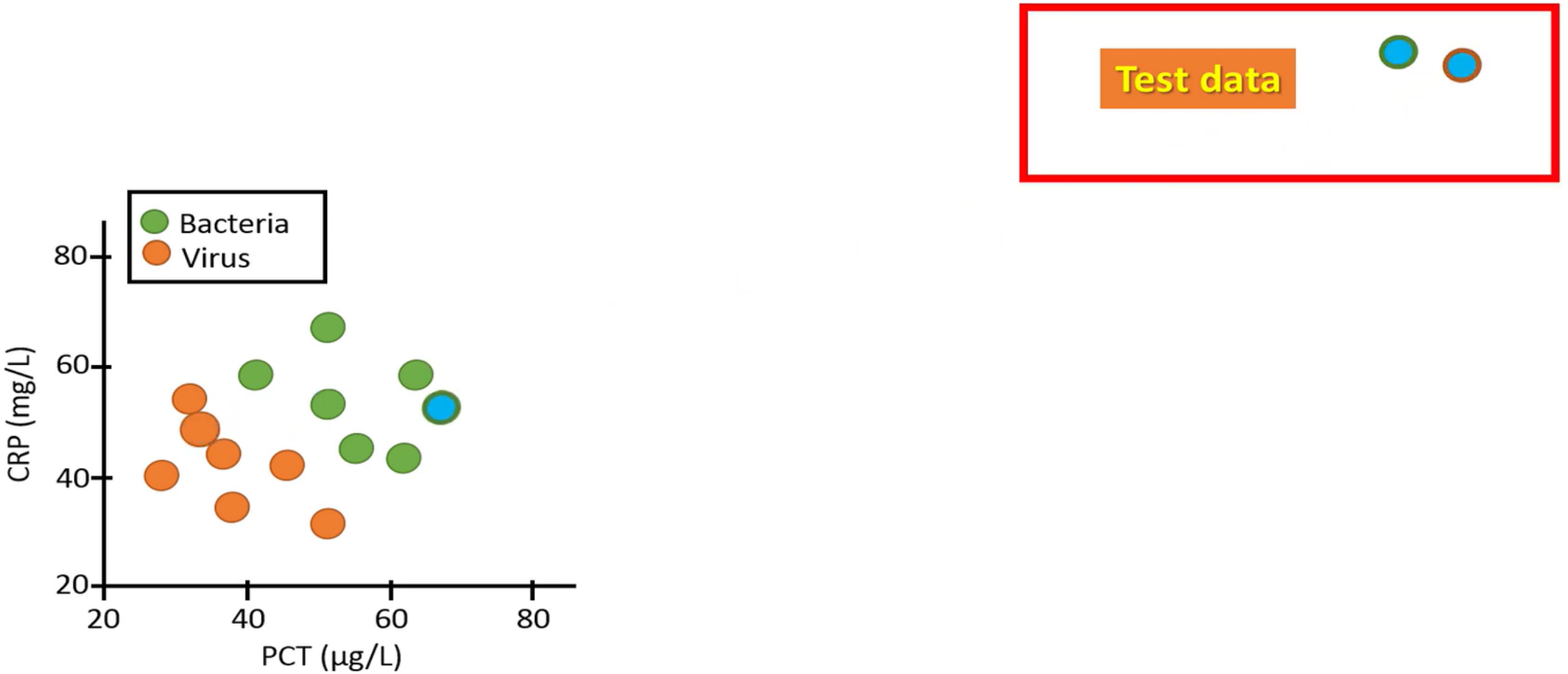
Training data



Test data

We then put the training data like this,

How good is the classifier?

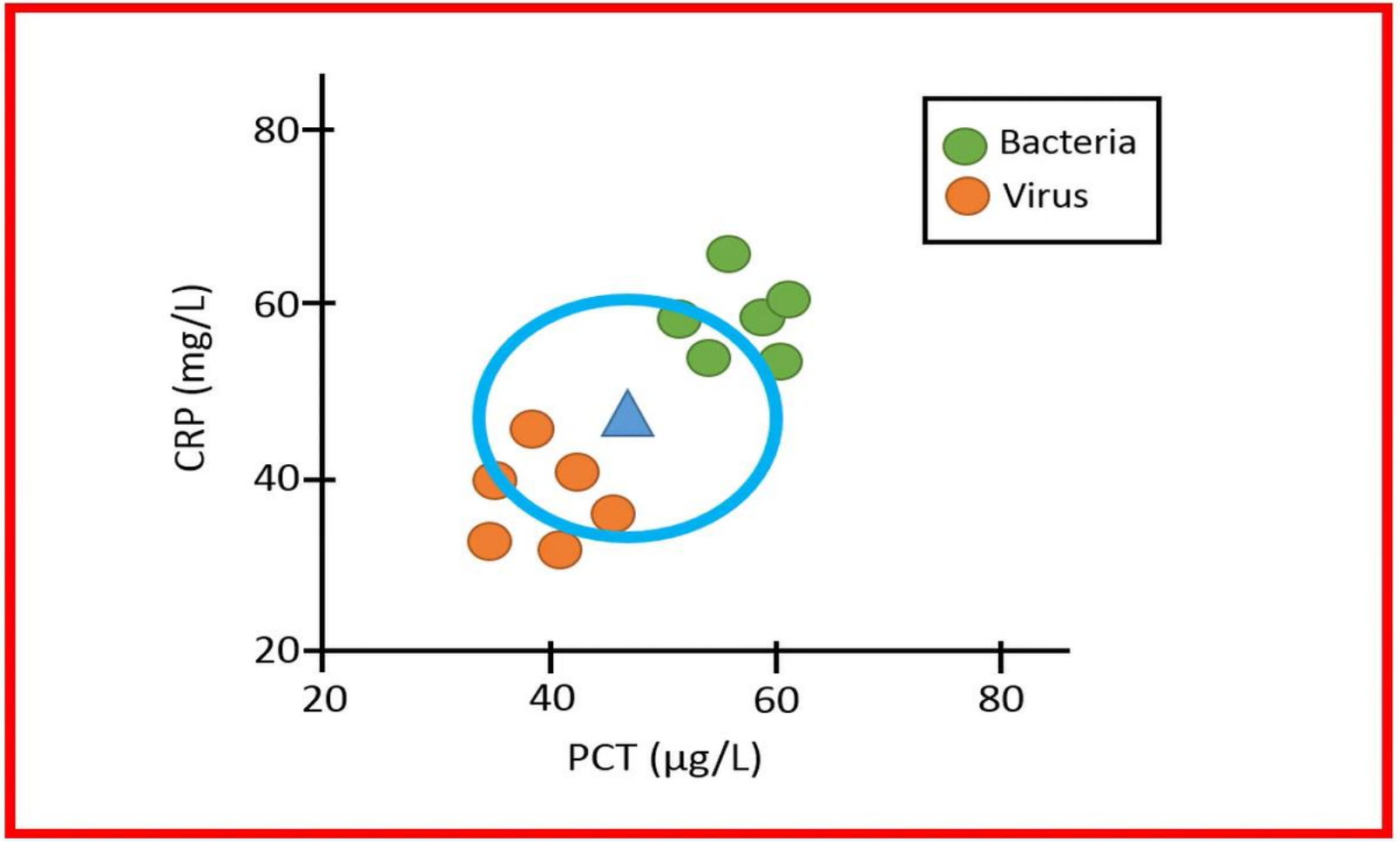


and predict the class of the test data based on the KNN.

