DSCI 100 - Introduction to Data Science

Lecture 7 - Classification continued

2019-02-14

Continuing with the classification problem

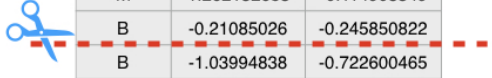
Can we use data we have seen in the past, to predict something about the future?

Unanswered questions from last week

1. Is our model any good?
2. How do we choose k?

1. Is our model any good?

Creating the training and validation sets



original data collected

Y = Class	*X1* = concavity	*X2* = perimeter
B	-0.39874785	-0.332451413
B	-0.87966023	-1.253810272
M	1.987839169	-0.249719577
B	-0.97490322	-1.488018741
M	1.262132653	-0.114908349
B	-0.21085026	-0.245850822
B	-1.03994838	-0.722600465
M	1.278909223	-0.031581320
B	-0.86355473	-0.702066304
M	-0.605339335	0.492188574

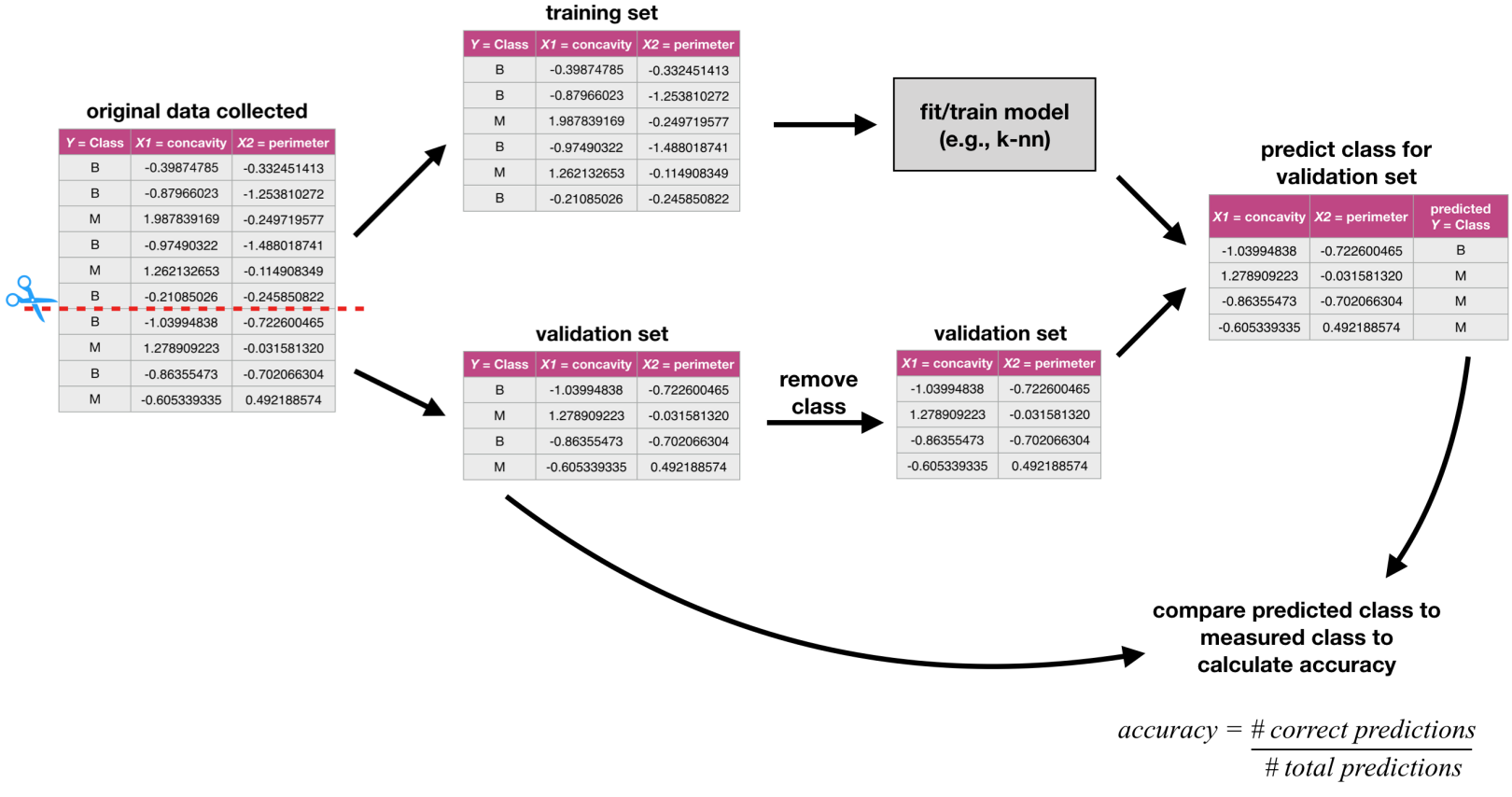
training set

Y = Class	*X1* = concavity	*X2* = perimeter
B	-0.39874785	-0.332451413
B	-0.87966023	-1.253810272
M	1.987839169	-0.249719577
B	-0.97490322	-1.488018741
M	1.262132653	-0.114908349
