Text Classification

Learning Objectives

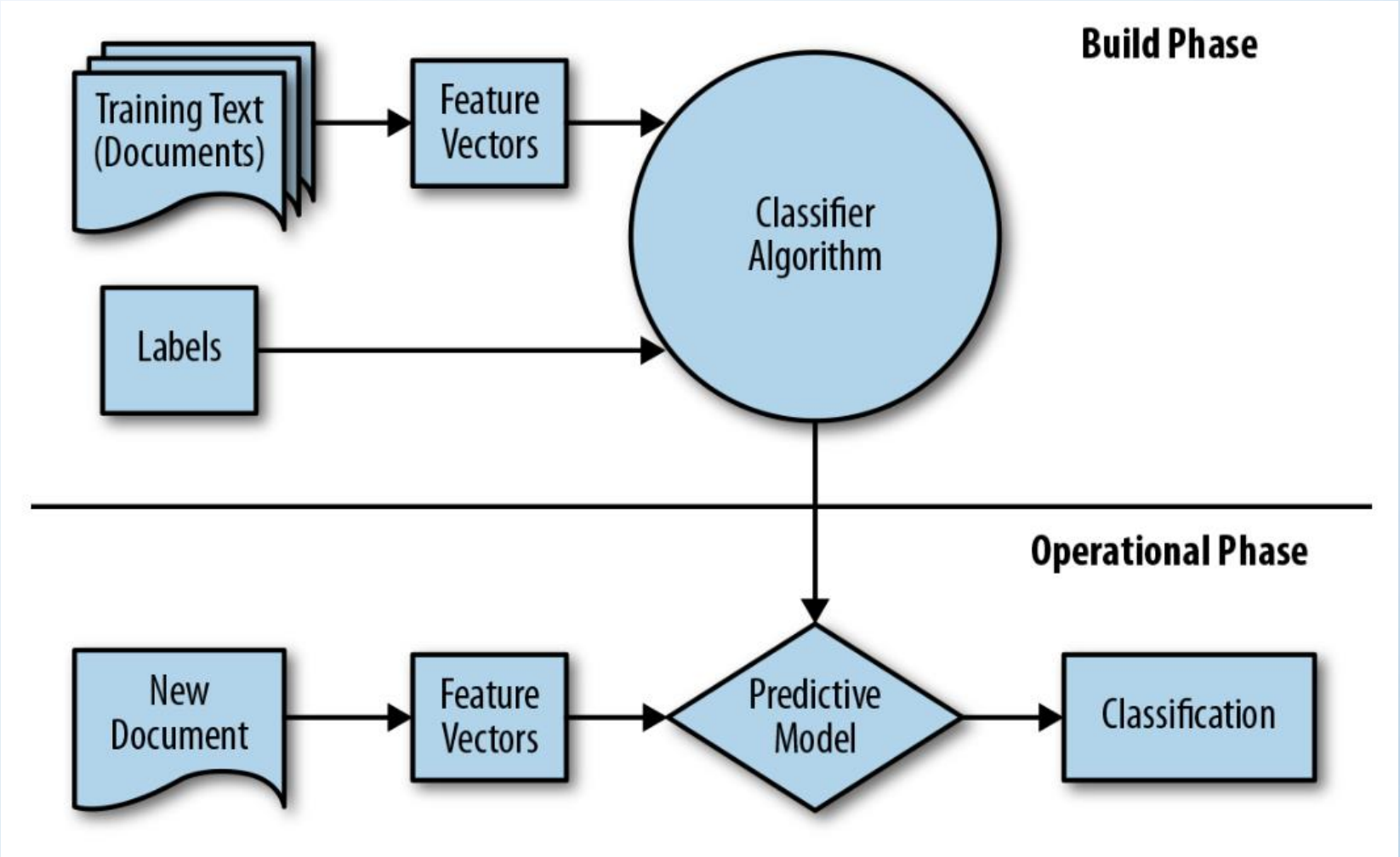
- Describe each component in Text Classification System
- Calculate classification evaluation metric
- Build and use Logistic Regression model in Text Classification System
- Build and use Naive Bayes model in Text Classification System