KNN

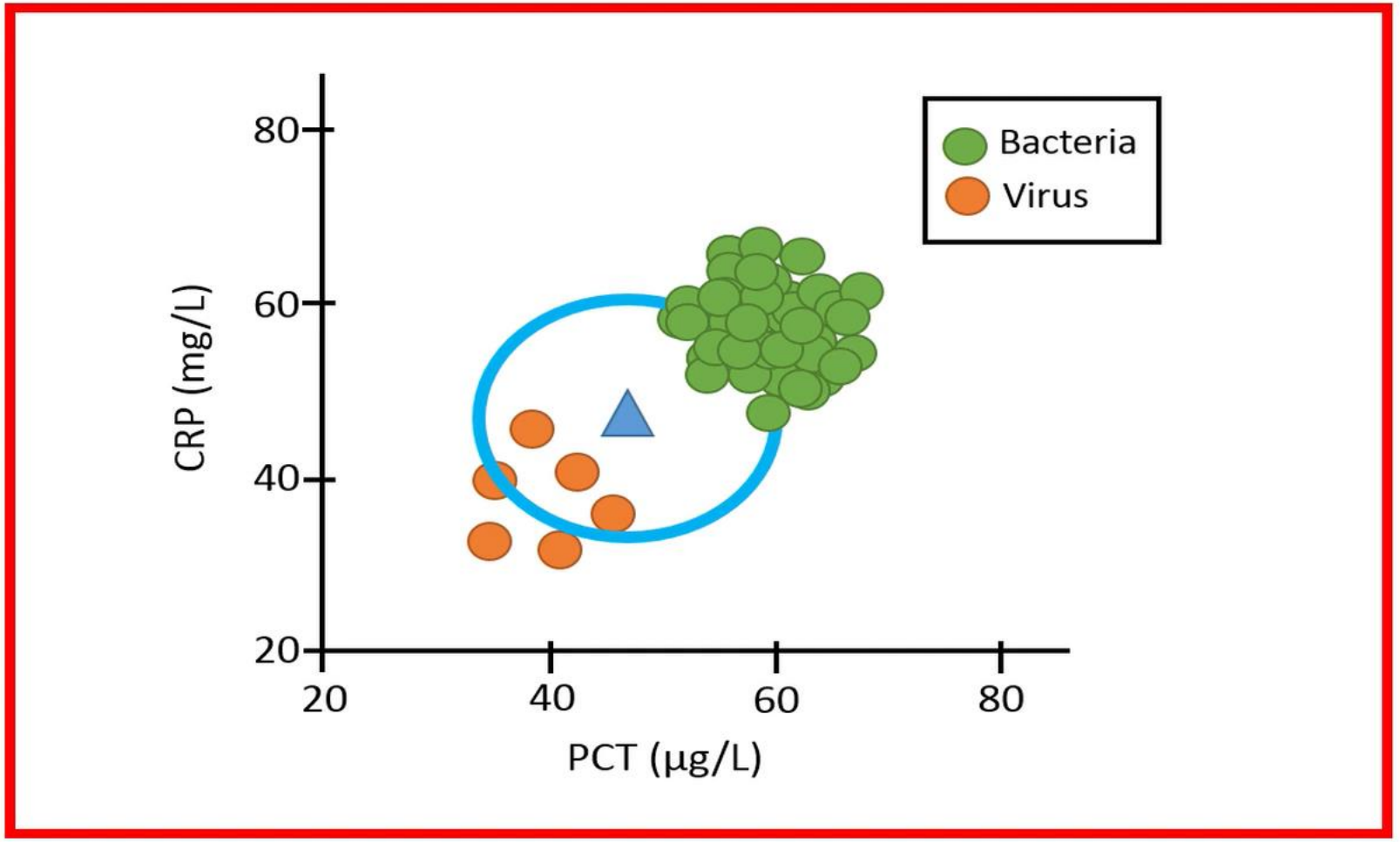
Problem with imbalanced data sets

Problem with imbalanced data sets



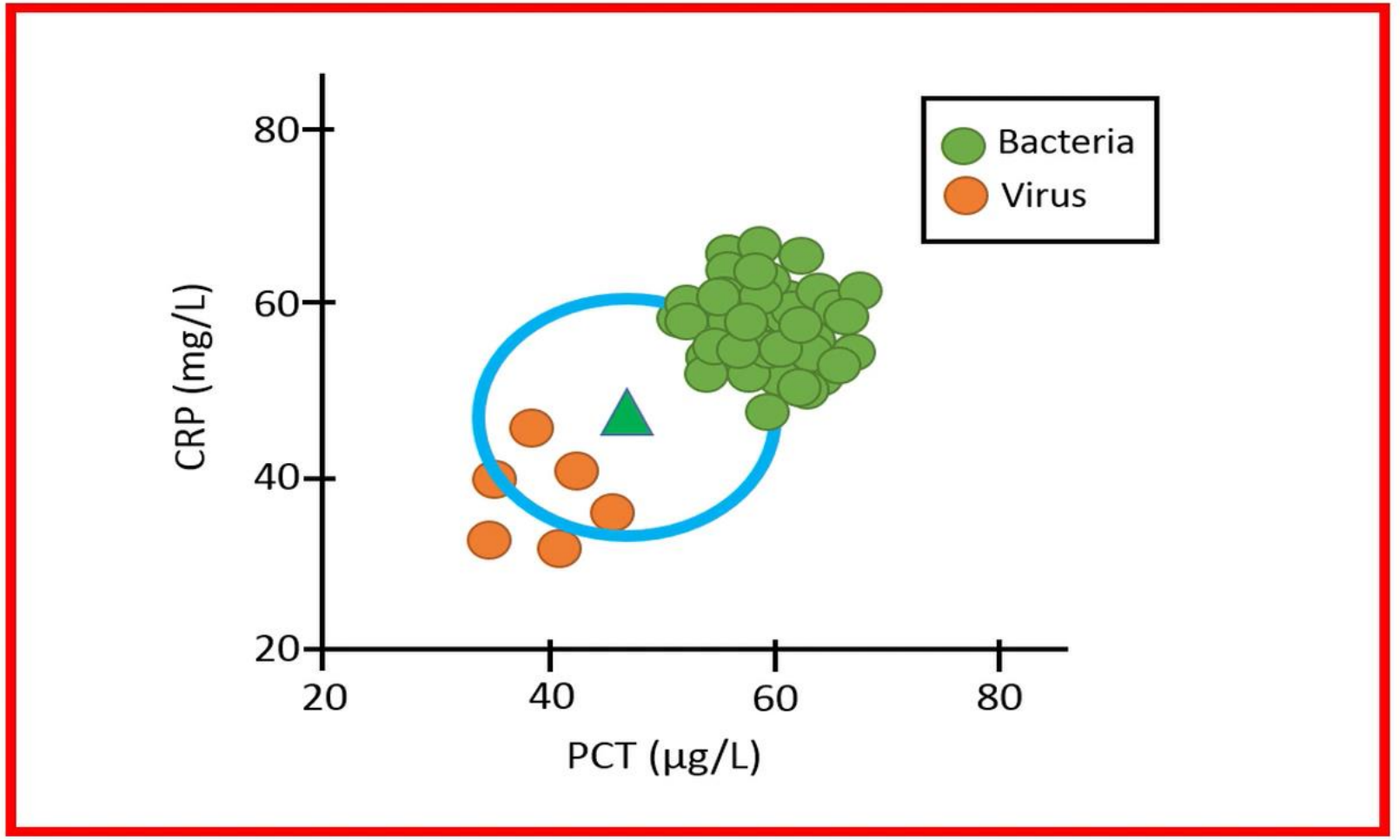
When the groups are of equal size, KNN is unbiased. For example, if *k* is set to five, we would predict the unknown case as having a viral infection since three out of the five closest neighbors are of class “virus”.

Problem with imbalanced data sets



However, if the bacteria group includes more data points than the virus group, the classifier will favor the bacteria group because of its higher density of data points in space.

Problem with imbalanced data sets

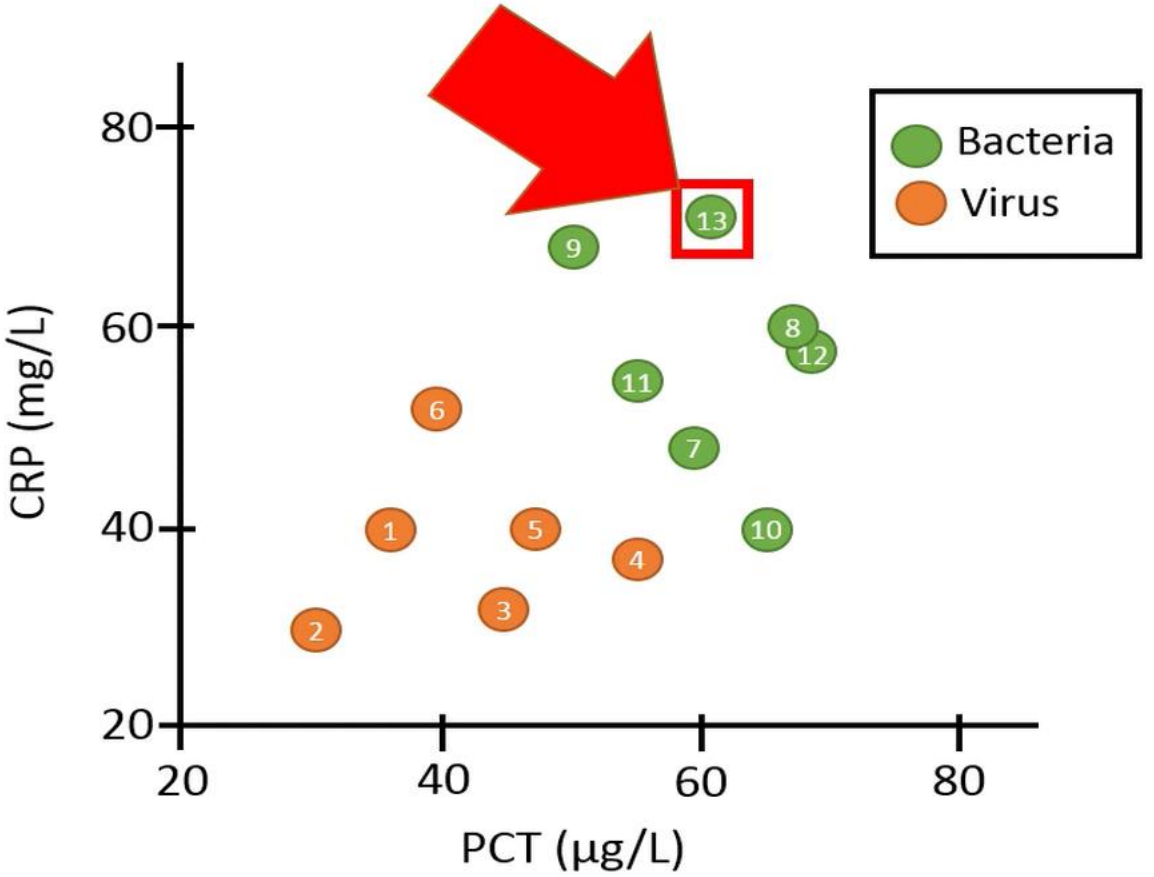


There are many types of methods to deal with an imbalanced data set. For example, one can put more weight on the group with fewer data points when calculating the distances between the unknown observation and the data points.

KNN

How do we find an optimal value for *k*?

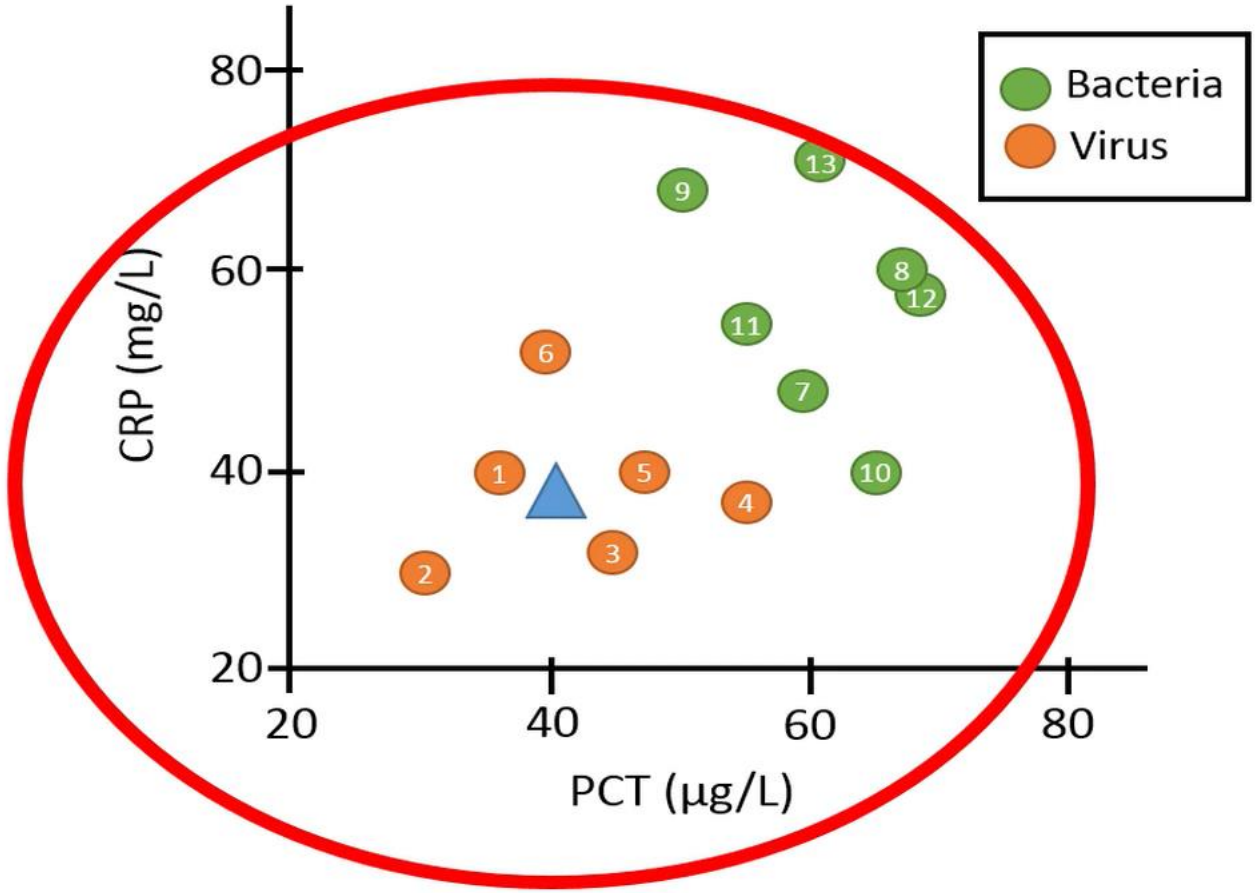
How do we find an optimal value for *k*?



