1

00:00:00,171 --> 00:00:04,190

[MUSIC]

2

00:00:06,708 --> 00:00:10,470

This lecture is about the mixture

of unigram language models.

3

00:00:11,900 --> 00:00:16,280

In this lecture we will continue

discussing probabilistic topic models.

4

00:00:16,280 --> 00:00:20,950

In particular, what we introduce

a mixture of unigram language models.

5

00:00:20,950 --> 00:00:24,230

This is a slide that

you have seen earlier.

6

00:00:24,230 --> 00:00:29,189

Where we talked about how to

get rid of the background

7

00:00:29,189 --> 00:00:34,271

words that we have on top of for

one document.

8

00:00:36,540 --> 00:00:38,440

So if you want to solve the problem,

9

00:00:38,440 --> 00:00:44,090

it would be useful to think about

why we end up having this problem.

10

00:00:44,090 --> 00:00:49,570

Well, this obviously because these

words are very frequent in our data and

11

00:00:49,570 --> 00:00:52,730

we are using a maximum

likelihood to estimate.

12

00:00:52,730 --> 00:00:56,170

Then the estimate obviously would

have to assign high probability for

13

00:00:56,170 --> 00:00:59,284

these words in order to

maximize the likelihood.

14

00:00:59,284 --> 00:01:03,390

So, in order to get rid of them that

would mean we'd have to do something

15

00:01:03,390 --> 00:01:04,030

differently here.

16

00:01:05,740 --> 00:01:09,290

In particular we'll have

to say this distribution

17

00:01:09,290 --> 00:01:12,300

doesn't have to explain all

the words in the tax data.

18

00:01:12,300 --> 00:01:13,620

What were going to say is that,

19

00:01:13,620 --> 00:01:19,760

these common words should not be

explained by this distribution.

20

00:01:19,760 --> 00:01:25,750

So one natural way to solve the problem is

to think about using another distribution

21

00:01:25,750 --> 00:01:29,350

to account for just these common words.

22

00:01:29,350 --> 00:01:33,940

This way, the two distributions can be

mixed together to generate the text data.

23

00:01:33,940 --> 00:01:38,390

And we'll let the other model which

we'll call background topic model

24

00:01:38,390 --> 00:01:40,700

to generate the common words.

25

00:01:40,700 --> 00:01:47,040

This way our target topic theta

here will be only generating

26

00:01:47,040 --> 00:01:51,439

the common handle words that are

characterised the content of the document.

27

00:01:52,880 --> 00:01:54,310

So, how does this work?

28

00:01:54,310 --> 00:01:58,210

Well, it is just a small

modification of the previous setup

29

00:01:58,210 --> 00:02:01,050

where we have just one distribution.

30

00:02:01,050 --> 00:02:02,870

Since we now have two distributions,

31

00:02:02,870 --> 00:02:07,810

we have to decide which distribution

to use when we generate the word.

32

00:02:07,810 --> 00:02:12,670

Each word will still be a sample

from one of the two distributions.

33

00:02:13,730 --> 00:02:16,940

Text data is still

generating the same way.

34

00:02:16,940 --> 00:02:20,770

Namely, look at the generating

of the one word at each time and

35

00:02:20,770 --> 00:02:23,300

eventually we generate a lot of words.

36

00:02:23,300 --> 00:02:24,840

When we generate the word,

37

00:02:24,840 --> 00:02:29,820

however, we're going to first decide

which of the two distributions to use.

38

00:02:29,820 --> 00:02:34,910

And this is controlled by another

probability, the probability of

39

00:02:34,910 --> 00:02:39,639

theta sub d and

the probability of theta sub B here.

40

00:02:41,850 --> 00:02:47,170

So this is a probability of enacting

the topic word of distribution.

41

00:02:47,170 --> 00:02:51,150

This is the probability of

enacting the background word

42

00:02:52,150 --> 00:02:54,500

of distribution denoted by theta sub B.

43

00:02:55,500 --> 00:02:59,890

On this case I just give example

where we can set both to 0.5.

44

00:02:59,890 --> 00:03:03,800

So you're going to basically flip a coin,

a fair coin,

45

00:03:03,800 --> 00:03:05,740

to decide what you want to use.

46

00:03:05,740 --> 00:03:09,850

But in general these probabilities

don't have to be equal.

47

00:03:09,850 --> 00:03:15,590

So you might bias toward using

one topic more than the other.