Logistic Regression

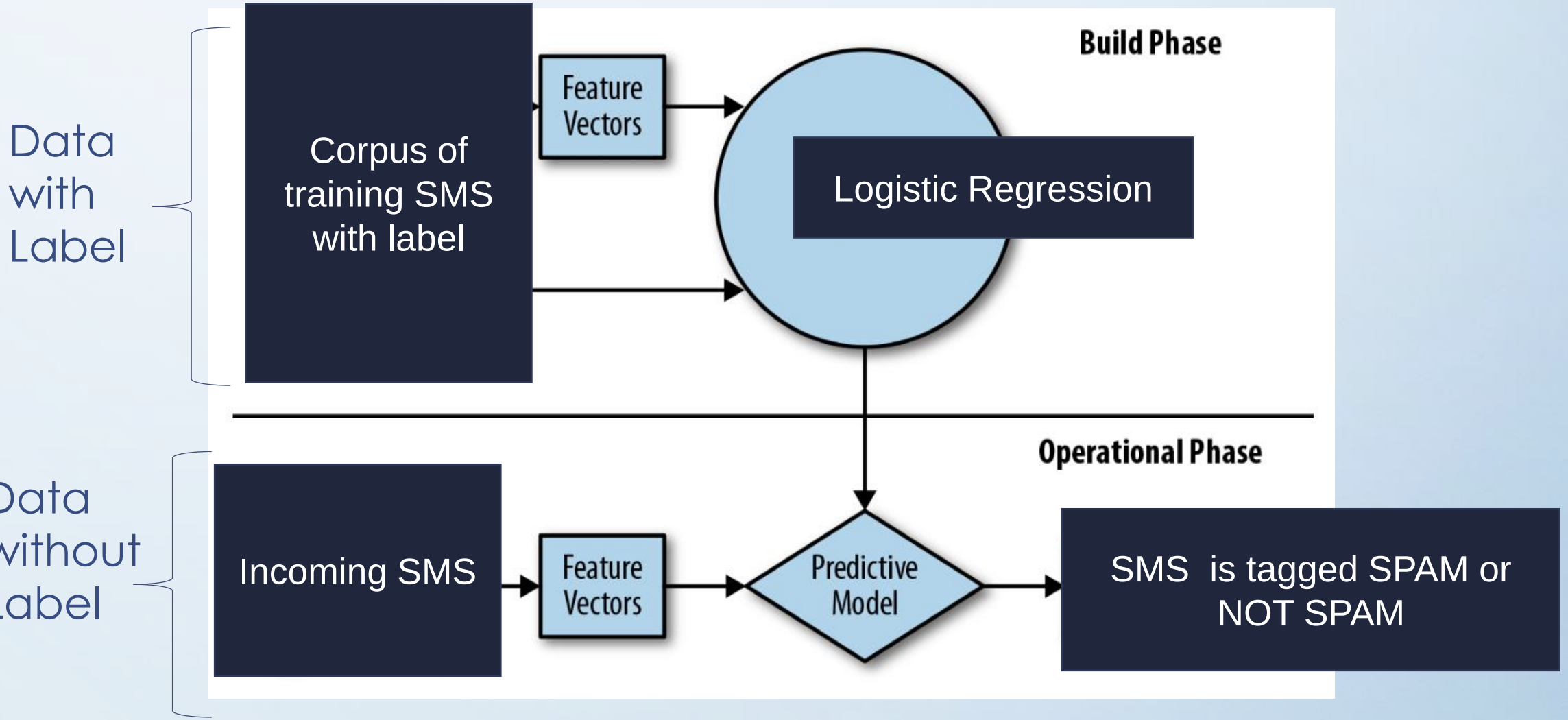
Workflow

Data with Label

Data without Label



Example


Data with Label
Corpus of training SMS with label
Feature Vectors
Build Phase
Logistic Regression
Operational Phase
Data without Label
Incoming SMS
Feature Vectors
Predictive Model
SMS is tagged SPAM or NOT SPAM