Let's say that we have an additional data point here, which means that the "bacteria" group now includes seven observations whereas the "virus" group only includes six.

How do we find an optimal value for *k*?

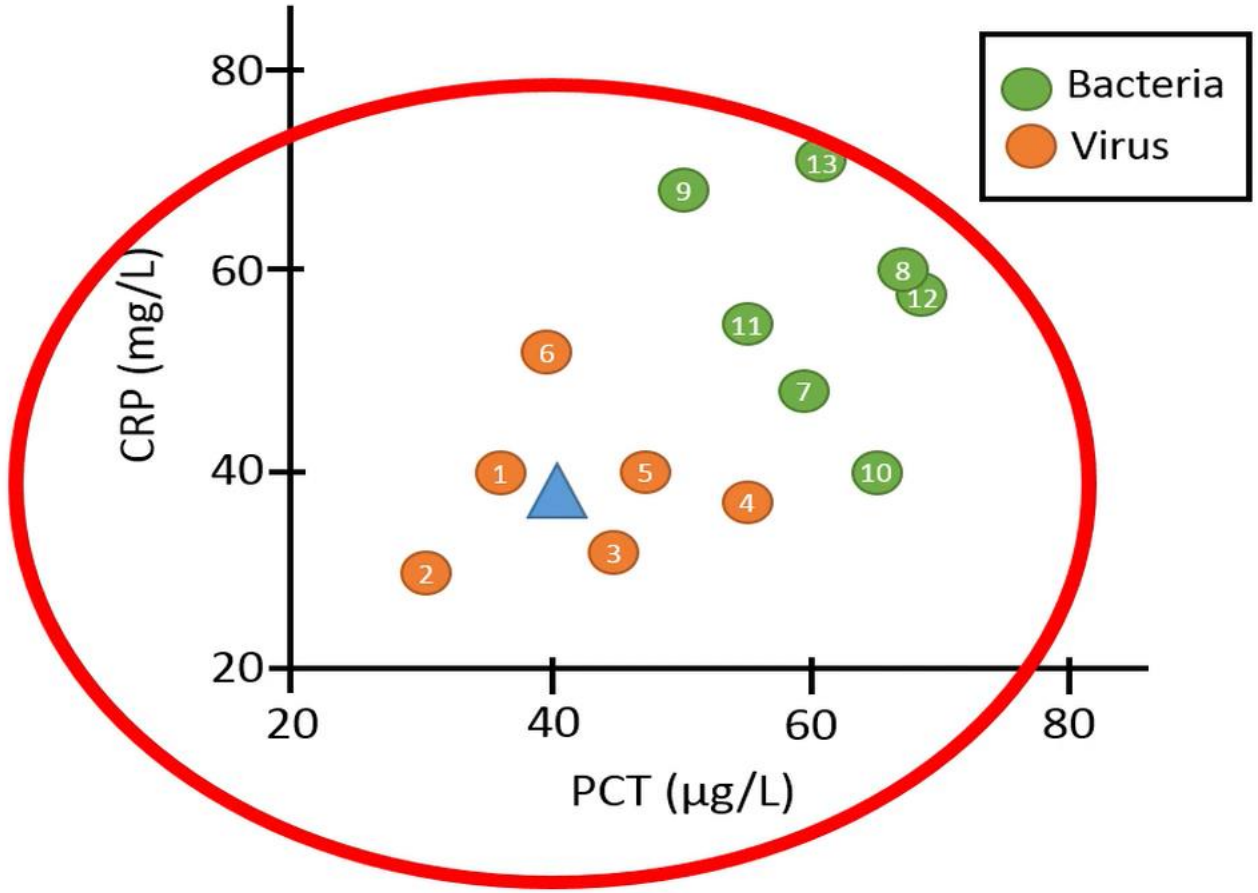
The value of *k* should not be too high



If we then would set *k* to 13, and predict the class of the new observation, the class will always be predicted as bacteria since we have in total 13 data points where the majority class will always be of type “bacteria”.

How do we find an optimal value for *k*?

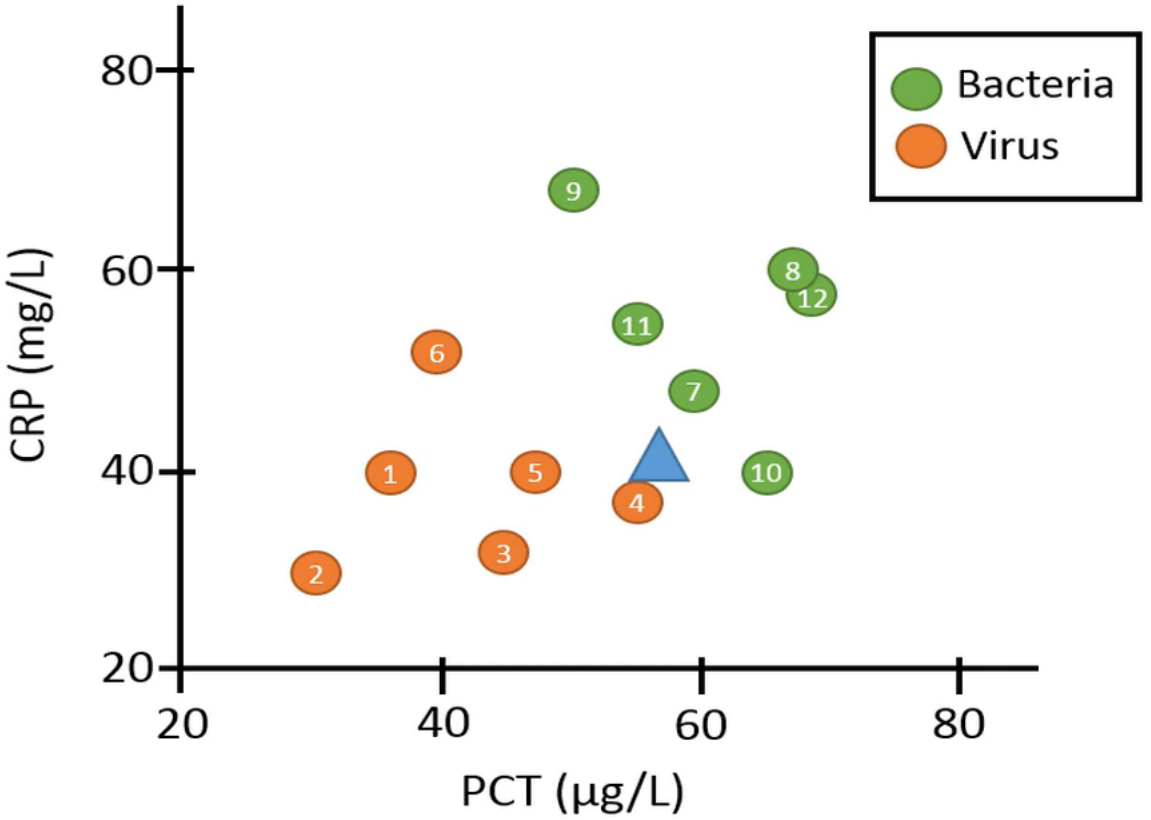
The value of *k* should not be too high



A high value of *k* is especially a problem with imbalanced data sets, where one group includes more observations than the other group.

How do we find an optimal value for *k*?

