Introduction to Data Science

DSA1101

Semester 1, 2018/2019
Week 4

Diagnostics of Classifiers

Diagnostics of Classifiers

- We have studied the k-nearest neighbor algorithm as an example of a classifier
- However, there is a need to evaluate the performance of the classifiers

Diagnostics of Classifiers

- k-nearest neighbor is often used as a classifier to assign class labels to a person, item, or transaction.
- In general, for two class labels, C and $\neg C$, where $\neg C$ denotes "not C," some working definitions and formulas follow:

- True Positive: Predict C, when actually C
- True Negative: Predict $\neg C$, when actually $\neg C$
- False Positive: Predict C, when actually $\neg C$
- False Negative: Predict $\neg C$, when actually C

Diagnostics of Classifiers

- We will study the *confusion matrix* which is a specific table layout that allows visualization of the performance of a classifier.
- In a two-class classification, a preset threshold may be used to separate positives from negatives (e.g. we used the majority rule, $\hat{Y} < 0.5$, in the k-nearest neighbor example).
-

		Predicted Class	
		Positive	**Negative**
Actual Class	**Positive**	True Positives (TP)	False Negatives (FN)
	Negative	False Positives (FP)	True Negatives (TN)

Diagnostics of Classifiers

- TP and TN are the correct guesses.
- A good classifier should have large TP and TN and small (ideally zero) numbers for FP and FN.
-

		Predicted Class	
		Positive	**Negative**
Actual Class	**Positive**	True Positives (TP)	False Negatives (FN)
	Negative	False Positives (FP)	True Negatives (TN)

Result + Predicted Answer

Diagnostics of Classifiers: example

- A testing set of 100 emails (with their spam or non-spam label known)
- Example confusion matrix of a k-nearest neighbor classifier to predict if each email is spam or not

		Predicted Class		
		Spam	**Non-Spam**	Total
Actual Class	**Spam**	3	8	11
	Non-Spam	2	87	89
Total		5	95	100

Diagnostics of Classifiers

- The *accuracy* (or the overall success rate) is a metric defining the rate at which a model has classified the records correctly.
- It is defined as the sum of TP and TN divided by the total number of instances:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%$$

Diagnostics of Classifiers

- A good model should have a high accuracy score, but having a high accuracy score alone does not guarantee the model is well established.
- We will introduce more fine-grained measures better evaluate the performance of a classifier.

Diagnostics of Classifiers

- The true positive rate (TPR) shows the proportion of positive instances the classifier correctly identified:

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

		Predicted Class	
		Positive	**Negative**
Actual Class	**Positive**	True Positives (TP)	False Negatives (FN)
	Negative	False Positives (FP)	True Negatives (TN)

Diagnostics of Classifiers

- The false positive rate (FPR) shows what percent of negatives the classifier marked as positive.
- The FPR is also called the false alarm rate or the type I error rate

$$\text{FPR} = \frac{FP}{FP + TN}$$

		Predicted Class	
		Positive	**Negative**
Actual Class	**Positive**	True Positives (TP)	False Negatives (FN)
	Negative	False Positives (FP)	True Negatives (TN)

Diagnostics of Classifiers

- The false negative rate (FNR) shows what percent of positives the classifier marked as negatives.
- It is also known as the miss rate or type II error rate.

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

		Predicted Class	
		Positive	**Negative**
Actual Class	**Positive**	True Positives (TP)	False Negatives (FN)
	Negative	False Positives (FP)	True Negatives (TN)

Diagnostics of Classifiers

- Precision is the percentage of instances marked positive that really are positive:

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

		Predicted Class	
		Positive	**Negative**
Actual Class	**Positive**	True Positives (TP)	False Negatives (FN)
	Negative	False Positives (FP)	True Negatives (TN)

Diagnostics of Classifiers

- A well-performed model should have a high TPR that is ideally 1 and a low FPR and FNR that are ideally 0.
- In reality, it is rare to have TPR $= 1$, FPR $= 0$, and FNR $= 0$, but these measures are useful to compare the performance of multiple models that are designed for solving the same problem.
- Note that in general, the model that is more preferable may depend on the business situation.

Diagnostics of Classifiers

- During the discovery phase of the data analytics lifecycle, the team should have learned from the business what kind of errors can be tolerated.
- Some business situations are more tolerant of type I errors, whereas others may be more tolerant of type II errors.