Build Phase

Steps for classification with NLP

- Prepare the data: Read in labelled data and pre-process the data
- Split the data: Separate inputs and outputs into a training set and a test set

Numerically encode inputs: Use either Count Vectorizer or TF-IDF Vectorizer

- Fit a model: Fit a model on the training data and apply the fitted model to the test set
- Evaluate the model: Decide how good the model is by calculating various error metrics
- Save the model: Output the model and vectorizer to external files

Exercise 1: Build Logistic Regression model to predict ham or spam for SMS message

Step 1:

Read the SMSSpamCollection.txt and pre-process the data

- Remove all the words with numbers and pure numbers. e.g. 21st, 2005
- Remove all the punctuation. e.g. !, ?
- Convert capital letter to small letter

	label	new_text
0	ham	ok lar joking wif u oni
1	spam	free entry in a wkly comp to win fa cup final tkts may text fa to to receive entry questionstd txt ratetcs apply s
2	ham	u dun say so early hor u c already then say
3	ham	nah i dont think he goes to usf he lives around here though
4	spam	freemsg hey there darling its been weeks now and no word back id like some fun you up for it still tb ok xxx std chgs to send £ to rcv

Exercise 1: Build Logistic Regression model to predict ham or spam for SMS message

Step 2:

Split the dataset into training set and test set

- Test dataset = 30% of observation and Training dataset = 70% of observation
- Random state =42, so we all get the same random train and test split

```
The size of original dataset: (5571,)
The size of training dataset: (3899,)
The size of test dataset: (1672,)
```

Exercise 1: Build Logistic Regression model to predict ham or spam for SMS message

Step 3:

Convert the text to vectors using count vectorizer

```
The dimensions of the training set: (3899, 6663)
The dimensions of the test set: (1672, 6663)
The features:
 ['aa' 'aah' 'aaooooright' ... 'zoom' 'zouk' 'üll']
```

Exercise 1: Build Logistic Regression model to predict ham or spam for SMS message

Step 4:

Fit Logistic Regression model on training data and apply the fitted model to test data. Predict the test data.

```
1   print(list(y_test[:10]))
2   print(list(y_pred_cv[:10]))
```

```
['ham', 'spam', 'ham', 'spam', 'ham', 'ham', 'ham', 'ham', 'ham', 'ham']
['ham', 'spam', 'ham', 'spam', 'ham', 'ham', 'ham', 'ham', 'ham', 'ham']
```

Exercise 1: Build Logistic Regression model to predict ham or spam for SMS message

Step 5:

Evaluate the mode. Decide how good the model is by calculating various metrics

- Confusion Metrix
- Precision
- Recall
- F1-score
- Accuracy

Exercise 1: Build Logistic Regression model to predict ham or spam for SMS message

Step 5 (cont.):

Evaluate the mode. Decide how good the model is by calculating various metrics

```
array([[1451,    2],
       [  32,  187]])
```

	precision	recall	f1-score	support
ham	0.98	1.00	0.99	1453
spam	0.99	0.85	0.92	219
accuracy			0.98	1672
macro avg	0.98	0.93	0.95	1672
weighted avg	0.98	0.98	0.98	1672

Exercise 1: Build Logistic Regression model to predict ham or spam for SMS message

Step 6:

Save the model and counter vectorizer

- Save count vectorizer so that we can retain the vocabulary list and other setting to get the features. New text will have to be transformed through the count vectorizer
- Save the model for predict the new text
- The model name is: `lr-2022-<MM>-<DD>.pkl`
- The vectorizer name: `countvectoriser-2022-<MM>-<DD>.pkl`

Operational Phase

Steps for operation the model

- Reload the model : Load in the regression model that was saved during the modelling stage
- Reload the Vectorizer: Load in the vectorizer that was used to encode the training set
- Pre-process the new text: Clean the input data in the SAME way as it was done during modelling
- Numerically encode the input : Convert the text to vectors using the previous Vectorizer
- Predict the label : Reuse the model to predict the label

Operational Phase

Steps for operation the model

- Reload the model
- Reload the Count Vectorizer
- Pre-process the new text
- Numerically encode the input
- Predict the label