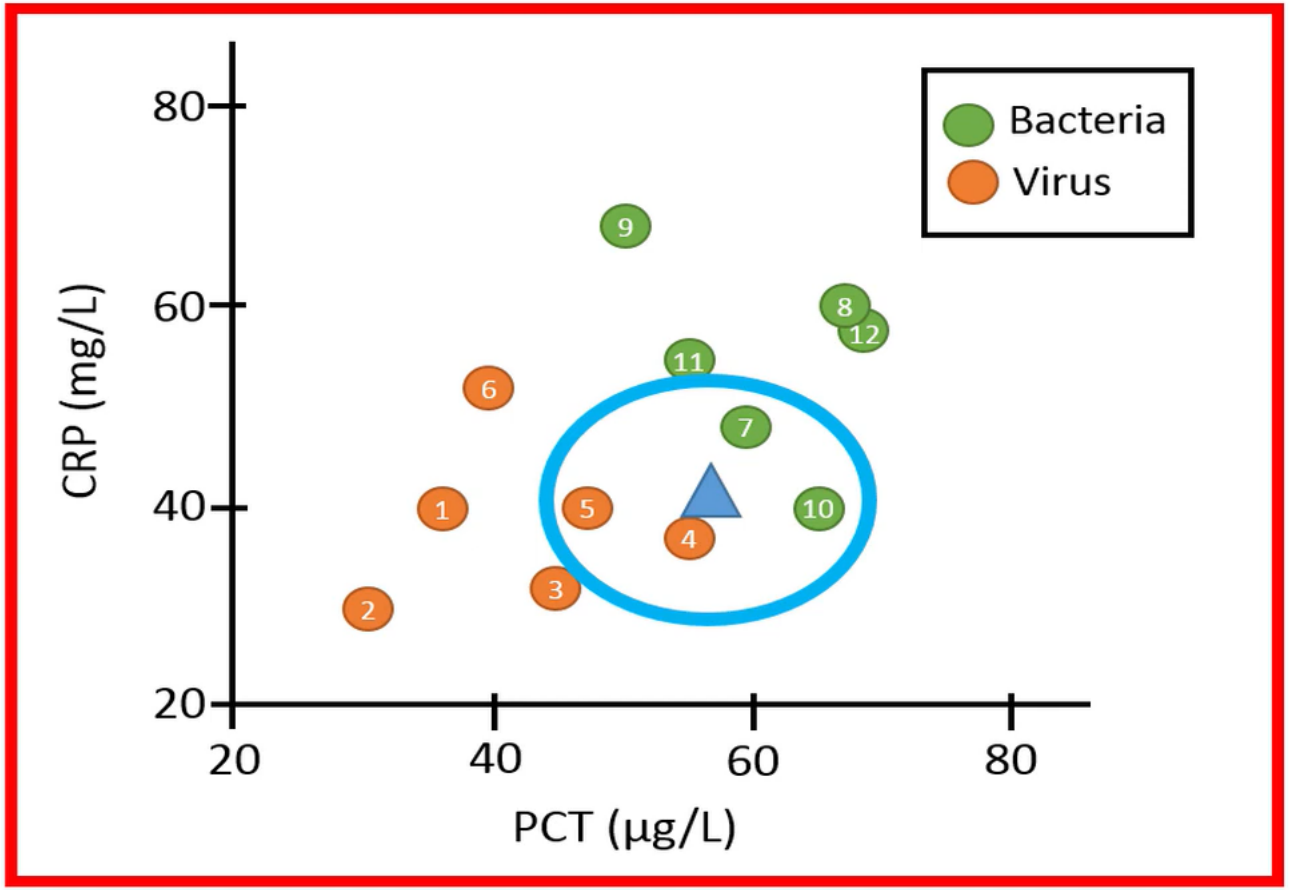
The value of *k* should be an odd number



If we like to predict the class based on two groups, it is recommended that *k* is an odd number since we then avoid possible ties.

How do we find an optimal value for *k*?

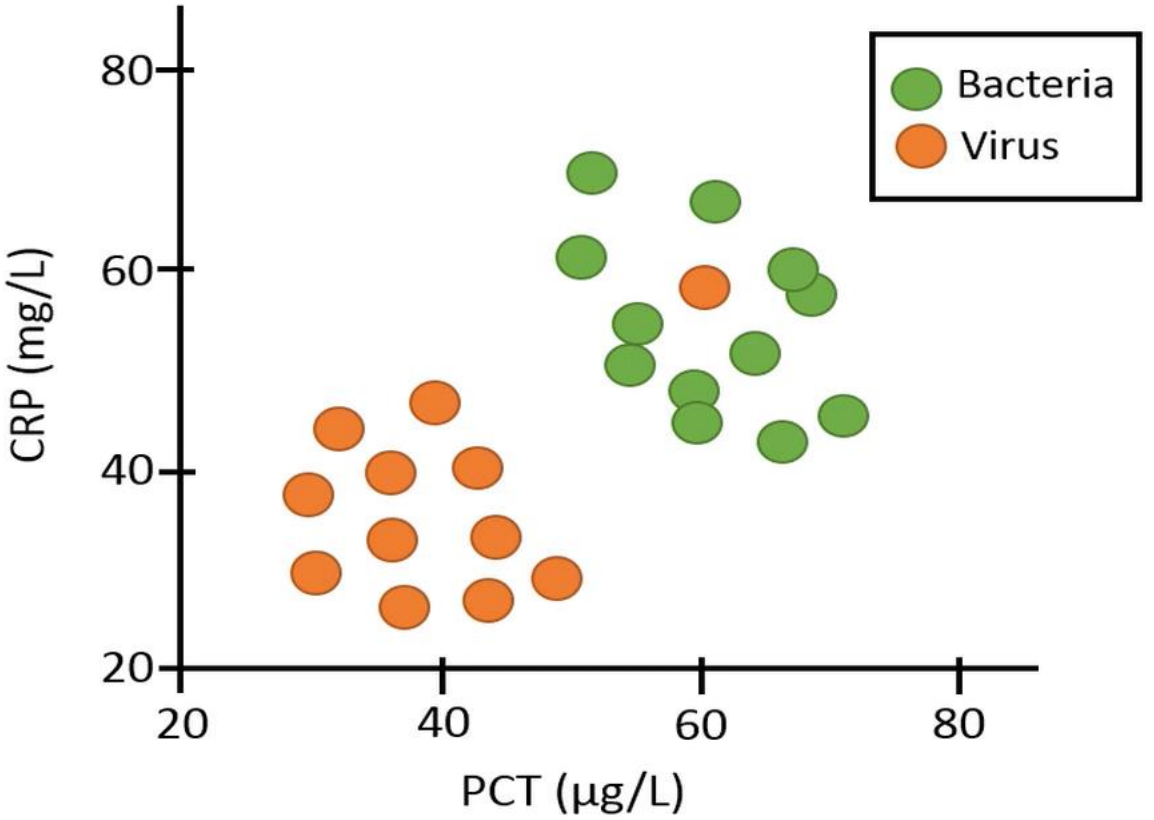
The value of *k* should be an odd number



For example, if we would set *k* to four, this means that we should check the four closest neighbors.

How do we find an optimal value for *k*?

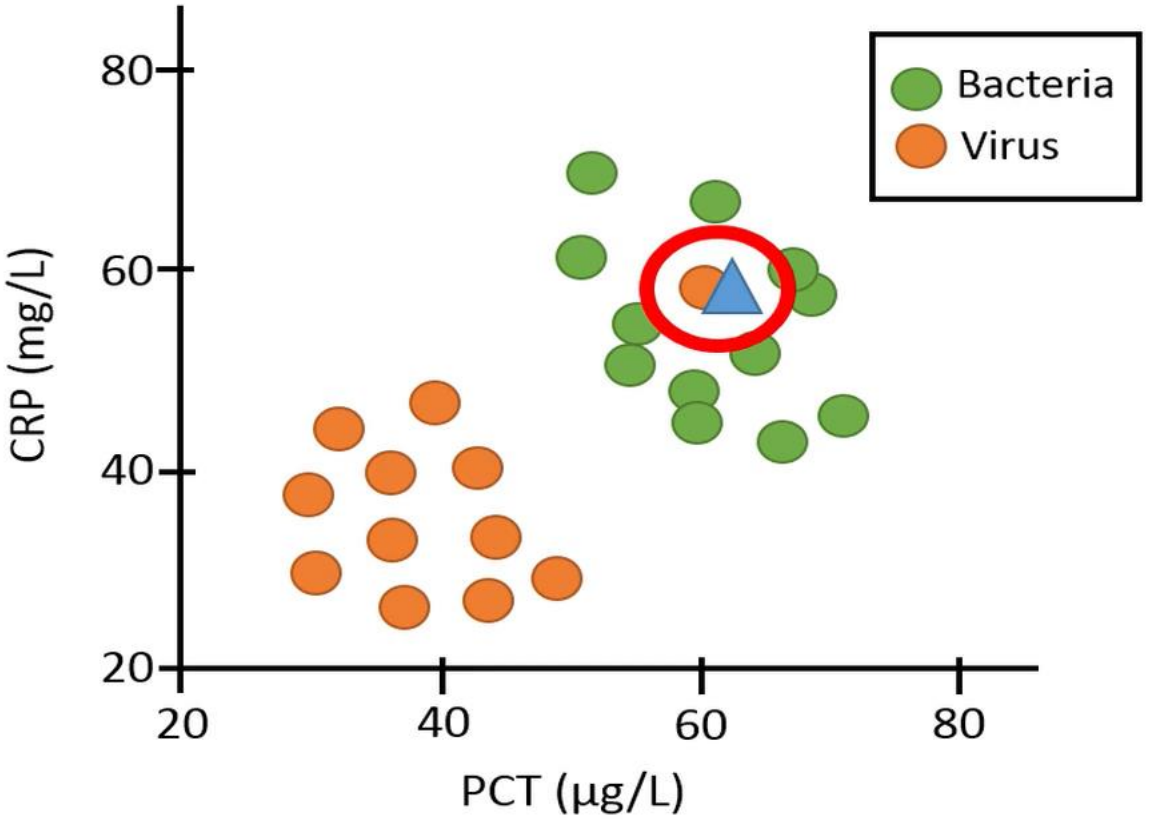
The value of *k* should not be equal to one



Also, if the value of *k* is set to one, the classification will be very sensitive to extreme values.

How do we find an optimal value for *k*?

