Machine Learning 101

- Got tweets about Ebola and Justin Bieber
- Let's train an algorithm to tell them apart
- If we removed the words “bieber” and “ebola”

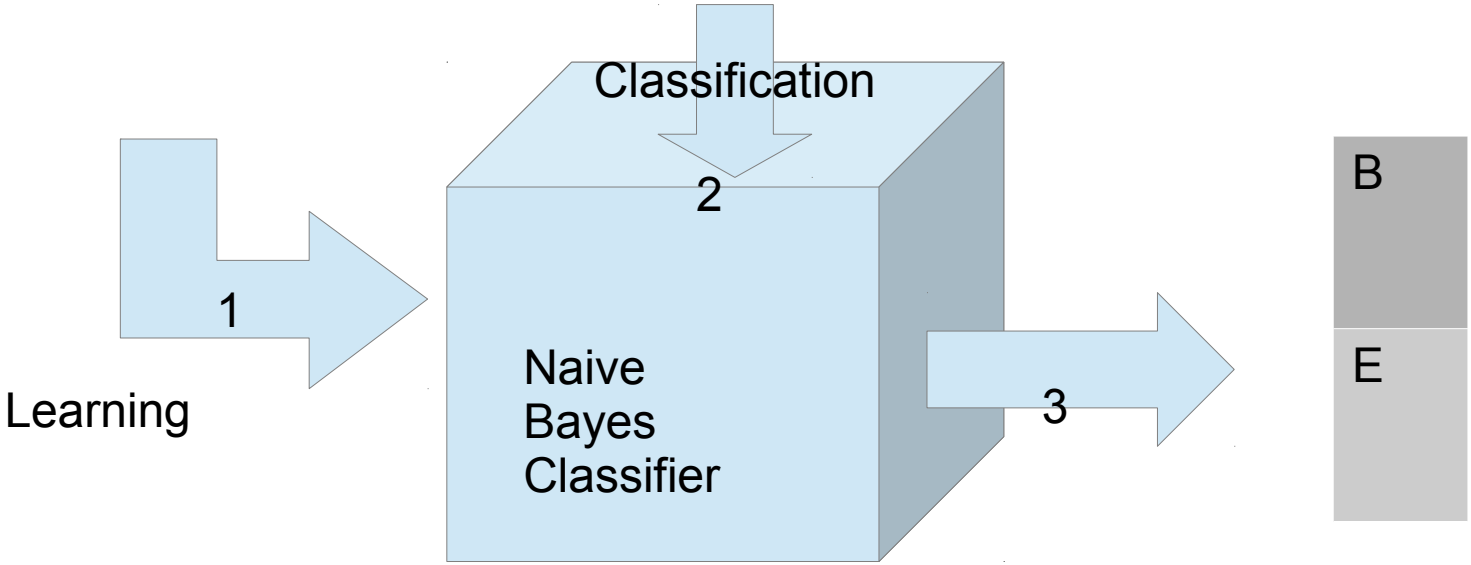
- Statistical model for classification
- The Naive Bayes learning algorithm
- Representing tweets as vectors

Training set

Omg biebs is awesome	B
If you have faith you will kiss bieber	B
Bieber posts naked photos of his girl friend	B
Another case of ebola in dallas, TX	E

New Data

If bieber doesnt come to nyc I will kill myself	?
If ebola kills me dont steal my stereo	?



http://ignignokt.bio.nyu.edu:8080/

Some tweets

Biebs

- Justin Bieber Shares New Pic Of Selena Gomez Kissing His Arm Tattoo: Justin Bieber has just post... http://bit.ly/1s7aRXX ff: gospelgee
- Petition for Justin Bieber to release these three songs from believe movie music #EMABiggestFansJustinBieber
- "if you faith you will kiss Justin Bieber" #EMABiggestFansJustinBieber
- Who is excited about @CodySimpson & @JustinBieber's music project coming out in early 2015?

Ebo

- Dallas Ebola patient's disease condition improving slightly http://ti.me/1nYFSOe
- ebola tweets spreadin quicker than the disease
- Liberia burns its bodies as Ebola fears run rampant http://ti.me/1oNVwXk
- Burning of a human body prevents contagious diseases like Ebola very well. Liberia learning the hard way. Hindus of India knew for ages!

Statistical model

Binary Classification

- P(T is about bieber | **w**)
- P(T is about ebola | **w**) = 1 – P(T is about bieber | **w**)

Statistical model

Binary Classification

- P(T is about bieber | **w**)
- P(T is about ebola | **w**) = 1 – P(T is about bieber | **w**)

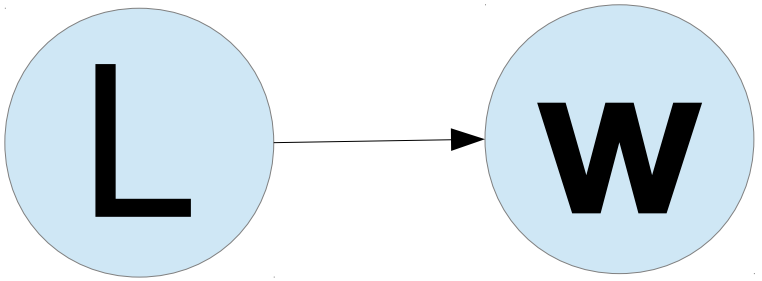
- First I pick the subject, **then** I write the tweet!

