* LDA & Topic Modeling : (Continue)

↳ Latent dirichlet Allocation → generative model.

- In ML it is used for the task of Topic Modeling. Topic Modeling is clustering of doc. into topics which are latent.

ML → discriminative $P(Y|X)$; generative $P(x, y)$

→ Some Preliminaries

Multinomial distⁿ

- Generalization of binomial distⁿ

- Suppose we have vector X in R^k such that only one coordinate is 1 & rest all 0. ↳ $\{x : x \in \{0,1\}$ & $\|x\|_1 = 1\}$
- If we denote $P(x_k = 1) = \mu_k$ distribution of X is given as

$$P(X|\mu) = \prod_{k=1}^{K} \mu_k^{x_k}$$

↳ k dimensional vector. (represent 1st one prob(1), 2nd prob(2) ...)

where $\mu = \{\mu_1 \dots \mu_k\}$

$\mu_k \geq 0$ &

$$\sum_k \mu_k = 1$$

ex $X_A = 0, 0, 1$

$X_B = 0, 1, 0$

$X_1 = 1, 0, 0$

$\mu_1 = 0.7$

$\mu_2 = 0.2$

$\mu_3 = 0.1$

$P(X_A|\mu) = (0.7)^0 (0.2)^0 (0.1)^1$

→ Now consider denote D of N independent observation $x_1 \dots x_N$

Likelihood

$$P(D|\mu) = \prod_{i=1}^{N} \prod_{k=1}^{K} \mu_k^{x_{nk}}$$

$$= \prod_{k=1}^{K} \mu_k^{(\sum_n x_{nk})}$$

↳ $a^x \cdot a^y = a^{x+y}$

$$= \prod_{k=1}^{K} \mu_k^{m_k}$$

$(\because m_k = \sum_n x_{nk})$

(no. of observation of $x_k = 1$)

* Now max^m likelihood sol^n for μ is

$$\mu_k^{ML} = \frac{m_k}{N}$$

→ we consider joint dist^n of quantities $m_1 \cdots m_K$ conditioned on μ & N.

$$\text{Multinomial dist}^n(m_1, m_2 \cdots m_K \mid \mu, N) = \binom{N}{m_1, m_2 \cdots m_K} \prod_{k=1}^{K} \mu_k^{m_k}$$

↳ Ex $\quad \text{mult}(10, 20 \cdots 50 \mid \mu_i)$

where $\binom{N}{m_1, m_2 \cdots m_K} = \frac{N!}{m_1!\, m_2! \cdots m_K!}$

& $N = \sum_{k=1}^{K} m_k = N$

- this is multinomial dist^n

* Dirichlet dist^n:-

Conjugate prior for multinomial dist^n:

$p(\mu \mid \alpha)$ varies as $\prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$

where $0 \le \mu_k \le 1$

$\sum_k \mu_k = 1$

Gaussian dist^n is conjugate prior for itself

dirichlet is conjugate prior for multinomial

$\alpha_1 \cdots \alpha_K$ are the parameters of dist^n.

$\alpha = (\alpha_1 \cdots \alpha_K)$

↳ (prob. dist^n over prob. dist^n)

Ex (10 coins ~~flipped~~ flipping 0.7 0.3 ... 0.1 & prob. of head & we need find what is u prob. of head is given prob. dist^n)

* Dirichlet list^{n} is given. as

$$Dir(\mu|\alpha) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{i=1}^{K} \mu_k^{\alpha_k - 1}$$

$$\alpha_0 = \sum_{k=1}^{K} \alpha_k$$

$$\Gamma(x) = \int_0^\infty u^{x-1} e^{-u}\, du$$

(extension of factorial fun^{n} for reals)

$$\Gamma(x+1) = x\,\Gamma(x)$$

$$\Gamma(1) = 1$$

Multinomial → vector (0 1 0 0 ...)