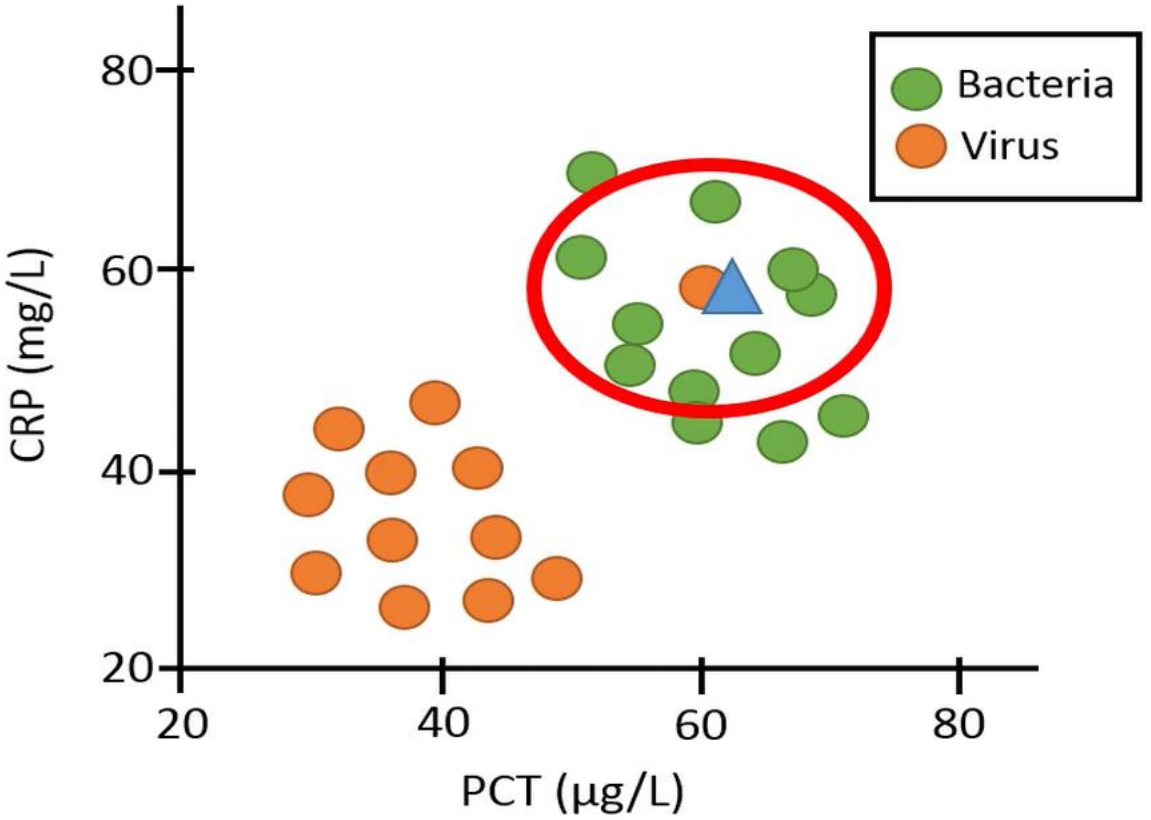
The value of *k* should not be equal to one



If the new observation that we like to classify happens to be close to such an outlier, it will be predicted as the same class as the outlier.

How do we find an optimal value for *k*?

The value of *k* should not be equal to one



If we increase *k* from, for example, 1 to 9, we see that 8 out of the nine closest neighbors are of class “bacteria”.

How do we find an optimal value for *k*?

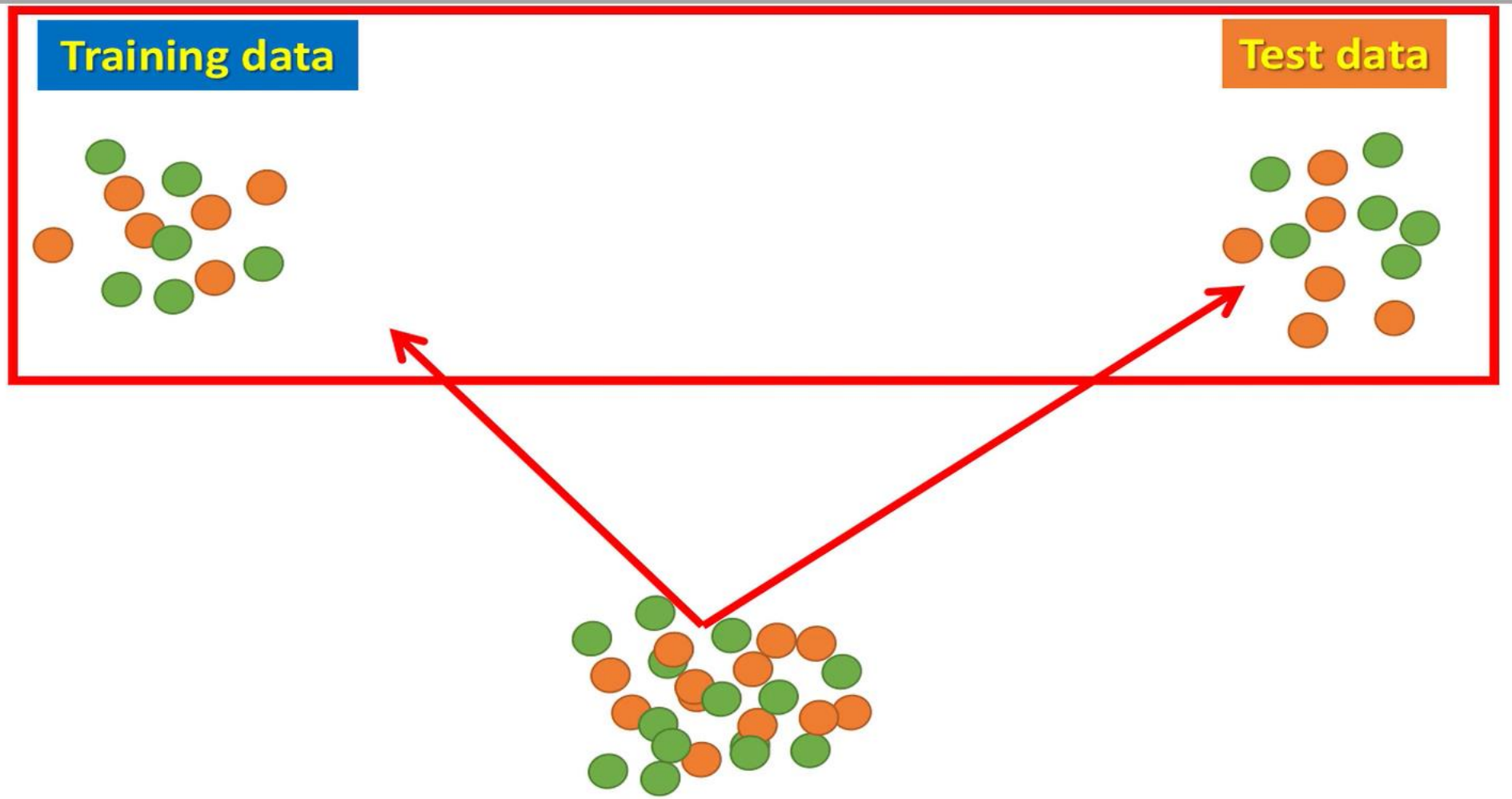
The value of *k* should therefore not be too small or too large!

How do we find an optimal value for *k*?

A rule of thumb is to set *k* equal to:

$$k = \sqrt{n}$$

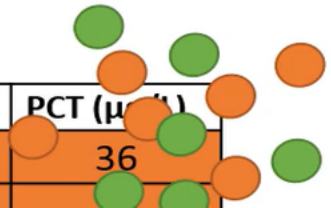
How do we find an optimal value for *k*?



We can also train the KNN to find an optimal value of *k* and at the same time perform validation of the model. To do this, we could, for example, use the hold-out method where we split the data into a training data set and a test data set.

How do we find an optimal value for *k*?

Training data



Infection	CRP (mg/L)	PCT (μ
Viral	40	36
Viral	30	[illegible]
Viral	32	45
Viral	37	55
Viral	40	47
Viral	52	40
Bacterial	48	59
Bacterial	60	67
Bacterial	68	50
Bacterial	40	65
Bacterial	55	55
Bacterial	58	68

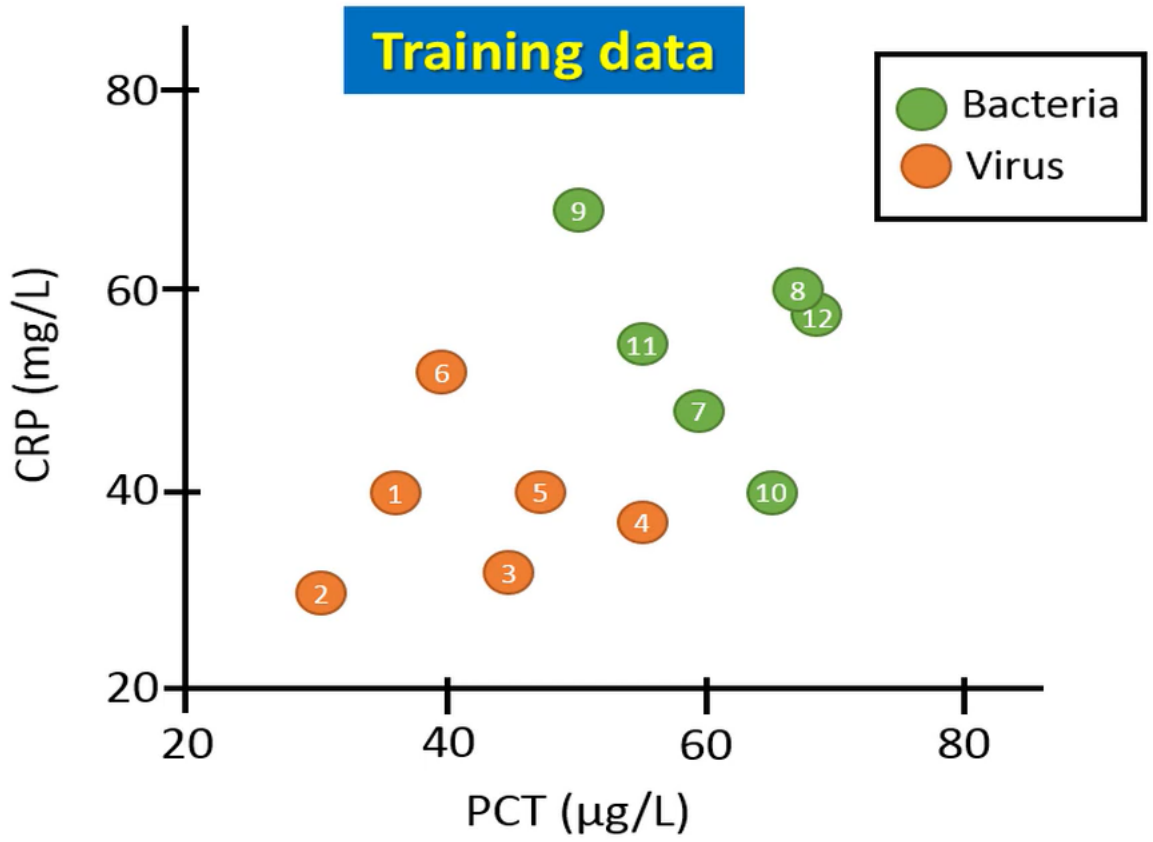
Test data

Let's say that our previous example data represents our training data set in this case.