  Type 1: dư thừa
  Type 2: thiếu sót

Diagnostics of Classifiers

- Consider the example of e-mail spam filtering.
- Some people (such as busy executives) only want important e-mail in their inbox and are tolerant of having some less important e-mail end up in their spam folder as long as no spam is in their inbox.
- In this case, a higher false positive rate (FPR) or type I error can be tolerated.

Diagnostics of Classifiers

- Other people may not want any important or less important e-mail to be specified as spam and are willing to have some spam in their inboxes as long as no important e-mail makes it into the spam folder.
- In this case, a higher false negative rate (FNR) or type II error can be tolerated.

Diagnostics of Classifiers

- Another example involves medical screening during an infectious disease outbreak.
- The cost of having a person, who has the disease, to be instead diagnosed as disease-free is extremely high, since the disease may be highly contagious.
- Therefore, the false negative rate (FNR) or type II error needs to be low.
- A higher false positive rate (FPR) or type I error can be tolerated.

Diagnostics of Classifiers

- Third example involves security screening at the airport.
- The cost of a false negative in this scenario is extremely high (not detecting a bomb being brought onto a plane could result in hundreds of deaths) whilst the cost of a false positive is relatively low (a reasonably simple further inspection)
- Therefore, a higher false positive rate (FPR) or type I error can be tolerated, in order to keep the false negative rate (FNR) or type II error low.

Diagnostics of Classifiers: example

- $$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%$$
$$= \frac{3 + 87}{3 + 87 + 2 + 8} \times 100\% = 90\%$$

		Predicted Class		
		Spam	**Non-Spam**	Total
Actual Class	**Spam**	3	8	11
	Non-Spam	2	87	89
Total		5	95	100

Diagnostics of Classifiers: example

- $$\text{TPR} = \frac{TP}{TP + FN} = \frac{3}{3 + 8} \approx 0.273$$

		Predicted Class		
		Spam	**Non-Spam**	Total
Actual Class	**Spam**	3	8	11
	Non-Spam	2	87	89
Total		5	95	100

Diagnostics of Classifiers: example

- $$\text{FPR} = \frac{FP}{FP + TN} = \frac{2}{2 + 87} \approx 0.022$$

		Predicted Class		
		Spam	**Non-Spam**	Total
Actual Class	**Spam**	3	8	11
	Non-Spam	2	87	89
Total		5	95	100

Diagnostics of Classifiers: example

- $$\text{FNR} = \frac{FN}{TP + FN} = \frac{8}{3 + 8} \approx 0.727$$

		Predicted Class		
		Spam	**Non-Spam**	Total
Actual Class	**Spam**	3	8	11
	Non-Spam	2	87	89
Total		5	95	100

Diagnostics of Classifiers: example

- $$\text{Precision} = \frac{TP}{TP + FP} = \frac{3}{3 + 2} = 0.6$$

		Predicted Class		
		Spam	**Non-Spam**	Total
Actual Class	**Spam**	3	8	11
	Non-Spam	2	87	89
Total		5	95	100