* Now, $P(\mu|D,\alpha)$ ← alpha

$\propto$ (Ratio symbol) (Proportional)

$P(D|\mu)\, P(\mu|\alpha)$

Posterior also dirichlet

$$\left(\prod_{k=1}^{K} \mu_k^{m_k}\right) \left(\prod_{k=1}^{K} \mu_k^{\alpha_k - 1}\right)$$

$$\propto \prod_{k=1}^{K} \mu_k^{\alpha_k + m_k - 1}$$

So $\left(a^{bx} \cdot a^{y} : a^{x+y}\right)$

Thus $P(\mu|D,\alpha) = Dir(\mu|\alpha+m)$

$$= \frac{\Gamma(\alpha_0 + N)}{\Gamma(\alpha_1+m_1)\cdots\Gamma(\alpha_n+m_n)}$$

(pause if any error)

~~+ lecture in DSA books~~

* lecture in DSA book

the posterior prob. $p(k|x)$ will given as

$$\gamma_k(n) : p(k|x)$$

$$= \frac{p(k)\,p(x|k)}{\sum_l p(l)\,p(x|l)}$$

$$= \frac{w_k\, N(x|\mu_k, \Sigma_k)}{\sum_l w_l\, N(x|\mu_l, \Sigma_l)}$$

Parameters of GMM will w, μ, Σ

$w = (w_1 \dots w_k)$

$\mu = (\mu_1 \dots \mu_k)$

$\Sigma = [\Sigma_1 \dots \Sigma_k)$

* How to get values??

Log of likelihood is given by $\ln p(X|w, \mu, \Sigma)$

$$= \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} w_k\, N(x_n|\mu_k, \Sigma_k) \right\}$$

* Another interpretation :

z is K-dim. binary random var. where at a time some z_k is 1 & all other coordinates will 0.

$\therefore z_k \in \{0,1\}$ and $\sum_k z_k = 1$

- K possible states of var. z
- Try to define joint dist. $p(x, z)$ in terms of $p(z)$ & $p(x|z)$

$\longrightarrow$ Now, $p(z_k = 1) = w_k$ $\quad [\because 0 \le w_k \le 1$ & $\sum_k w_k = 1]$

$\longrightarrow$ $p(z) = \prod_{k=1}^{K} w_k^{z_k}$

(∵ here there is assumption that every z in z is independent so it's Multiplication of one of the vector is value is 1 & others will 0's.) gaussian dist" just like one hot encoding (1 0 0 ⋯)

Also $p(x|z_k = 1) = N(x|\mu_k, \Sigma_k)$

$\therefore p(x|z) = \prod_{k=1}^{K} \left(N(x|\mu_k, \Sigma_k)\right)^{z_k}$

Now $P(X) = \sum_z P(z) P(X|z)$ (Marginal Prob.)

$= \sum_{k=1}^{K} \omega_k \cancel{P(X)} N(x|\mu_k, \Sigma_k)$

∴ Marginal of X is GMM if we have observations $x_1 \dots x_N$ and as $P(X) = \sum_z P(X, z)$ for every observed data point x_n there is corresponding latent var. z_n.

(Means which come from GMM which is hidden)