k=5

Infection	CRP (mg/L)	PCT (µg/L)	Predict
Viral	40	36	Viral
Viral	30	30	Viral
Viral	32	45	Viral
Viral	37	55	Bacterial
Viral	40	47	Viral
Viral	52	40	Bacterial
Bacterial	48	59	Bacterial
Bacterial	60	67	Bacterial
Bacterial	68	50	Bacterial
Bacterial	40	65	Bacterial
Bacterial	55	55	Bacterial
Bacterial	58	68	Bacterial

$$Accuracy = \frac{10}{12} = 0.83$$

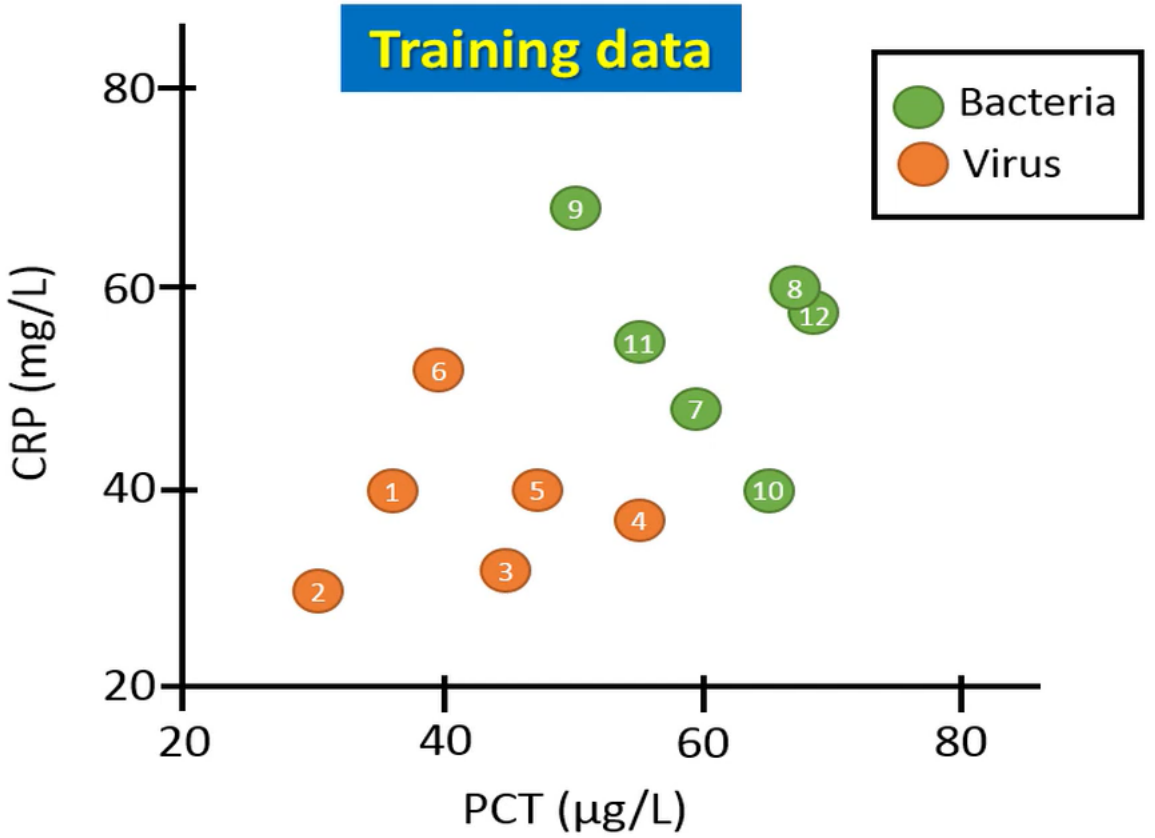


We know that the accuracy, or the proportion of correct predictions, is equal to about 83% when we set *k* to five for this data set.

k=3

Infection	CRP (mg/L)	PCT (µg/L)	Predict
Viral	40	36	Viral
Viral	30	30	Viral
Viral	32	45	Viral
Viral	37	55	Viral
Viral	40	47	Viral
Viral	52	40	Viral
Bacterial	48	59	Bacterial
Bacterial	60	67	Bacterial
Bacterial	68	50	Bacterial
Bacterial	40	65	Viral
Bacterial	55	55	Bacterial
Bacterial	58	68	Bacterial

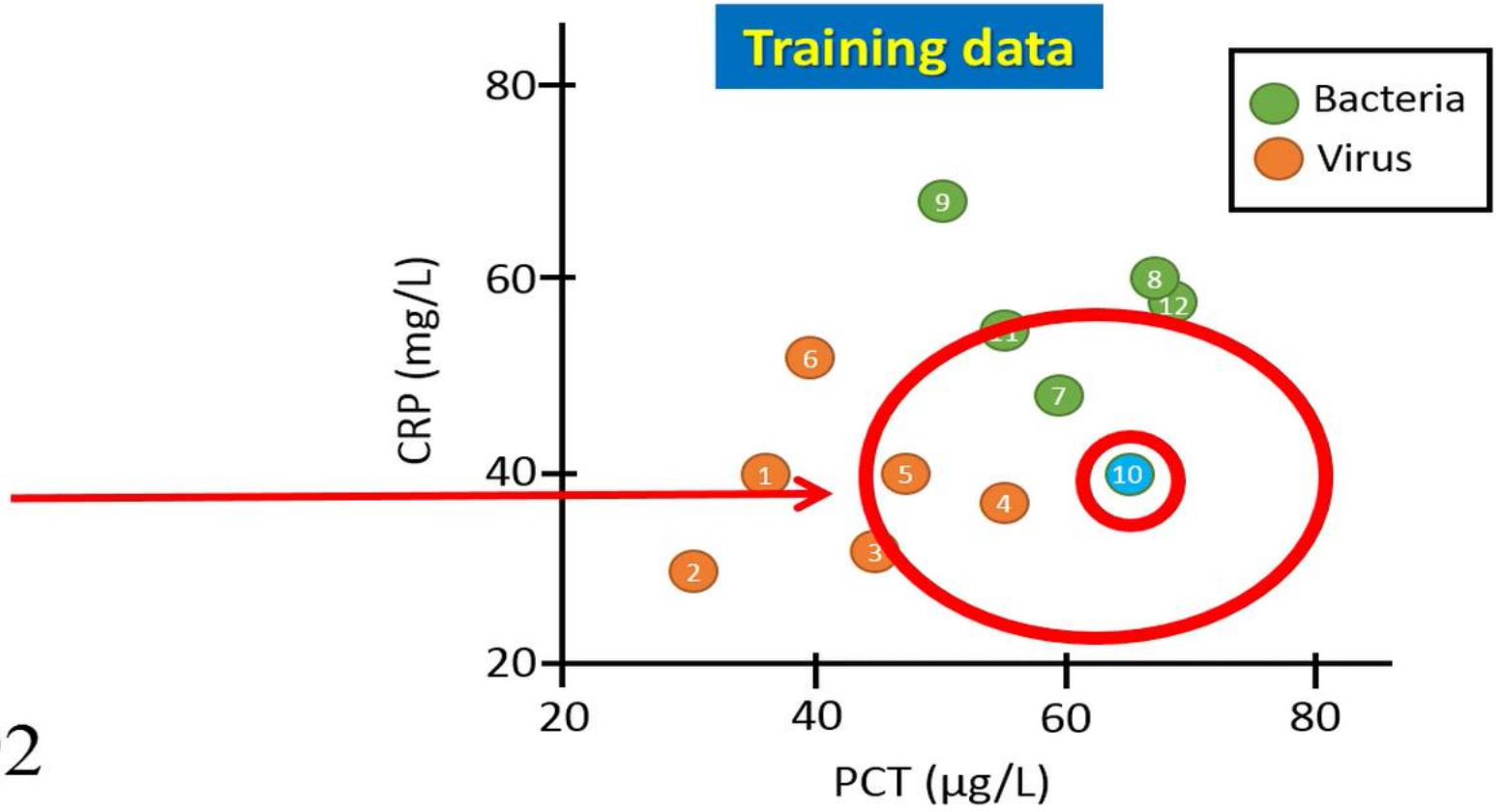
$$Accuracy = \frac{11}{12} = 0.92$$



If we set *k* to three, and compute the KNN, we see that we have increased the accuracy to about 92%.

k=3

Infection	CRP (mg/L)	PCT (µg/L)	Predict
Viral	40	36	Viral
Viral	30	30	Viral
Viral	32	45	Viral
Viral	37	55	Viral
Viral	40	47	Viral
Viral	52	40	Viral
Bacterial	48	59	Bacterial
Bacterial	60	67	Bacterial
Bacterial	68	50	Bacterial
Bacterial	40	65	Viral
Bacterial	55	55	Bacterial
Bacterial	58	68	Bacterial



