Text classification - Topic modelling

- Naive Bayes classifier - multiset model (bag of words)

  $P(c)$ - category c

  $P(w_i|c_j)$ $w_i \in V$, $c_j \in C$

  └ $P(c|d)$

- Semisupervised Approach

  - Label a small fraction of D
  - Estimate $P(c)$, $P(w_i|c_j)$

- Assign fractional $P(c_j|d)$ to each unlabelled d
- Recompute $P(c)$, $P(w_j|c_i)$ using fractional sums

Iterate - label unlabelled d again ---

Expectation Maximization.

- Unsupervised topic labelling

Generative model

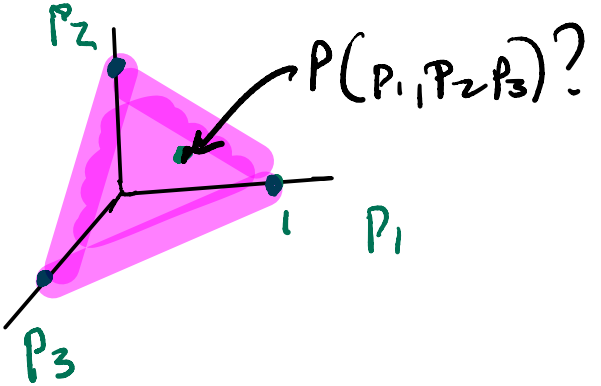
- Pick c with $P(c)$
- Generate n words with $P(w_i|c)$

$P(c_i) = p_i$ $\quad \sum p_i = 1$

k topics $(p_1, p_2, \ldots, p_k)$ $\quad$ Prob $((p_1, p_2, \ldots, p_k))$

Dirichlet distribution

Parameter α: $D(k, \alpha)$



Generative model

Pick $(p_1, \ldots p_k)$ from $D(k, \alpha)$

Assign random d, this mixed topic $\quad$ 1/3 sports, 2/3, food

Generate n words — pick c_i for this mood

choose w_j with $P(w_j | c_i)$

Work backwards.

Given $D = \{d_1, \ldots, d_n\}$ & k

Assign $c_1, \ldots, c_k$ to each $d \in D$ as best we can

Example: $D = \{d_1, d_2, d_3, d_4, d_5\}$ $k = 2$

- I like to eat broccoli and bananas.
- I ate a banana and spinach smoothie for breakfast.
- Chinchillas and kittens are cute.
- My sister adopted a kitten yesterday.
- Look at this cute hamster munching on a piece of broccoli.

] Input

- **Sentences 1 and 2**: 100% Topic A
- **Sentences 3 and 4**: 100% Topic B
- **Sentence 5**: 60% Topic A, 40% Topic B
- **Topic A**: 30% broccoli, 15% bananas, 10% breakfast, 10% munching, … (at which point, you could interpret topic A to be about food)
- **Topic B**: 20% chinchillas, 20% kittens, 20% cute, 15% hamster, … (at which point, you could interpret topic B to be about cute animals)

Goal

Start with an initial estimate (p_1, p_2) for each sentence

- Assign a topic to each word
- Fractional counts — $P(c_j|d)$
  $P(w_i|c_j)$ ←

c_1 c_2 c_m

w_1 w_2 ... w_m is a document

"Resample"

GIBB'S SAMPLING

$P(c_j|w_i)$

$c_1 \rightarrow c_1'$ → Recompute all $P(c_j|d)$, $P(w_i|c_j)$

Move to w_2, do the same

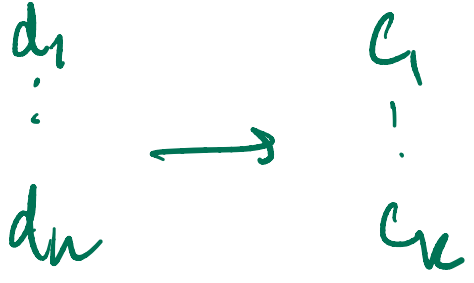
Repeat until counts stabilize

Latent Dirichlet Allocation (LDA)

Unsupervised

Find the right k

Evaluating k_1 vs k_2



Example

Customer complaints

What are they complaining about?

Unsupervised preprocessing → Relevant interpretation

LDA