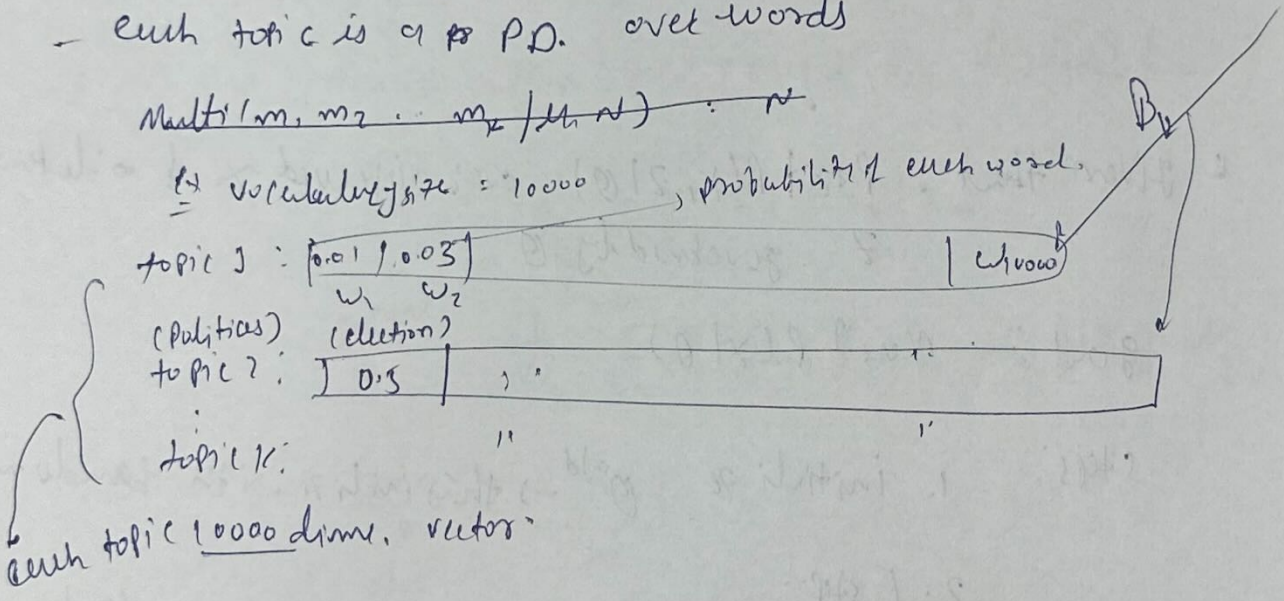
Now, $\gamma(z_k)$ i.e. P
(responsibility)

→ Maximizing log likelihood of GMM is more complex problem than the case of single Gaussian, because of summation over K inside log, log fn" does act directly on Gaussian. wl

(Matrix cookbook vlb)

* Topic Modeling & LDA

- given a corpus of doc. we consider that each word in each doc. has latent assignment into one of K topics. LDA is standard algo. for the task.
- each doc. is mixture of topics.
- each topic is a PD. over words



* High level idea of Generative model

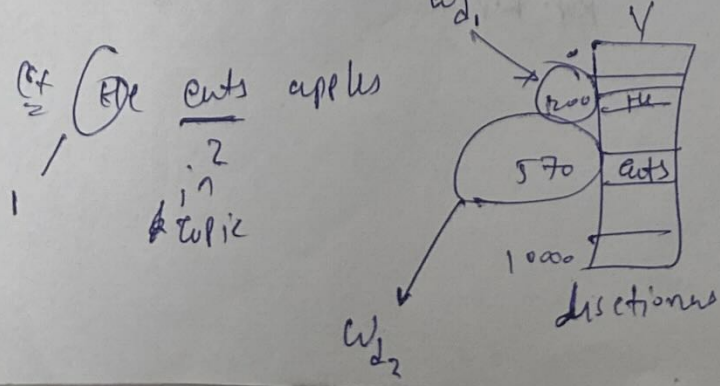
- each doc. is generate by choosing topic
- each

→ D- documents V = Vocabulary / dictionary

. we represent a doc. d with a vector $w_d \in N^{n_d}$ where d^{n_d} is no. of terms doc.

$w_d = (w_{d_1}, w_{d_2} \dots w_{d_{n_d}})$ where

w_{d_i} stands for index of word in position i of doc. in the V.

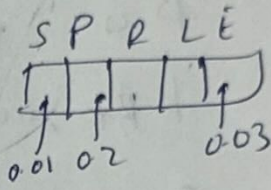


- Let K - num. of latent topics in corpus
- θ_d a vector K-dim. simplex Δ^{K-1} $\quad \theta_{dk} \in [0,1]$

Shows topic k in doc. d.

$\beta_k \in \Delta^{V-1}$ vector of term possibilities in topic k.