$$Accuracy = \frac{11}{12} = 0.92$$

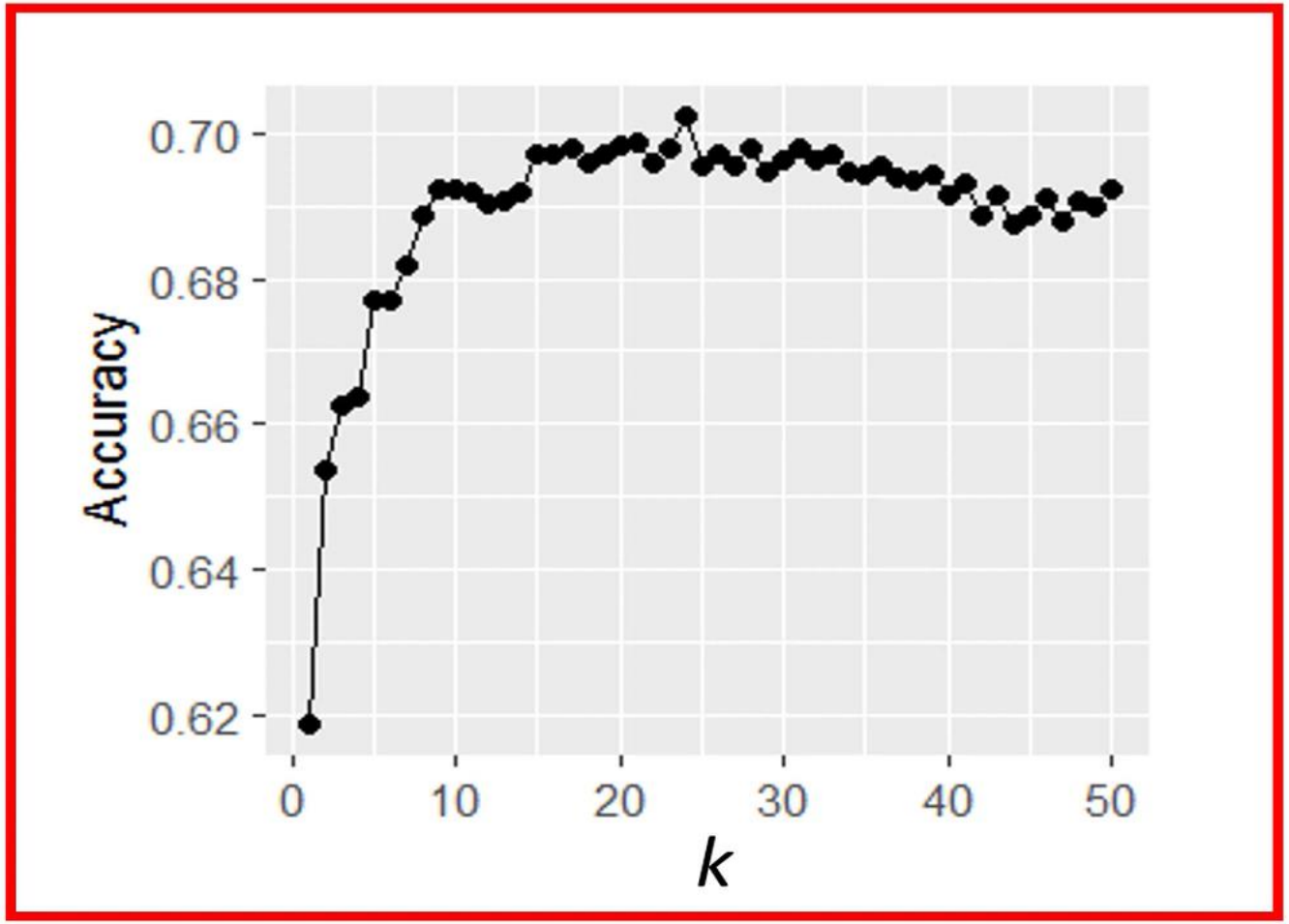
The only case that is incorrectly classified, if we set *k* to three, is patient number 10, where two of its three closest neighbors are of class “virus” and only one of class “bacteria”.

k

k	Accuracy
3	92%
5	83%
7	100%
9	83%

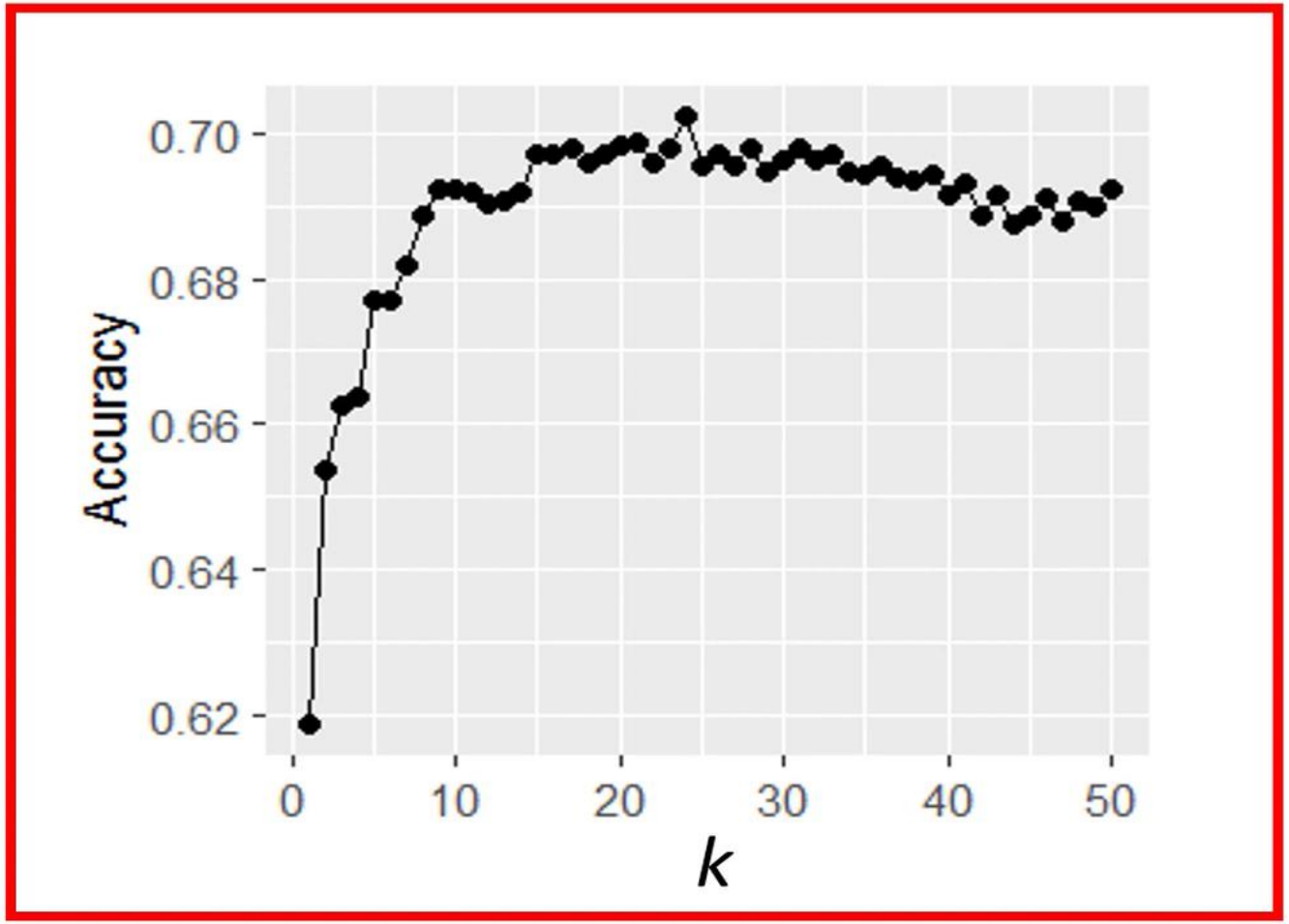
If we try four different odd numbers for *k*, we can select the value of *k* that results in the best performance.

k



If we have more data, we could make a plot like this one where the accuracy of the KNN is plotted against different values of *k*.

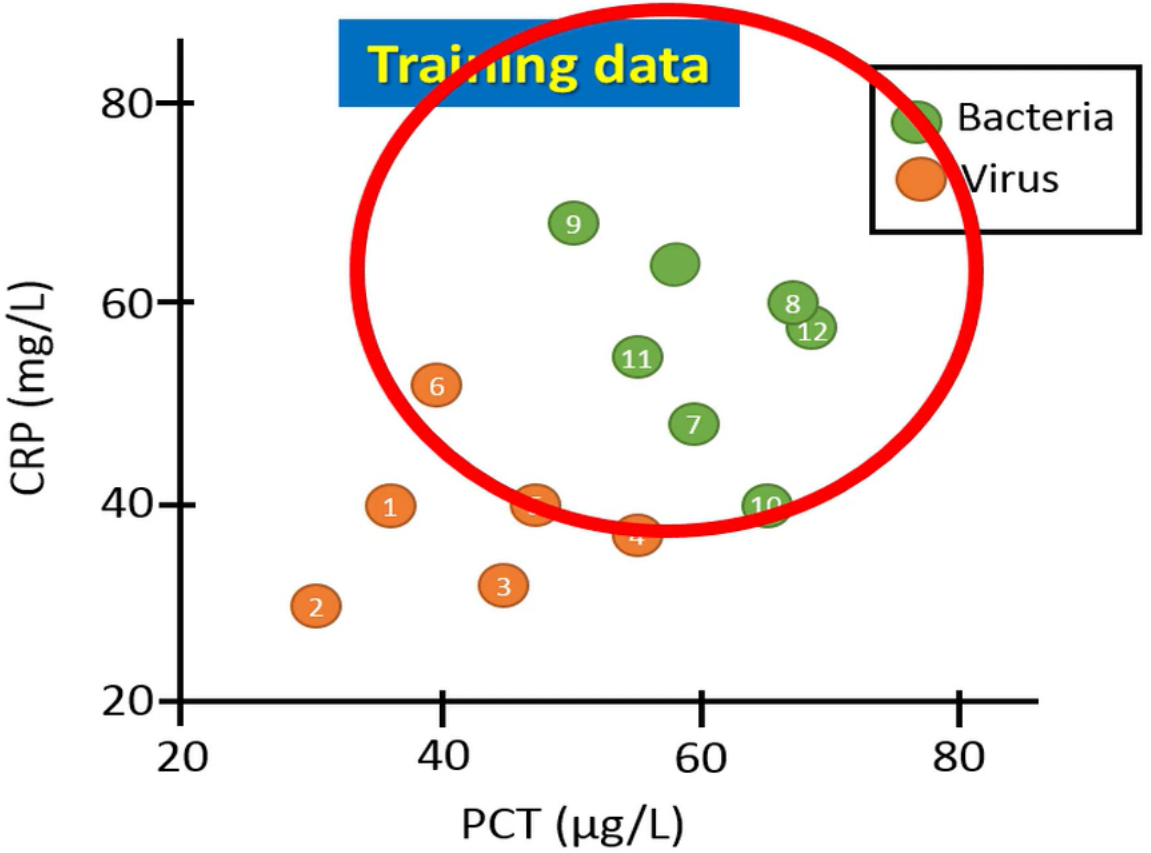
k



If we have more data, we could make a plot like this one where the accuracy of the KNN is plotted against different values of *k*.

Validate

Test data



We then test if the KNN correctly predicts this observation as bacteria based on the seven nearest neighbors.