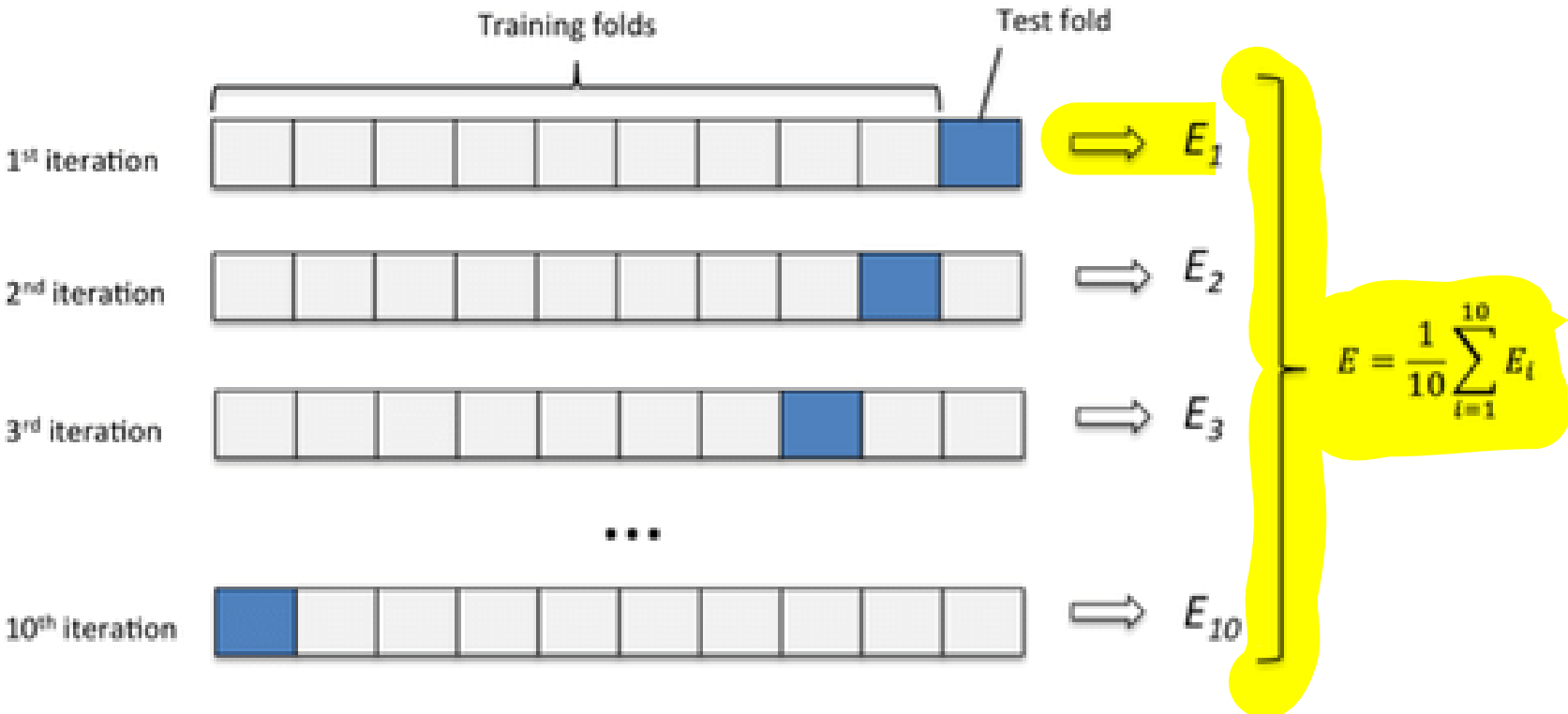
Diagnostics of Classifiers

- We have studied a number of measures that can be used to evaluate the performance of a classifier.
- In practice, when we are presented with a dataset, how should we go about estimating these performance measures?
- A common practice is to perform *N*-Fold Cross-Validation