$\therefore \beta_{kv}$ is the prob. of observing word v in topic k

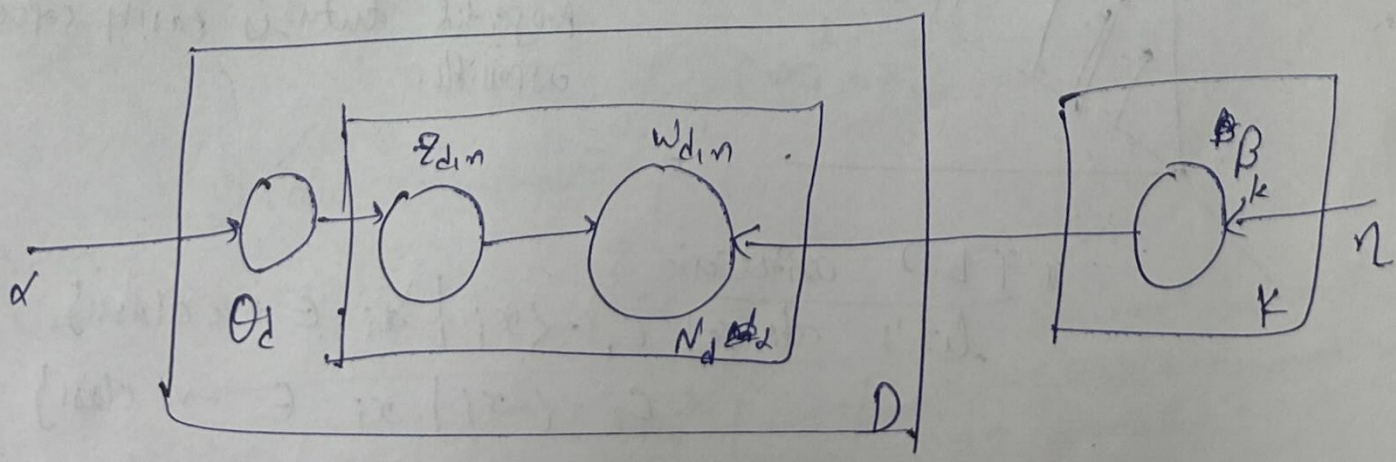
θ_d = 

↳ probability of choosing k^{th} topic for generating word

Consider the latent var. $z_{d,n} \in (1 \ldots K)$ is topic assigned to n^{th} word in document d. (pertains doc. what is proportion of topic)

Algo.

1. Draw K vectors $\beta_k \in \Delta^{V-1}$ from $Dir(\beta_k \mid \eta)$
2. Draw D vectors $\theta_d \in \Delta^{K-1}$ from $Dir(\theta_d \mid \alpha)$
3. Each word $w_{d,n}$ in doc d is generated in two steps.
   i. Draw $z_{n,d} \in (1, \ldots K)$ according to topic prob. θ_d.
   ii. Draw $w_{d,n}$ using term prob. $\beta_{z_{n,d}}$



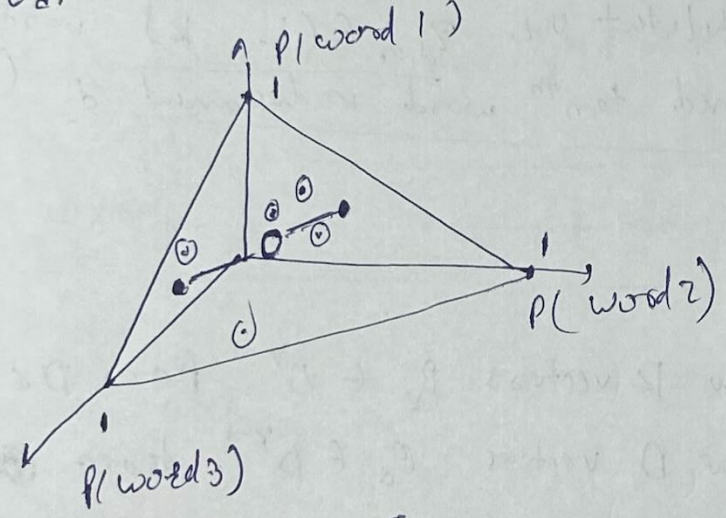
- the lattice for observed corpus $w_{d,n}$ given η, α:

$$P(z, \theta, \beta \mid w, \alpha, \eta) = \left[\prod_{d,n} P(w_{d,n} \mid z_{d,n}, \beta)\right]\left(\prod_{d,n} P(z_{d,n} \mid \theta_d\right)$$

$$= \left[\prod_{d,n} P(w_{d,n} \mid z_{d,n}, \beta) \right] \left[\prod_{d,n} P(z_{d,n} \mid \theta_d) \right]$$

$$\cdot \left[\prod_{d} P(\theta_d \mid \alpha) \right] \left[\prod_{k} P(\beta_k \mid \eta) \right]$$