Exercise 2: Use Logistic Regression model to predict ham or spam for SMS message

Step 1:

Load in the regression model that was saved during the modelling stage

Step 2:

Load in the vectorizer that was used to encode the training set

Exercise 2: Use Logistic Regression model to predict ham or spam for SMS message

Step 3:

Pre-process the new text: Clean the input data in the SAME way as it was done during modelling

- Create function preprocess(text) to clean the data

```
1  new_text="SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, 6days, 16+ \
2  TsandCs apply Reply HL 4 info"
3  new_text=preprocess(new_text)
4  new_text
```

```
'six chances to win cash from  to  pounds txt  and send to  cost day  tsandcs apply reply hl  info'
```

Exercise 2: Use Logistic Regression model to predict ham or spam for SMS message

Step 4:

Numerically encode the input : Convert the text to vectors using the previous Vectorizer

- Create function encode_text_to_vector(cv, text) to encode the new text

```
1   new_text_vector=encode_text_to_vector(trained_cv,new_text)
2   print(new_text_vector)
```

```
(0, 248)      1
(0, 859)      1
(0, 900)      1
(0, 1187)     1
(0, 1346)     1
(0, 2521)     1
(0, 2746)     1
(0, 4335)     1
(0, 4685)     1
(0, 4972)     1
(0, 5971)     1
(0, 5997)     1
(0, 6400)     1
```

Exercise 2: Use Logistic Regression model to predict ham or spam for SMS message

Step 5:

Predict the label

```
predicted_label = (model.predict(new_text_vector))[0]
print ("The new text:\n",new_text)
print("Predicted label is:\n", predicted_label)
```

```
The new text:
 SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, 6days, 16+ TsandCs apply Reply HL 4 info
Predicted label is:
 spam
```

Test other SMS messages

Message	Predicted Label
I'm gonna be home soon and i don't want to talk about this stuff anymore tonight, k? I've cried enough today.	ham
I've been searching for the right words to thank you for this breather. I promise i wont take your help for granted and will fulfil my promise. You have been wonderful and a blessing at all times.	ham
Oh k...i'm watching here:)	ham
Eh u remember how 2 spell his name... Yes i did. He v naughty	ham
Fine if thats□the way u feel. Thats□the way its gota b	ham
Is that seriously how you spell his name?	ham

Test other SMS messages

Message	Predicted Label
SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 a nd send to 87575. Cost 150p/day, 6days, 16+ TsandCs apply Reply HL 4 info	spam
URGENT! You have won a 1 week FREE membership in our £100,00 0 Prize Jackpot! Txt the word: CLAIM to No: 81010 T&C www.dbuk. net LCCLTD POBOX 4403LDNW1A7RW18	spam
XXXMobileMovieClub: To use your credit, click the WAP link in the next txt message or click here>> http://wap. xxxmobilemovieclub. com?n=QJKGIGHJJGCBL	spam
England v Macedonia - dont miss the goals/team news. Txt ur national team to 87077 eg ENGLAND to 87077 Try:WALES, SCOTLAND 4txt/ú1.20 POBOXox365 04W45WQ 16+	spam

Naïve Bayes

Exercise 3: Training dataset

Text	Label
a great game	sports
the election was over	not sports
very clean match	sports
it was a close election	not sports
a clean but forgettable game	Sports
a very close game	??

What is the label for a new text “a very close game”?
Would it be “Sports” or “Not Sports”?

Naïve Bayes Approach

Text	Label
a great game	sports
the election was over	not sports
very clean match	sports
it was a close election	not sports
a clean but forgettable game	sports

Find:

P(sports | “a very close game")

P(not sports | “a very close game”)

- If P(sports | “a very close game") is the larger value, then the label is sports
- If P(not sports | “a very close game”) is the larger value, then the label is not sports

Exercise 4:

Can you implement the **Build Phase and Operational Phase** using Naïve Bayes to predict ham or spam for SMS messages

Hint: Modify the code from "Logistic Regression". A very small changes only!