B	-0.21085026	-0.245850822

validation set

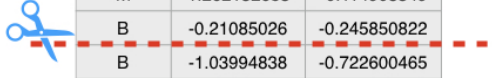
Y = Class	*X1* = concavity	*X2* = perimeter
B	-1.03994838	-0.722600465
M	1.278909223	-0.031581320
B	-0.86355473	-0.702066304
M	-0.605339335	0.492188574

Assumption -> there is no order to the original data collected



Is one accuracy measurement good enough?

Creating the training and validation sets



original data collected

Y = Class	*X1* = concavity	*X2* = perimeter
B	-0.39874785	-0.332451413
B	-0.87966023	-1.253810272
M	1.987839169	-0.249719577
B	-0.97490322	-1.488018741
M	1.262132653	-0.114908349
B	-0.21085026	-0.245850822
B	-1.03994838	-0.722600465
M	1.278909223	-0.031581320
B	-0.86355473	-0.702066304
M	-0.605339335	0.492188574

training set

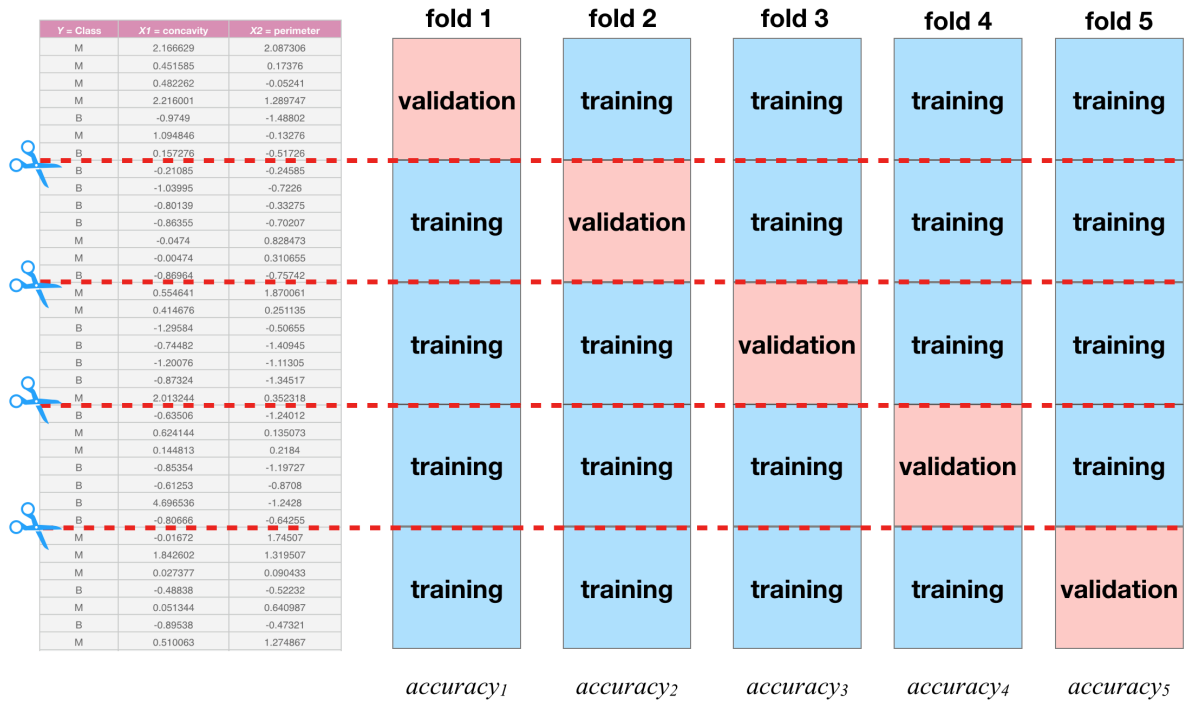
Y = Class	*X1* = concavity	*X2* = perimeter
B	-0.39874785	-0.332451413
B	-0.87966023	-1.253810272
M	1.987839169	-0.249719577
B	-0.97490322	-1.488018741
M	1.262132653	-0.114908349
B	-0.21085026	-0.245850822