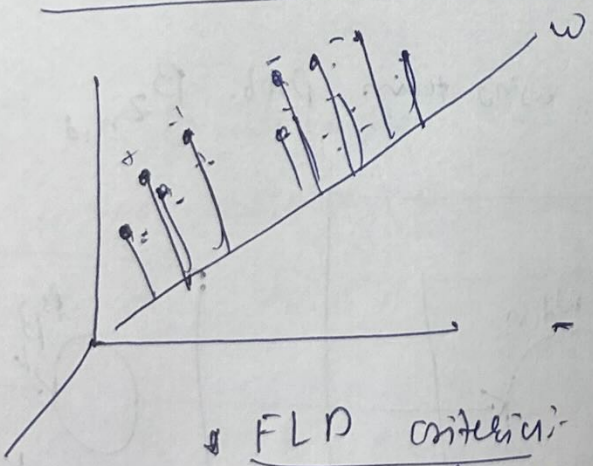
- β is $V \times K$ matrix of term prob. with column k given to β_k
- θ is $K \times D$ matrix of topic membership with column given to θ_d.



● - topic
◉ - Observed topic
○ - generated doc.

many $w^T x$

* Fischer's linear discriminante :-



Aim :- get a w for which projected data is easily seperable as possible.

* FLD criteria :-

Let's set $C_1 = \{x_i \mid x_i \in +ve \text{ class}\}$

$C_2 = \{x_i \mid x_i \in -ve \text{ class}\}$

$|C_1| = n_1, \quad |C_2| = n_2$

$n = n_1 + n_2$

$m_1 = \frac{1}{n_1} \sum_{x_i \in C_1} x_i$, $m_2 = \frac{1}{n_2} \sum_{x_i \in C_2} x_i$

vectors of those classes with d dim Point $\in R^d$

Sign of Perpendicular $\left(\vec{x_i^{\perp}}\right) = x_i^T w \quad \forall i$

(Means projected x_i in w) which is scalar

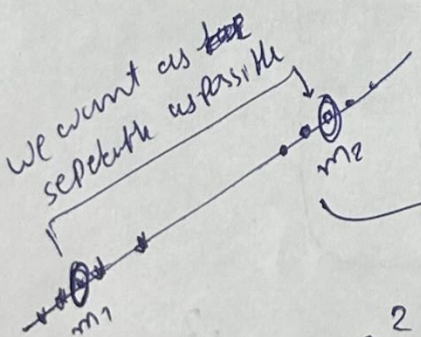
→ if $\frac{1}{n_1} \sum_{x_i \in C_1} x_i^{\perp} = \frac{1}{n_1} \sum_{x_i \in C_1} x_i^T w$

$= \left(\frac{1}{n_1} \sum_{x_i \in C_1} x_i^T\right) w$

$= m_1^T w$

$= m_1^{\perp}$

* similarly $m_2 = m_2^{\perp} = m_2^T w$



→ $|m_1^{\perp} - m_2^{\perp}| = |w^T (m_1 - m_2)|$

$S_1^2 = \sum_{x_i \in C_1} (m_1^{\perp} - x_i^{\perp})^2$

(class 1 variance)

Scatter of Class $C_1 \& C_2$ → $S_2^2 = \sum_{x_i \in C_2} (m_2^{\perp} - x_i^{\perp})^2$

→ we want $(m_1^{\perp} - m_2^{\perp})^2$ as large & denominator is small.

$$J(w) = \frac{|m_1^{\perp} - m_2^{\perp}|^2}{S_1^2 + S_2^2}$$

scalar

one of the 2 is larger

↳ known as FLD (criterion)