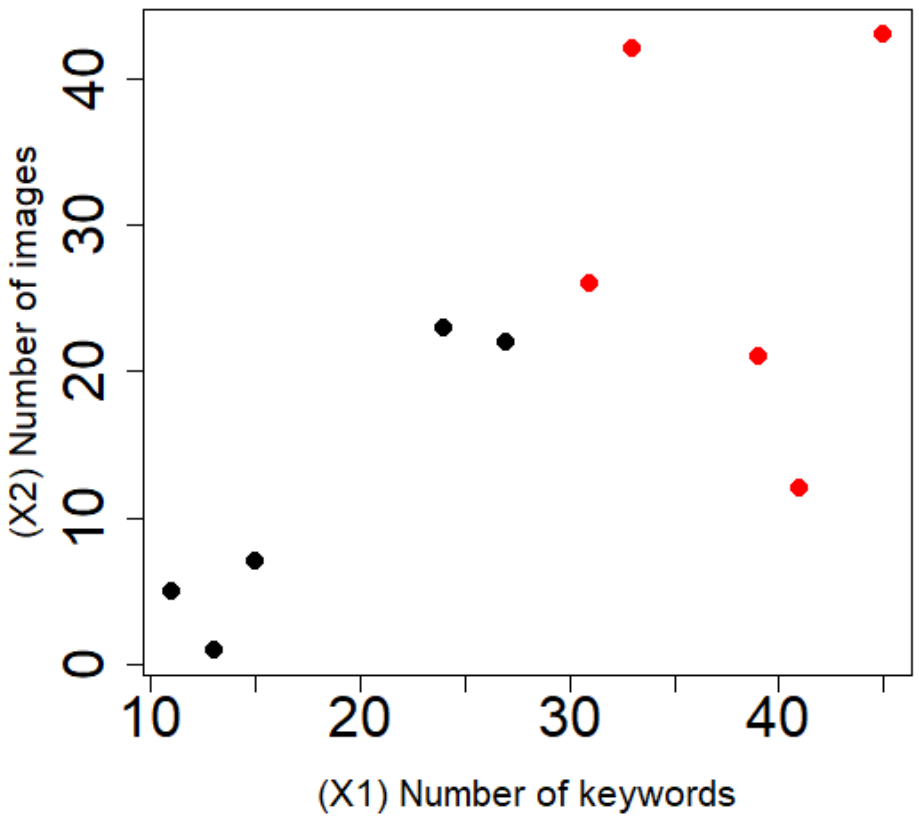
Diagnostics of Classifiers

- The entire dataset is randomly split into N datasets of approximately equal size.
- N-1 of these datasets are treated as the training dataset, while the remaining one is the test dataset. A measure of the model error is obtained.
- This process is repeated across the various combinations of N datasets taken $N - 1$ at a time.
- The observed N models errors are averaged across the N folds.

Diagnostics of Classifiers



Example: Anti-spam techniques

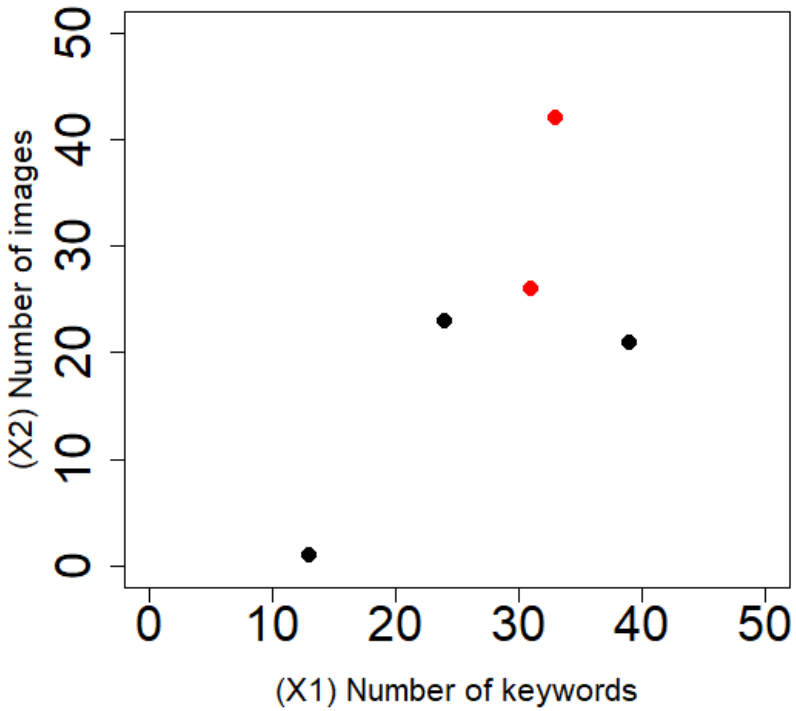


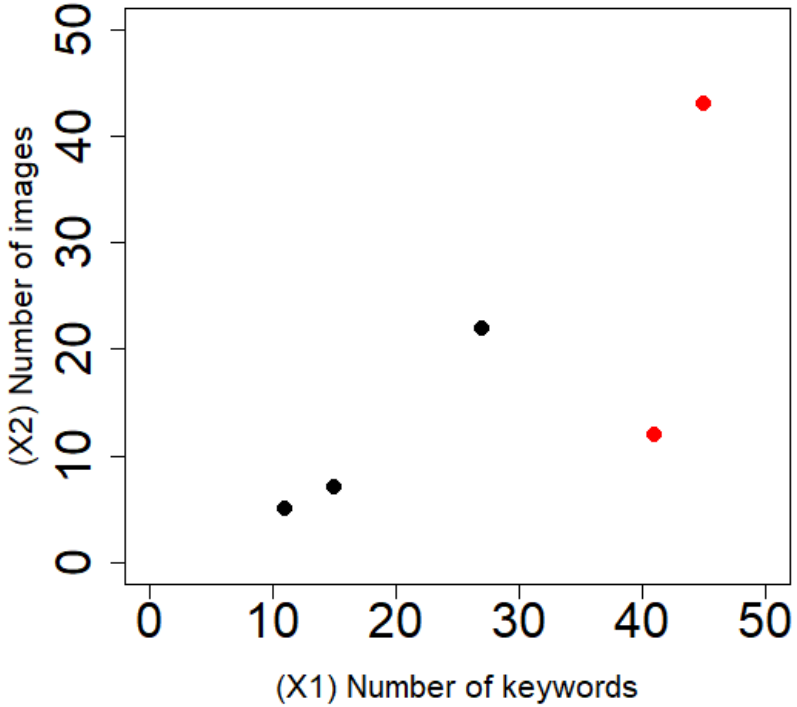
- Let us illustrate N-Fold Cross-Validation with an example with the k-nearest neighbor classfier for spams, where we specify $k = 1$.
- Suppose our dataset consists of 10 data points.