Statistical model

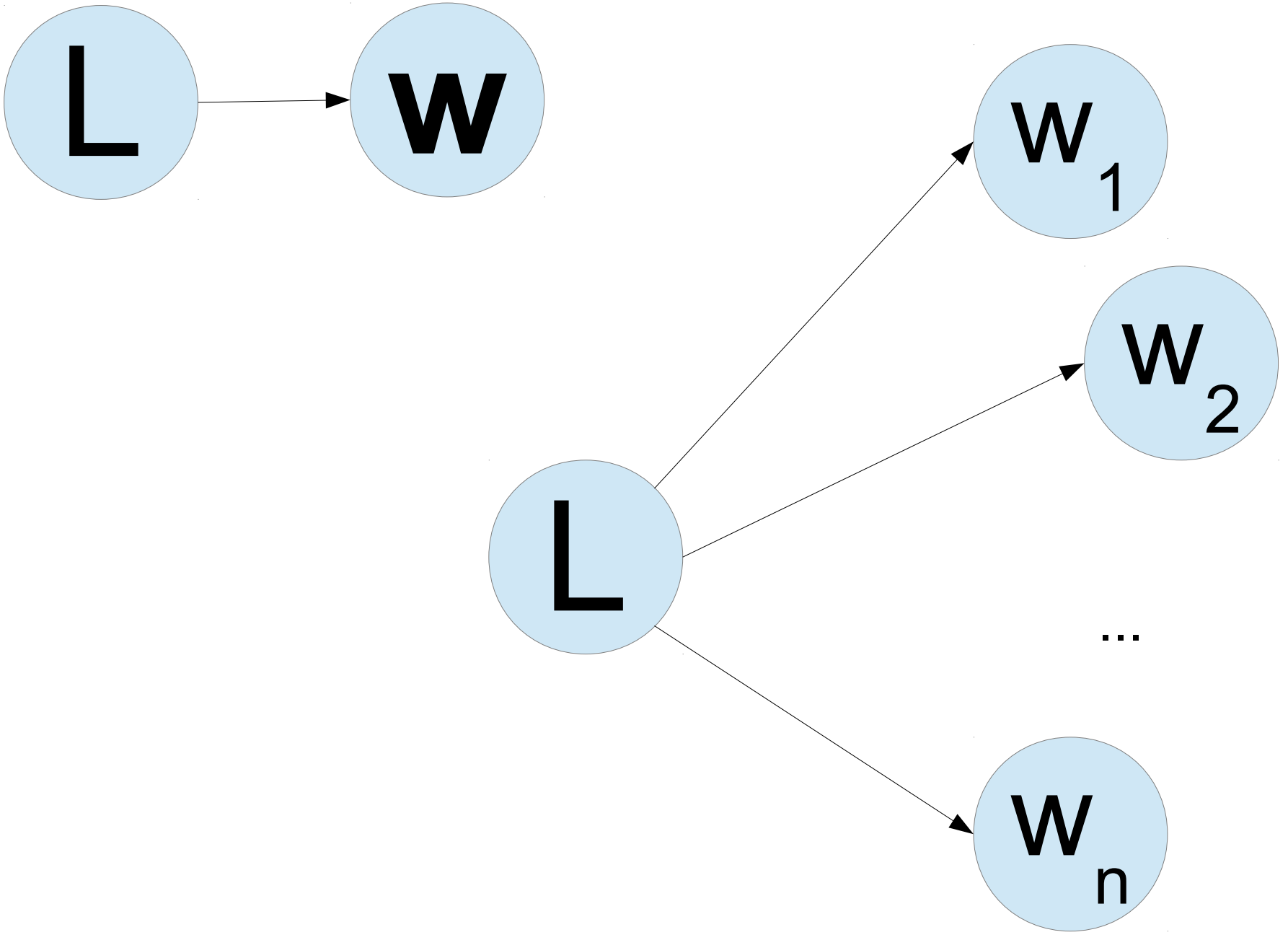
Binary Classification

- P(T is about bieber | **w**)
- P(T is about ebola | **w**) = 1 – P(T is about bieber | **w**)

- First I pick the subject, **then** I write the tweet!



