Topic Models

Review: The generative model for Naive Bayes

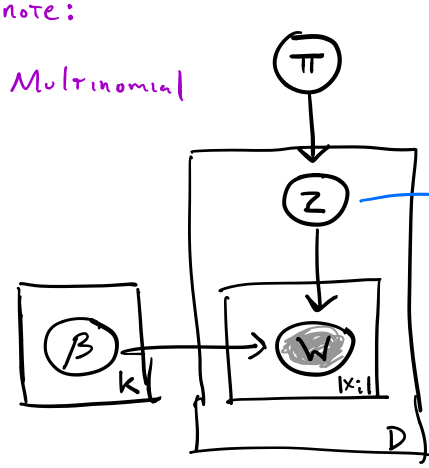
$X = \{x_1, \dots x_D\}$ $\quad$ K labels/classes

$z_d \sim \text{Categorical}^{*}(\pi)$

$x_{dn} \mid z_d \sim \text{Categorical}(\beta_{z_d})$

Probability over words w given class z_d.

☆ Terminology note:
"Categorical" = Multinomial w/ 1 trial.



one label ("Topic") per doc.

Assumes one topic z_d per document or instance.

Most documents will span multiple topics.

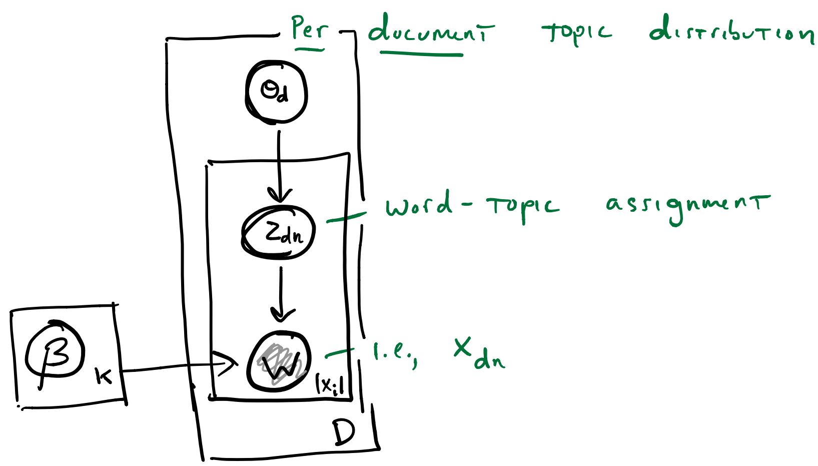
To respect this, we need a new model.

$$z_{d\underline{n}} \sim \text{Categorical}(\theta_d)$$

$$x_{dn} \mid z_{d\underline{n}} \sim \text{Categorical}(\beta_{z_{dn}})$$

distribution over words for topic z_{dn}.

Graphically



EM for Topic Models: PLSA

(Probabistic Latent Semantic Analysis)

$$z_{dn} \sim \text{Categorical}(\theta_d)$$

$$x_{dn} \mid z_{dn} \sim \text{Categorical}(\beta_{z_{dn}})$$

distribution over words for topic z_{dn}.

E-step

Update soft assignments

$$\phi_{dnk} = \frac{\theta_{dk}\,\beta_{kv}}{\sum_l \theta_{dl}\,\beta_{lv}} \qquad v \overset{\text{def}}{=} x_{dn}$$

P that word n in document d was drawn from topic k

M-step

$$\beta_{kv} = \frac{N_{kv}}{\sum_w N_{kw}} \qquad N_{kv} \overset{\text{def}}{=} \sum_{d,n} \phi_{dnk}\, x_{nv}$$

= expected count of v in topic k

$$\theta_{dk} = \frac{N_{dk}}{\sum_l N_{dl}} \qquad N_l \overset{\text{def}}{=} \sum_n \phi_{dnl}$$

= Expected number of words from topic l in d