Diagnostics of Classifiers

- For 2-fold cross validation, we randomly split the whole dataset of 10 points into two datasets of 5 points each

Example: Anti-spam techniques



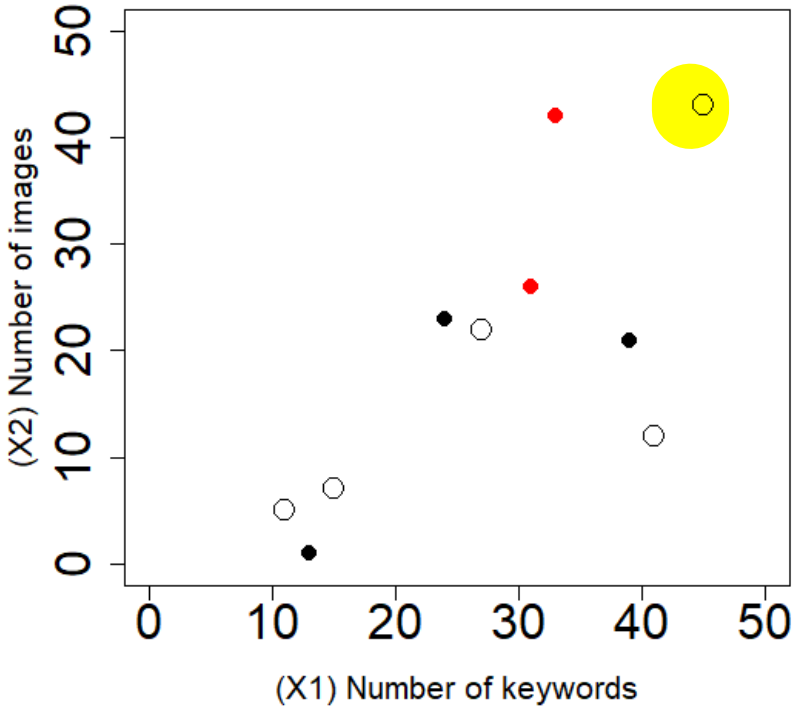


Diagnostics of Classifiers

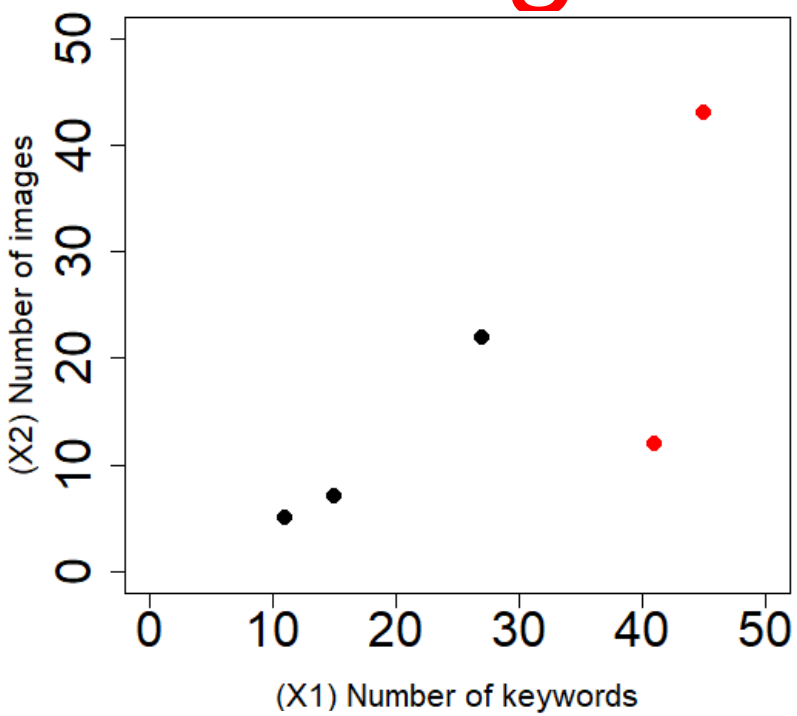
- For the first iteration, we use the first dataset as the training set and the second dataset as the testing set.