KNN

More than two groups

Standardization

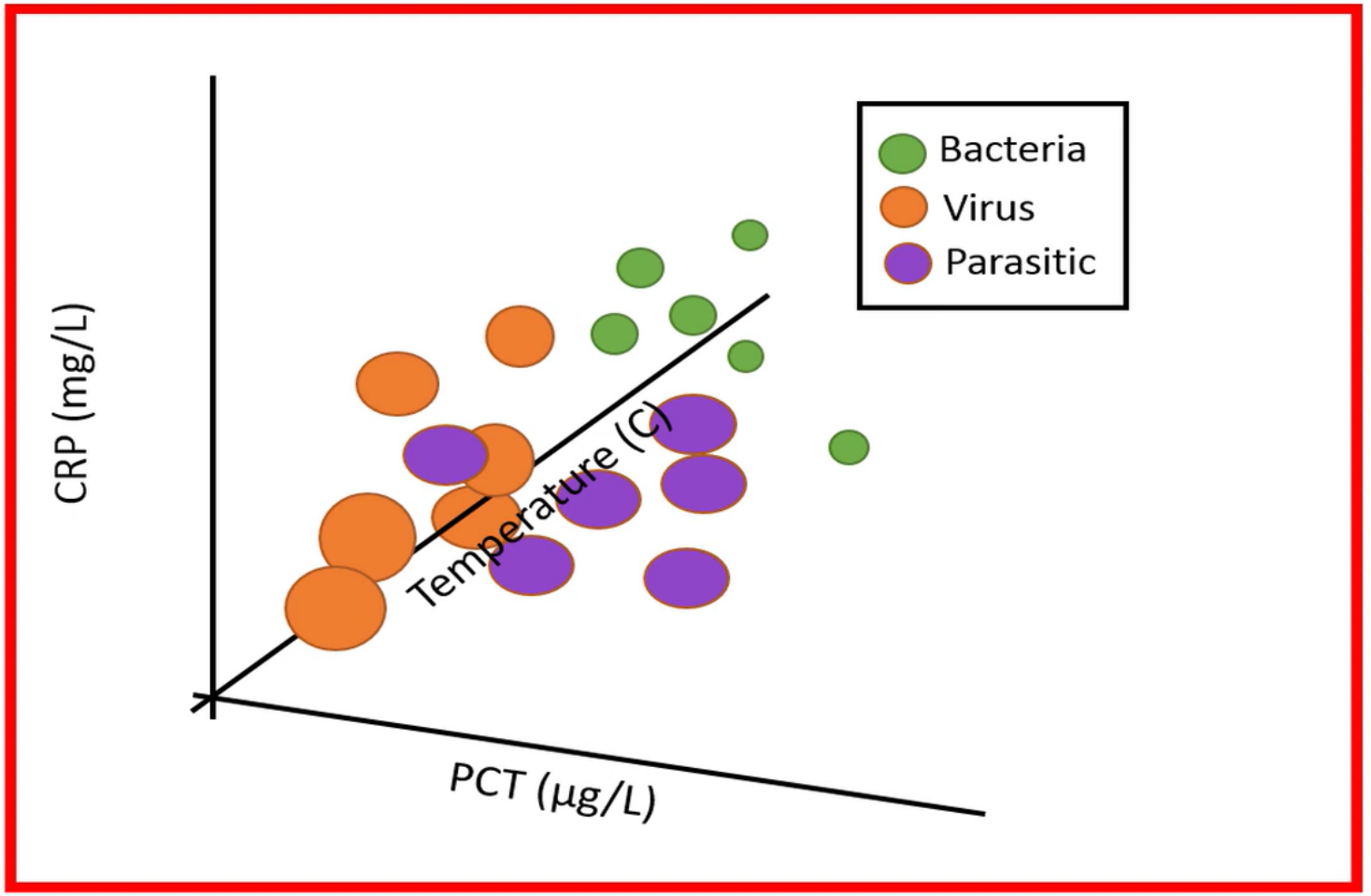
KNN

Infection	CRP (mg/L)	PCT (µg/L)	Temp (C)
Viral	40	36	37.9
Viral	30	30	36.8
Viral	32	45	37.5
Viral	37	55	38
Viral	40	47	39.1
Viral	52	40	37.5
Bacterial	48	59	38.2
Bacterial	60	67	38
Bacterial	68	50	39.2
Bacterial	40	65	40.8
Bacterial	55	55	39.2
Bacterial	58	68	41.1
Parasitic	40	40	41.1
Parasitic	33	40	36.9
Parasitic	20	38	37.1
Parasitic	70	38	40.2
Parasitic	28	37	40.4
Parasitic	29	70	38.7

as well as on other clinical variables, such as the body temperature,

KNN

Infection	CRP (mg/L)	PCT (µg/L)	Temp (C)
Viral	40	36	37.9
Viral	30	30	36.8
Viral	32	45	37.5
Viral	37	55	38
Viral	40	47	39.1
Viral	52	40	37.5
Bacterial	48	59	38.2
Bacterial	60	67	38
Bacterial	68	50	39.2
Bacterial	40	65	40.8
Bacterial	55	55	39.2
Bacterial	58	68	41.1
Parasitic	40	40	41.1
Parasitic	33	40	36.9
Parasitic	20	38	37.1
Parasitic	70	38	40.2
Parasitic	28	37	40.4
Parasitic	29	70	38.7



then we can use KNN to predict if someone has a viral, bacterial or parasitic infection based on three clinical variables.

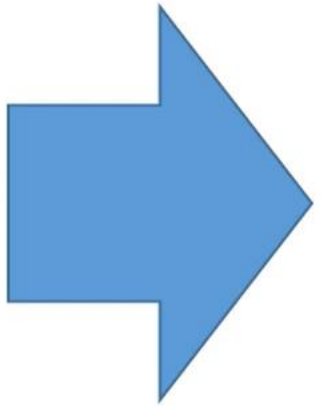
Standardize the data

Infection	CRP (mg/L)	PCT (µg/L)	Temp (C)
Viral	40	36	37.9
Viral	30	30	36.8
Viral	32	45	37.5
Viral	37	55	38
Viral	40	47	39.1
Viral	52	40	37.5
Bacterial	48	59	38.2
Bacterial	60	67	38
Bacterial	68	50	39.2
Bacterial	40	65	40.8
Bacterial	55	55	39.2
Bacterial	58	68	41.1
Parasitic	40	40	41.1
Parasitic	33	40	36.9
Parasitic	20	38	37.1
Parasitic	70	38	40.2
Parasitic	28	37	40.4
Parasitic	29	70	38.7

For example, the body temperature only spans between 36.8 to 41.1, which means that the distance in this dimension will have a very small impact compared to the other variables during the classification.

Standardize the data

Infection	CRP (mg/L)	PCT (µg/L)	Temp (C)
Viral	40	36	37.9
Viral	30	30	36.8
Viral	32	45	37.5
Viral	37	55	38
Viral	40	47	39.1
Viral	52	40	37.5
Bacterial	48	59	38.2
Bacterial	60	67	38
Bacterial	68	50	39.2
Bacterial	40	65	40.8
Bacterial	55	55	39.2
Bacterial	58	68	41.1
Parasitic	40	40	41.1
Parasitic	33	40	36.9
Parasitic	20	38	37.1
Parasitic	70	38	40.2
Parasitic	28	37	40.4
Parasitic	29	70	38.7



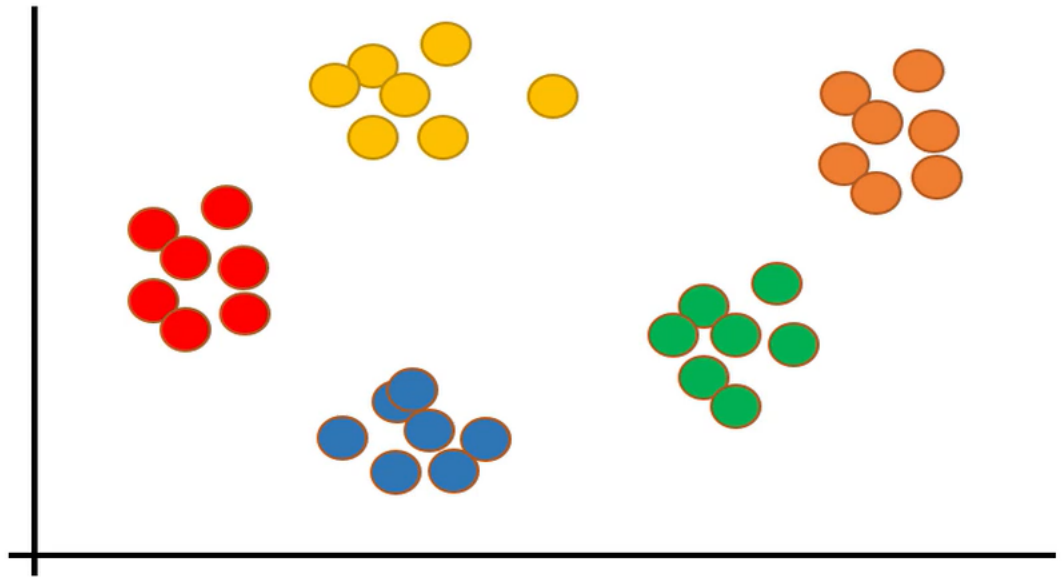
$$Z_i = \frac{X_i - \overline{X}}{\mathrm{SD}}$$

Infection	CRP (mg/L)	PCT (µg/L)	Temp (C)
Viral	-0,23	-1,02	-0,59
Viral	-0,93	-1,49	-1,35
Viral	-0,79	-0,31	-0,87
Viral	-0,44	0,48	-0,52
Viral	-0,23	-0,15	0,23
Viral	0,60	-0,70	-0,87
Bacterial	0,32	0,80	-0,39
Bacterial	1,16	1,43	-0,52
Bacterial	1,71	0,09	0,30
Bacterial	-0,23	1,27	1,40
Bacterial	0,81	0,48	0,30
Bacterial	1,02	1,51	1,61
Parasitic	-0,23	-0,70	1,61
Parasitic	-0,72	-0,70	-1,28
Parasitic	-1,62	-0,86	-1,14
Parasitic	1,85	-0,86	0,99
Parasitic	-1,06	-0,94	1,13
Parasitic	-0,99	1,66	-0,04

To give the variables equal weights in the classification process, we can first standardize the data so that all variables have a mean of zero and a standard deviation of one.

Advantages

- It is a very simple method and easy to understand.
- It is based on local data points, which might be beneficial for data sets involving many groups with local clusters.



Disadvantages

- All training data is used every time we should predict something. This means that the data must be stored everywhere where you like to use the classifier.
- For very large data sets, the classification might be computationally expensive for prediction.
- It is sensitive to imbalanced data sets.