48

00:03:15,590 --> 00:03:19,960

So now the process of generating a word

would be to first we flip a coin.

49

00:03:19,960 --> 00:03:26,500

Based on these probabilities choosing

each model and if let's say the coin

50

00:03:26,500 --> 00:03:31,920

shows up as head, which means we're going

to use the topic two word distribution.

51

00:03:31,920 --> 00:03:37,620

Then we're going to use this word

distribution to generate a word.

52

00:03:37,620 --> 00:03:40,649

Otherwise we might be

going slow this path.

53

00:03:41,680 --> 00:03:45,530

And we're going to use the background

word distribution to generate a word.

54

00:03:46,910 --> 00:03:51,330

So in such a case,

we have a model that has some uncertainty

55

00:03:51,330 --> 00:03:54,630

associated with the use

of a word distribution.

56

00:03:54,630 --> 00:03:59,420

But we can still think of this as

a model for generating text data.

57

00:03:59,420 --> 00:04:01,220

And such a model is

called a mixture model.

58

00:04:02,760 --> 00:04:03,860

So now let's see.

59

00:04:03,860 --> 00:04:07,020

In this case, what's the probability

of observing a word w?

60

00:04:07,020 --> 00:04:10,460

Now here I showed some words.

61

00:04:10,460 --> 00:04:12,280

like "the" and "text".

62

00:04:12,280 --> 00:04:13,820

So as in all cases,

63

00:04:13,820 --> 00:04:17,910

once we setup a model we are interested

in computing the likelihood function.

64

00:04:17,910 --> 00:04:19,550

The basic question is, so

65

00:04:19,550 --> 00:04:23,040

what's the probability of

observing a specific word here?

66

00:04:23,040 --> 00:04:27,870

Now we know that the word can be observed

from each of the two distributions, so

67

00:04:27,870 --> 00:04:29,840

we have to consider two cases.

68

00:04:29,840 --> 00:04:32,660

Therefore it's a sum over these two cases.

69

00:04:34,410 --> 00:04:40,040

The first case is to use the topic for

the distribution to generate the word.

70

00:04:40,040 --> 00:04:46,150

And in such a case then

the probably would be theta sub d,

71

00:04:46,150 --> 00:04:48,550

which is the probability

of choosing the model

72

00:04:48,550 --> 00:04:53,760

multiplied by the probability of actually

observing the word from that model.

73

00:04:53,760 --> 00:04:56,970

Both events must happen

in order to observe.

74

00:04:56,970 --> 00:05:02,050

We first must have choosing

the topic theta sub d and then,

75

00:05:02,050 --> 00:05:07,650

we also have to actually have sampled

the word the from the distribution.

76

00:05:07,650 --> 00:05:11,100

And similarly,

the second part accounts for

77

00:05:11,100 --> 00:05:13,880

a different way of generally

the word from the background.

78

00:05:15,190 --> 00:05:20,970

Now obviously the probability of

text the same is all similar, right?

79

00:05:20,970 --> 00:05:25,040

So we also can see the two

ways of generating the text.

80

00:05:25,040 --> 00:05:29,720

And in each case, it's a product of the

probability of choosing a particular word

81

00:05:29,720 --> 00:05:34,530

is multiplied by the probability of

observing the word from that distribution.

82

00:05:35,640 --> 00:05:38,890

Now whether you will see,

this is actually a general form.

83

00:05:38,890 --> 00:05:43,940

So might want to make sure that you have

really understood this expression here.

84

00:05:43,940 --> 00:05:48,130

And you should convince yourself that

this is indeed the probability of

85

00:05:48,130 --> 00:05:49,940

obsolete text.

86

00:05:49,940 --> 00:05:52,010

So to summarize what we observed here.

87

00:05:52,010 --> 00:05:57,270

The probability of a word from

a mixture model is a general sum

88

00:05:57,270 --> 00:05:59,500

of different ways of generating the word.

89

00:06:00,610 --> 00:06:01,990

In each case,

90

00:06:01,990 --> 00:06:07,898

it's a product of the probability

of selecting that component model.

91

00:06:07,898 --> 00:06:12,320

Multiplied by the probability of

actually observing the data point

92

00:06:12,320 --> 00:06:14,010

from that component of the model.

93

00:06:14,010 --> 00:06:20,940

And this is something quite general and

you will see this occurring often later.