Example: Anti-spam techniques

Test set



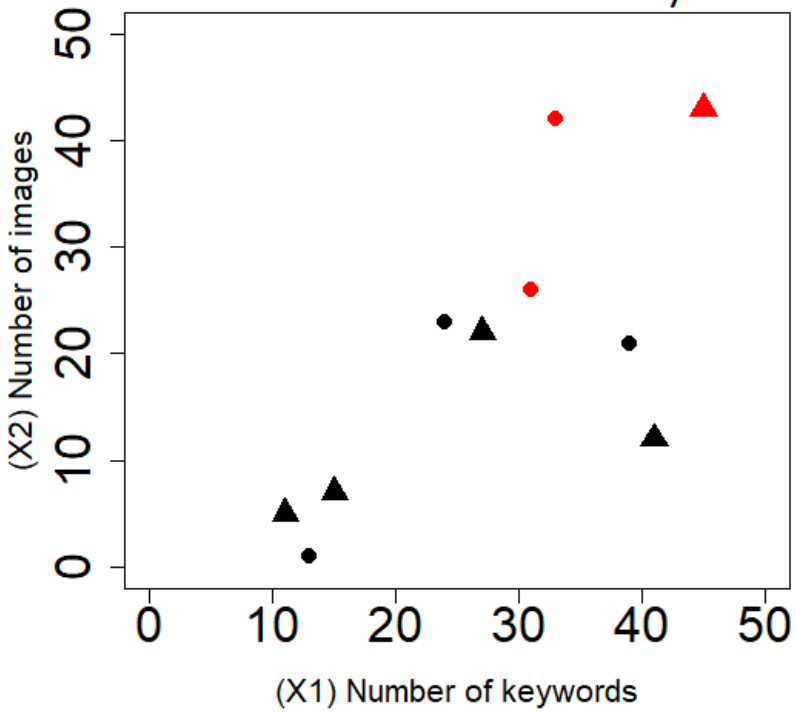
Training set



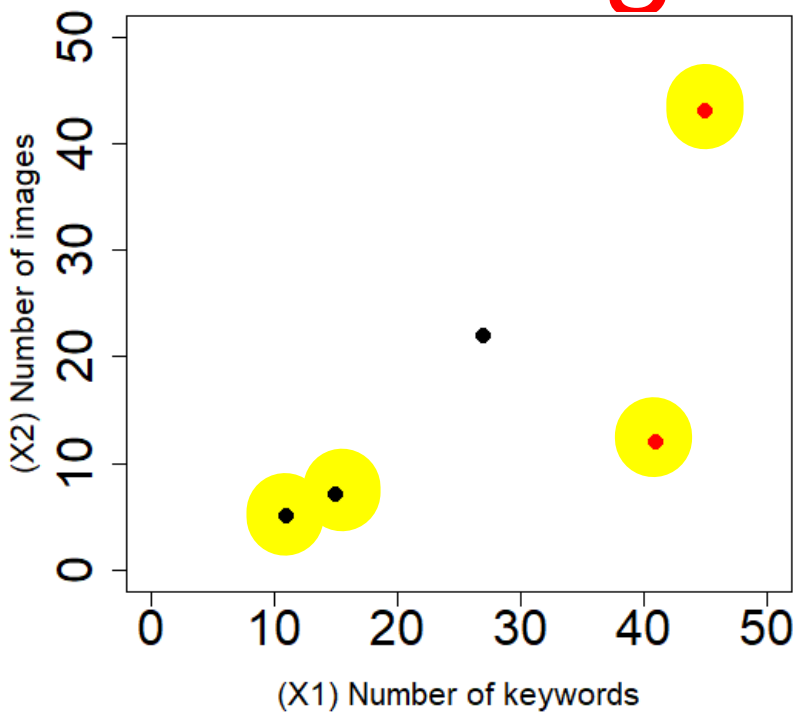
first iteration

Example: Anti-spam techniques

Test set (Triangles are predicted labels)