validation set

Y = Class	*X1* = concavity	*X2* = perimeter
B	-1.03994838	-0.722600465
M	1.278909223	-0.031581320
B	-0.86355473	-0.702066304
M	-0.605339335	0.492188574

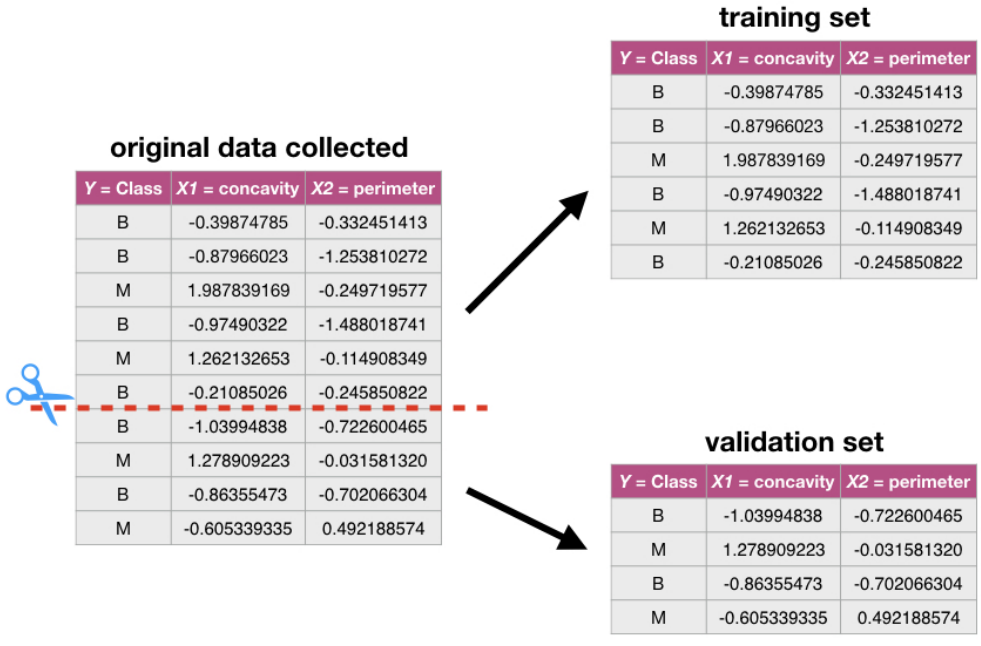
Assumption -> there is no order to the original data collected