94

00:06:20,940 --> 00:06:23,825

So the basic idea of a mixture

model is just to retrieve

95

00:06:23,825 --> 00:06:28,820

thesetwo distributions

together as one model.

96

00:06:28,820 --> 00:06:32,810

So I used a box to bring all

these components together.

97

00:06:32,810 --> 00:06:36,200

So if you view this

whole box as one model,

98

00:06:36,200 --> 00:06:38,610

it's just like any other generative model.

99

00:06:38,610 --> 00:06:41,260

It would just give us

the probability of a word.

100

00:06:42,850 --> 00:06:47,310

But the way that determines this

probability is quite the different from

101

00:06:47,310 --> 00:06:48,840

when we have just one distribution.

102

00:06:50,050 --> 00:06:54,710

And this is basically a more

complicated mixture model.

103

00:06:54,710 --> 00:06:57,710

So the more complicated is more

than just one distribution.

104

00:06:57,710 --> 00:06:58,740

And it's called a mixture model.

105

00:07:00,460 --> 00:07:04,450

So as I just said we can treat

this as a generative model.

106

00:07:04,450 --> 00:07:08,450

And it's often useful to think of

just as a likelihood function.

107

00:07:08,450 --> 00:07:10,140

The illustration that

you have seen before,

108

00:07:10,140 --> 00:07:14,210

which is dimmer now, is just

the illustration of this generated model.

109

00:07:14,210 --> 00:07:18,390

So mathematically,

this model is nothing but

110

00:07:18,390 --> 00:07:21,690

to just define the following

generative model.

111

00:07:21,690 --> 00:07:25,820

Where the probability of a word is

assumed to be a sum over two cases

112

00:07:26,840 --> 00:07:28,830

of generating the word.

113

00:07:28,830 --> 00:07:32,800

And the form you are seeing now

is a more general form that

114

00:07:32,800 --> 00:07:36,680

what you have seen in

the calculation earlier.

115

00:07:36,680 --> 00:07:41,150

Well I just use the symbol

w to denote any water but

116

00:07:41,150 --> 00:07:46,330

you can still see this is

basically first a sum.

117

00:07:46,330 --> 00:07:47,560

Right?

118

00:07:47,560 --> 00:07:53,080

And this sum is due to the fact that the

water can be generated in much more ways,

119

00:07:53,080 --> 00:07:55,070

two ways in this case.

120

00:07:55,070 --> 00:08:00,330

And inside a sum,

each term is a product of two terms.

121

00:08:00,330 --> 00:08:05,720

And the two terms are first

the probability of selecting a component

122

00:08:05,720 --> 00:08:07,280

like of D Second,

123

00:08:07,280 --> 00:08:12,730

the probability of actually observing

the word from this component of the model.

124

00:08:12,730 --> 00:08:18,770

So this is a very general description

of all the mixture models.

125

00:08:18,770 --> 00:08:23,020

I just want to make sure

that you understand

126

00:08:23,020 --> 00:08:27,154

this because this is really the basis for

understanding all kinds of on top models.

127

00:08:28,480 --> 00:08:31,350

So now once we setup model.

128

00:08:31,350 --> 00:08:34,310

We can write down that like

functioning as we see here.

129

00:08:34,310 --> 00:08:37,720

The next question is,

how can we estimate the parameter,

130

00:08:37,720 --> 00:08:40,080

or what to do with the parameters.

131

00:08:40,080 --> 00:08:41,540

Given the data.

132

00:08:41,540 --> 00:08:42,860

Well, in general,

133

00:08:42,860 --> 00:08:47,410

we can use some of the text data

to estimate the model parameters.

134

00:08:47,410 --> 00:08:50,470

And this estimation would allow us to

135

00:08:50,470 --> 00:08:55,350

discover the interesting

knowledge about the text.

136

00:08:55,350 --> 00:08:58,450

So you, in this case, what do we discover?

137

00:08:58,450 --> 00:09:01,120

Well, these are presented

by our parameters and

138

00:09:01,120 --> 00:09:03,320

we will have two kinds of parameters.

139

00:09:03,320 --> 00:09:07,400

One is the two worded distributions,

that result in topics, and

140

00:09:07,400 --> 00:09:10,380

the other is the coverage

of each topic in each.

141

00:09:12,560 --> 00:09:14,340

The coverage of each topic.