Training set



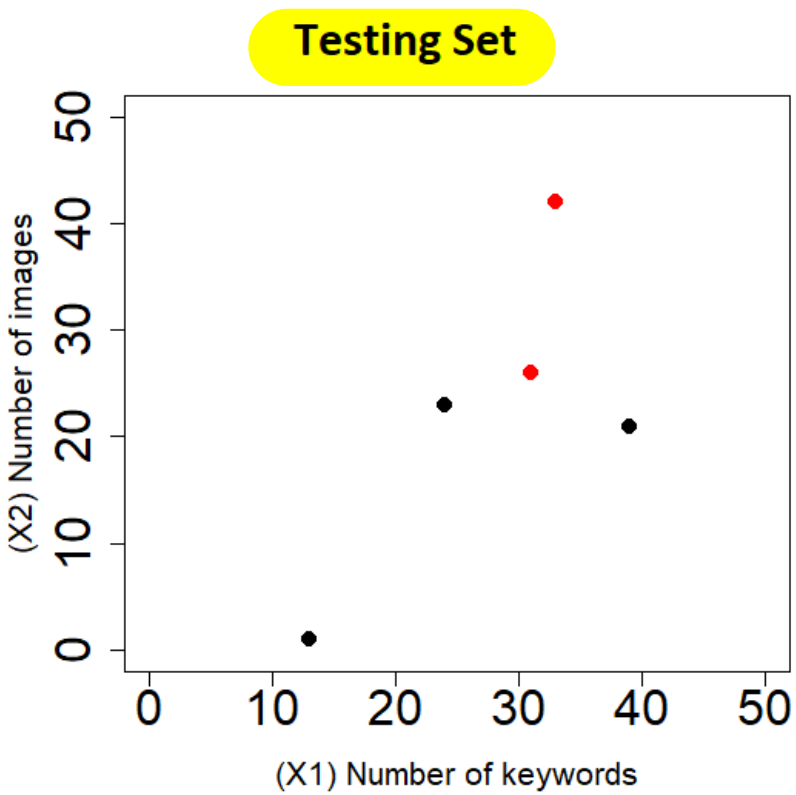
Diagnostics of Classifiers

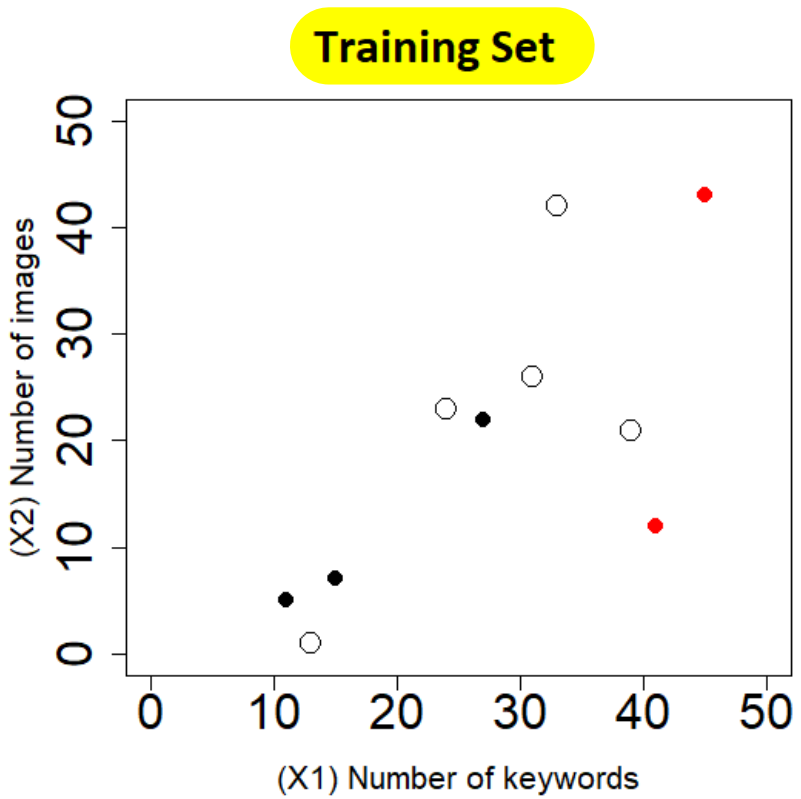
- In this iteration, we estimate the accuracy of the 1-nearest neighbor algorithm to be equal to $\frac{4}{5}$