Cross-validation as an alternative approach



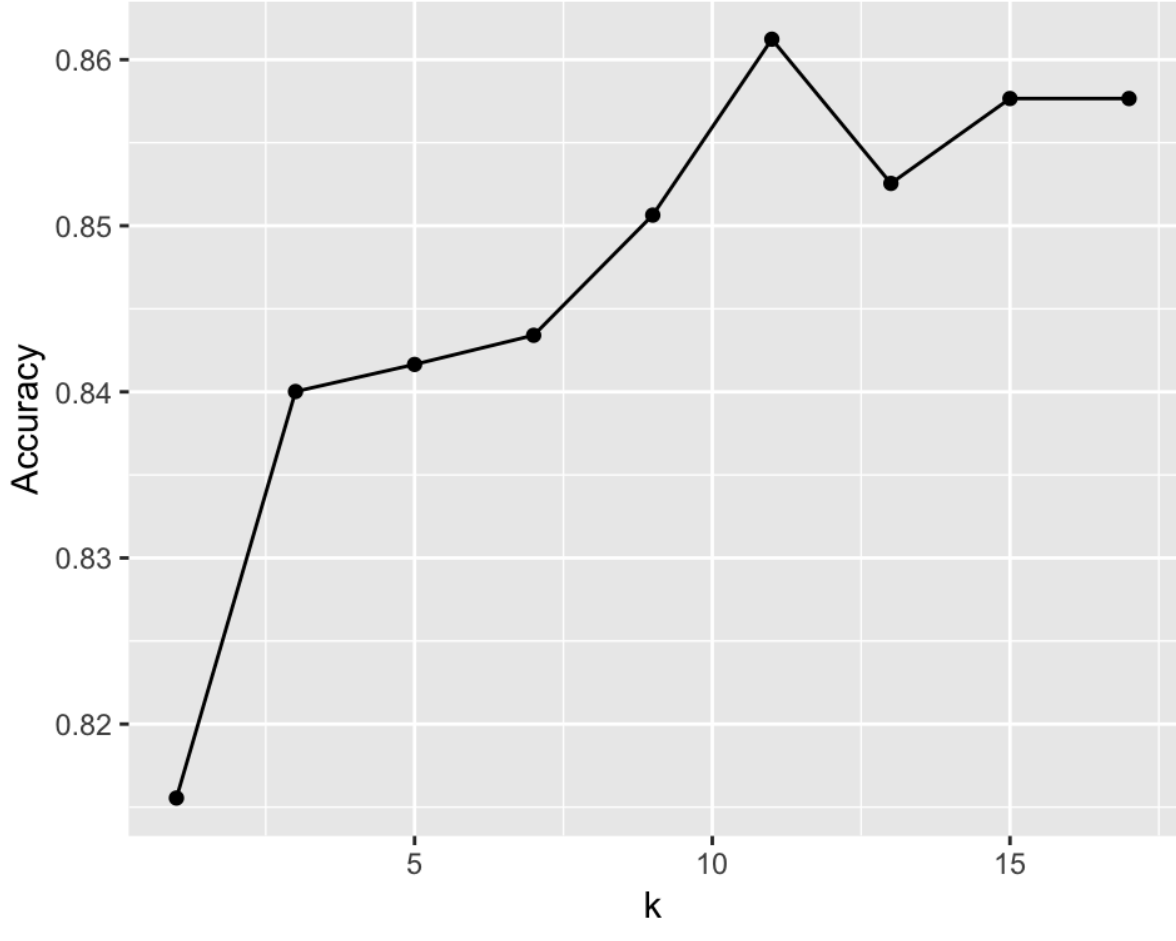
$$\text{cross-validation accuracy} = \frac{\text{accuracy}_1 + \text{accuracy}_2 + \text{accuracy}_3 + \text{accuracy}_4 + \text{accuracy}_5}{\#\,\text{folds}}$$