142

00:09:14,340 --> 00:09:17,630

And this is determined by

probability of C less of D and

143

00:09:17,630 --> 00:09:22,310

probability of theta, so this is to one.

144

00:09:22,310 --> 00:09:25,040

Now, what's interesting is

also to think about special

145

00:09:25,040 --> 00:09:29,540

cases like when we send one of

them to want what would happen?

146

00:09:29,540 --> 00:09:32,770

Well with the other, with the zero right?

147

00:09:32,770 --> 00:09:35,150

And if you look at

the likelihood function,

148

00:09:36,320 --> 00:09:40,640

it will then degenerate to the special

case of just one distribution.

149

00:09:40,640 --> 00:09:46,290

Okay so you can easily verify that by

assuming one of these two is 1.0 and

150

00:09:46,290 --> 00:09:47,940

the other is Zero.

151

00:09:49,130 --> 00:09:53,290

So in this sense,

the mixture model is more general than

152

00:09:53,290 --> 00:09:56,490

the previous model where we

have just one distribution.

153

00:09:56,490 --> 00:09:58,740

It can cover that as a special case.

154

00:09:59,960 --> 00:10:05,340

So to summarize, we talked about the

mixture of two Unigram Language Models and

155

00:10:05,340 --> 00:10:09,110

the data we're considering

here is just One document.

156

00:10:09,110 --> 00:10:13,420

And the model is a mixture

model with two components,

157

00:10:13,420 --> 00:10:16,830

two unigram LM models,

specifically theta sub d,

158

00:10:16,830 --> 00:10:22,810

which is intended to denote the topic of

document d, and theta sub B, which is

159

00:10:22,810 --> 00:10:28,500

representing a background topic that

we can set to attract the common

160

00:10:28,500 --> 00:10:32,840

words because common words would be

assigned a high probability in this model.

161

00:10:33,950 --> 00:10:36,870

So the parameters can

be collectively called

162

00:10:36,870 --> 00:10:40,380

Lambda which I show here you can again

163

00:10:41,560 --> 00:10:45,520

think about the question about how many

parameters are we talking about exactly.

164

00:10:45,520 --> 00:10:50,920

This is usually a good exercise to do

because it allows you to see the model in

165

00:10:50,920 --> 00:10:56,470

depth and to have a complete understanding

of what's going on this model.

166

00:10:56,470 --> 00:10:58,700

And we have mixing weights,

of course, also.

167

00:10:59,790 --> 00:11:02,340

So what does a likelihood

function look like?

168

00:11:02,340 --> 00:11:06,620

Well, it looks very similar

to what we had before.

169

00:11:06,620 --> 00:11:09,100

So for the document,

170

00:11:09,100 --> 00:11:14,260

first it's a product over all the words in

the document exactly the same as before.

171

00:11:14,260 --> 00:11:20,200

The only difference is that inside here

now it's a sum instead of just one.

172

00:11:20,200 --> 00:11:24,420

So you might have recalled before

we just had this one there.

173

00:11:25,420 --> 00:11:30,610

But now we have this sum

because of the mixture model.

174

00:11:30,610 --> 00:11:34,800

And because of the mixture model we

also have to introduce a probability of

175

00:11:34,800 --> 00:11:37,640

choosing that particular

component of distribution.

176

00:11:39,530 --> 00:11:44,470

And so

this is just another way of writing, and

177

00:11:44,470 --> 00:11:49,800

by using a product over all the unique

words in our vocabulary instead of

178

00:11:49,800 --> 00:11:52,878

having that product over all

the positions in the document.

179

00:11:52,878 --> 00:11:57,582

And this form where we look at

the different and unique words is

180

00:11:57,582 --> 00:12:04,720

a commutative that formed for computing

the maximum likelihood estimate later.

181

00:12:04,720 --> 00:12:09,965

And the maximum likelihood estimator is,

as usual,

182

00:12:09,965 --> 00:12:15,290

just to find the parameters that would

maximize the likelihood function.

183

00:12:15,290 --> 00:12:18,940

And the constraints here

are of course two kinds.

184

00:12:18,940 --> 00:12:24,125

One is what are probabilities in each

185

00:12:24,125 --> 00:12:29,142

[INAUDIBLE] must sum to 1 the other is

186

00:12:29,142 --> 00:12:35,343

the choice of each

[INAUDIBLE] must sum to 1.

187

00:12:35,343 --> 00:12:39,799

[MUSIC]