Statistical model



Statistical model

$$P(L|\boldsymbol{w}) = \frac{P(L)\,P(\boldsymbol{w}|L)}{P(\boldsymbol{w})}$$

(Bayes Rule)

Statistical model

$$P(L|\boldsymbol{w}) = \frac{P(L)\,P(\boldsymbol{w}|L)}{P(\boldsymbol{w})}$$ (Bayes Rule)

$$P(L|\boldsymbol{w}) \propto P(L)\,P(\boldsymbol{w}|L)$$

Statistical model

$$P(L|\boldsymbol{w}) \propto P(L)\,P(\boldsymbol{w}|L)$$
$$= P(L)\,P(w_1, w_2, \ldots w_n|L)$$

The Naive Assumption

- Words are chosen independently and interchangeably in order

$$P(w_i|L, w_j) = P(w_i|L)$$

The naive statistical model

$$P(L|\boldsymbol{w}) \propto P(L)\, P(w_1|L)\, P(w_2|w_1, L) \cdots P(w_n|w_1, \ldots, w_{n-1}, L)$$

+

$$P(w_i|L, w_j) = P(w_i|L)$$

$$P(L|\boldsymbol{w}) \propto P(L) \prod_{i=1}^{n} P(w_i|L)$$

The Naive Bayes Classifier

Classification with Naive Bayes

$$P(L|\boldsymbol{w}) \propto P(L) \prod_{i=1}^{n} P(w_i|L)$$

- Use this formula to get

  P(L=biebs | **w**=”if you are a belieber RT this!”)

  P(L=ebola | **w**=”if you are a belieber RT this!”)

  When we observe a new tweet **w**

- **Then, pick the L which maximizes that**

$$\hat{l} = \underset{c}{argmax}\, P(L = c|\boldsymbol{w})$$

Learning Naive Bayes

$$P(L|\boldsymbol{w}) \propto P(L) \prod_{i=1}^{n} P(w_i|L)$$

The prior probability of class L

The likelihood of word i given class L

Learning Naive Bayes

$$P(L|\boldsymbol{w}) \propto P(L) \prod_{i=1}^{n} P(w_i|L)$$

$$\hat{P}(L = biebs) = \frac{N_{biebs}}{N}$$

$$\hat{P}(L = ebola) = 1 - P(L = biebs)$$

Learning Naive Bayes

$$P(L|\boldsymbol{w}) \propto P(L) \prod_{i=1}^{n} P(w_i|L)$$

$$\hat{P}(w_i = sexy | L = biebs) = \frac{N_{sexy_{biebs}}}{N_{biebs}}$$

One such parameter for each word for each class

Learning Naive Bayes

$$\hat{P}(w_i = sexy | L = biebs) = \frac{N_{sexy_{biebs}}}{N_{biebs}}$$

- What if the word “sexy” doesn't appear in tweets about ebola?
  - “Ebola is one sexy disease” is still not about bieber
  - $\hat{P}(w_i = sexy | L = biebs) = \frac{0}{V} = 0$

Learning Naive Bayes

$$\hat{P}(w_i = sexy | L = biebs) = \frac{N_{sexy_{biebs}}}{N_{biebs}}$$

- What if the word “sexy” doesn't appear in tweets about ebola?
  - “Ebola is one sexy disease” is still not about bieber
  - $\hat{P}(w_i = sexy | L = biebs) = \frac{0}{V} = 0$

$$P(L|\boldsymbol{w}) \propto P(L) \prod_{i=1}^{n} P(w_i|L) = 0$$

Laplace smoothing

- Add “pseudocounts” to never-seen words

$$\hat{P}\left(w_i = sexy \middle| L = biebs\right) = \frac{N_{sexy_{biebs}} + \alpha}{N_{biebs} + \alpha V}$$

- V = vocabulary size

  (number of words in your dictionaty)

Bag of words representation

- *Justin Bieber Shares New Pic Of Selena Gomez Kissing His Arm Tattoo: http://bit.ly/1s7aRXX ff: gospelgee*
- *"if you have faith you will kiss Justin Bieber" #EMABiggestFansJustinBieber*

Dictionary:

0	1	2	3	4
share	pic	kiss	tattoo	faith

- 1st tweet: [1, 1, 1, 1, 0]
- 2nd tweet:[0, 0, 1, 0, 1]

Bag of words representation

How do choose the dictionary

- Choose words from our “corpus” of data

- Remove “stop words”
- Remove very common words
- Remove very rare words

Classifying

- Get a “training set” of ebola/bieber tweets
- Learn the Naive Bayes parameters from those
- Use to predict labels on tweets when we don't know in advance what they are about
  - This is called the “test set”
  - Removed the words “ebola” or “justin” or “bieber” to make more interesting

Training set

Omg biebs is awesome	B
If you have faith you will kiss bieber	B
Bieber posts naked photos of his girl friend	B
Another case of ebola in dallas, TX	E

New Data

If bieber doesnt come to nyc I will kill myself	?
If ebola kills me dont steal my stereo	?