Creating the training and validation sets

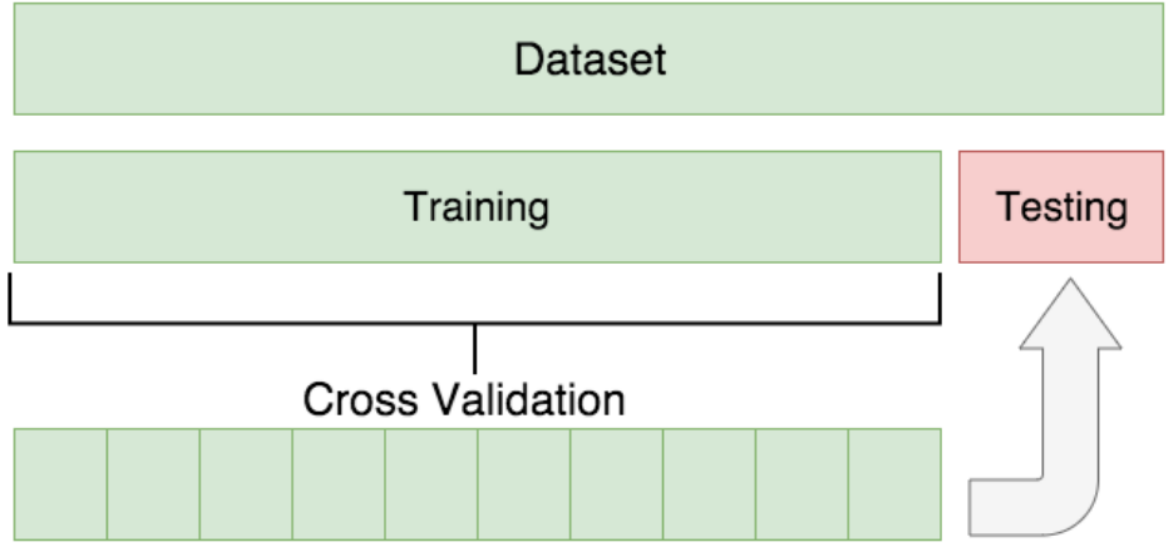


Assumption -> there is no order to the original data collected

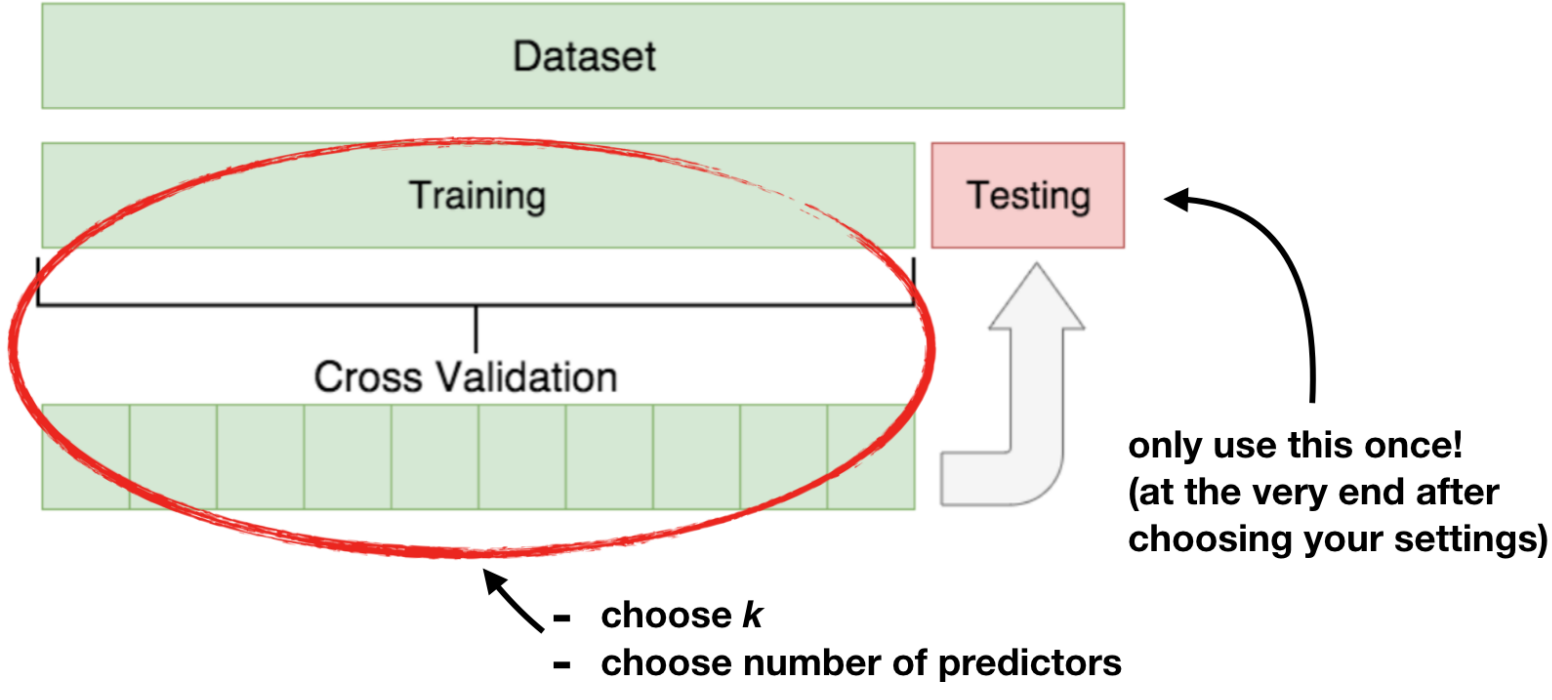
2. How do we choose k?


0.86
0.85
0.84
0.83
0.82
Accuracy
5
10
15
k

The big picture



The big picture



Why are we doing all this???

Our question:

- Can we use past information to predict the class labels of new observations we don't have labels for?

- We can always do this, but we might only want to do this if we have evidence we can do this well.

Class activity 1

- In your group, discuss and explain cross-validation in your own words. Post your group's answer as a response to this post in Piazza.

Class activity 2

- In your group, discuss and explain what a test, validation and training data set are in your own words. Post your group's answer as a response to this post in Piazza.