Diagnostics of Classifiers

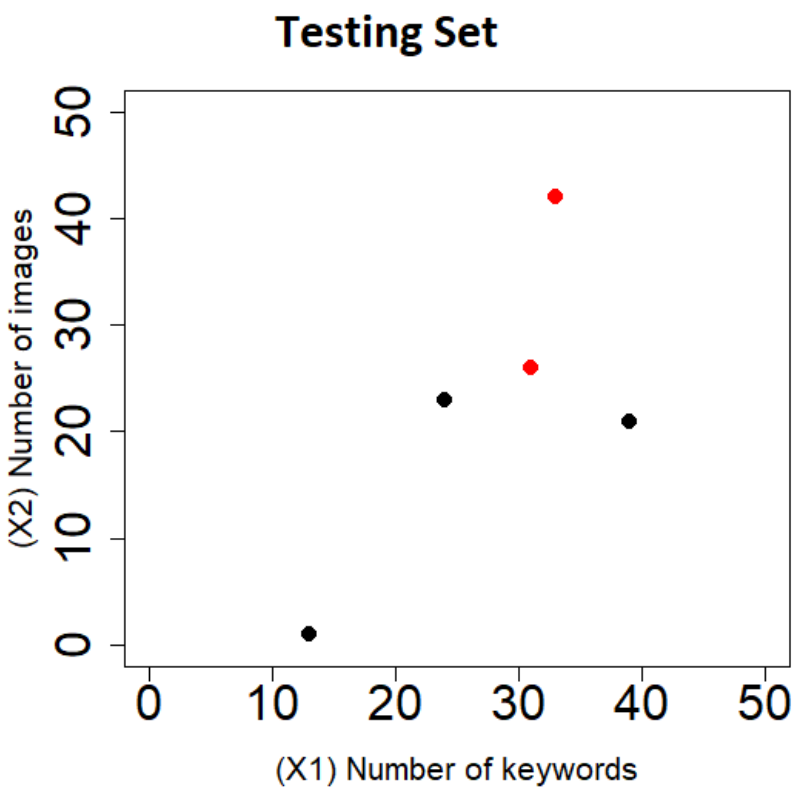
- For the second iteration, we use the second dataset as the training set and the first dataset as the testing set.

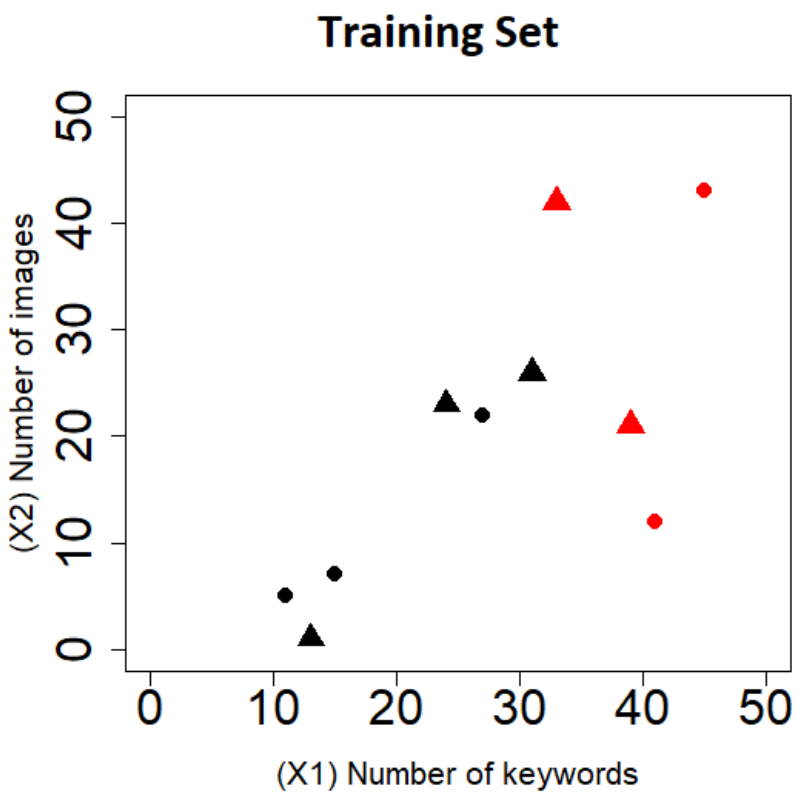
Example: Anti-spam techniques





Example: Anti-spam techniques





Diagnostics of Classifiers

- In this iteration, we estimate the accuracy of the 1-nearest neighbor algorithm to be equal to $\frac{3}{5}$

Diagnostics of Classifiers

- Therefore, based on 2-fold cross validation, the accuracy of the 1-nearest neighbor algorithm is estimated to be $(\frac{4}{5} + \frac{3}{5})/2 = \frac{7}{10}$.
- We will continue with more examples next week to ground ideas.