* $\tilde{m}^+ = w^T m^+ \rightarrow m_1$

$\tilde{m}^- = w^T m^- \rightarrow m_2$

$\therefore$ maximize $|\tilde{m}^+ - \tilde{m}^-|$

$= |w^T (m^+ - m^-)|$

$\therefore |\tilde{m}^+ - \tilde{m}^-|^2 = [w^T (m^+ - m^-)]^2$

$= w^T (m^+ - m^-) . w^T (m^+ - m^-)$

$= w^T \underbrace{(m^+ - m^-)(m^+ - m^-)^T} w$

S_B (Rank 1)

↳ between class scatter (Inter class) matrix

$\therefore |\tilde{m}^+ - \tilde{m}^-|^2 = w^T S_B w$

* $\tilde{s}^{+2} = \sum_{x_i \in C^+} (\tilde{x}_i - \tilde{m}^+)^2$

$\tilde{s}^{-2} = \sum_{x_i \in C^-} (\tilde{x}_i - \tilde{m}^-)^2$

Scatter for projected points. for +ve class

$\tilde{s}^{+2} = \sum_{x_i \in C^+} (w^T (x_i - m^+))^2$

$= \sum_{x_i \in C^+} \{ w^T (x_i - m^+)(x_i - m^+)^T w \}$

$= w^T \{ \sum_{x_i \in C^+} (x_i - m^+)(x_i - m^+)^T \} w$

Scatter matrix for +ve class.

Summation of rank 1 matrix

$\tilde{s}^{+2} = w^T S^+ w$

$$\tilde{s}^{-2} = \omega^T s^- \omega$$

$$\tilde{s}^{+2} + \tilde{s}^{-2} = \omega^T (s^+ + s^-) \omega$$

$$= \omega^T S_w \omega \quad \Rightarrow \{ \because S_w = s^+ + s^- \}$$

(S_w ↑ within class scatter class $s^+ + s^-$ (intra))

* FLD Criteria :-

$$J(\omega) = \frac{|\tilde{m}^+ - \tilde{m}^-|^2}{\tilde{s}^{+2} + \tilde{s}^{-2}} = \frac{\omega^T S_B \omega}{\omega^T S_w \omega}$$

(S_B, S_w all known)

$$\omega_{fisher} = \underset{\omega}{\operatorname{argmax}}\ J(\omega)$$

$J(\omega) : R^d \to R$

fisher hyperplan.

* Maximum the $J(\omega)$

$$\frac{\partial J(\omega)}{\partial \omega} = \frac{(\omega^T S_w \omega)(2 S_B \omega) - (\omega^T S_B \omega)(2 S_w \omega)}{(\omega^T S_w \omega)^2}$$

$$= 0 = \frac{2 [(\omega^T S_w \omega) S_B \omega - (\omega^T S_B \omega) S_w \omega]}{(\omega^T S_w \omega)^2}$$

$$= \frac{(\omega^T S_w \omega)(S_B \omega) - (\omega^T S_B \omega)(S_w \omega)}{(\omega^T S_w \omega)^2}$$

$$\left. \frac{\partial J(\omega)}{\partial \omega} \right|_{\omega = \omega^*} = 0$$

$$(w^{*T} S_w w^*)(S_B w^*) = (w^{*T} S_B w^*)(S_w w^*)$$

$$S_B w^* = \frac{(w^{*T} S_B w^*)(S_w w^*)}{(w^{*T} S_w w^*)}$$

$$S_B w^* = J(w^*)(S_w w^*)$$

$$= AV = \lambda BV$$

$$AV = \lambda IV$$

Generalized Eigen value Problem

$$S_B w^* = (m^+ - m^-)\underbrace{(m^+ - m^-)^T w^*}_{\text{Scalar}}$$

↳ is some vector along with direction $(m^+ - m^-)$

$$\therefore S_B w^* = J(w^*).(S_w w^*)$$

~~$J(w^*) S_w w^* = \alpha(m^+ - m^-)$~~

$$J(w^*) S_w w^* = (m^+ - m^-)\underbrace{(m^+ - m^-)^T w^*}$$

$$S_w w^* = \alpha(m^+ - m^-)$$

$$\left(\because \alpha = \frac{(m^+ - m^-)^T w^*}{J(w^*)}\right)$$

$$\therefore w^* \propto S_w^{-1}(m^+ - m^-)$$

It is invertible

direction of w^*