Modelling

→ Latent Dirichlet Allocation is generative model
↳ used for topic modelling

Topic modeling is clustering of documents into topics which are latent.

- Some preliminaries:-

- multinomial distribution

- generalization of binomial distribution
- suppose we have vector x into R^K such that only one coord. is 1 rest are 0

- If we denote $P(x_k = 1) = \mu_k$ distribution of x is given as

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

where

$\mu = (\mu_1, \ldots \mu_K)$ $\quad \sum_k \mu_k = 1$

$\mu_k \geq 0$

- now consider dataset D of N independent observations $x_1, \ldots x_N$

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

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

$= \prod_{k=1}^{K} \mu_k^{m_k} \quad \left[m_k = \sum_n x_{nk} \right]$

[no of observations of $x_k = 1$]

→ Now maximum likelihood solution for μ is

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

- we consider joint dist. of quantities $m_1, \ldots m_K$ conditioned on μ & N

$\text{mult}(m_1, \ldots m_K / \mu, N)$

$= \binom{N}{m_1, \ldots, m_K} \prod_{k=1}^{K} \mu_k^{m_k}$

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

$\sum_{K=1}^{k} m_K = N$ (mult. dist.)

- **Dirichlet Distribution:-**

- Conjugate prior for multinomial distribution

$p(\mu|\alpha)$ varies as $\prod_{K=1}^{k} \mu_K^{\alpha_K - 1}$

where $0 \leq \mu_K \leq 1$

$$\sum_{k} \mu_K = 1$$

$\alpha_1, \ldots \alpha_K$ are the parameters of distribution.

$\alpha = (\alpha_1, \ldots \alpha_K)$

The dirichlet dist. is given as

$$Dir(\mu|\alpha) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_K)} \cdot \prod_{i=1}^{k} \mu_K^{\mu_K - 1}$$

$$\alpha_0 = \sum_{K=1}^{K} \alpha_K$$

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

(extension of factorial func for real

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

Now Γ alpha

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

$\propto \quad P(D|\mu) \cdot P(\mu|\alpha)$

(varies symbol)

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

Thus

$$P(\mu|D,\alpha) = Dir(\mu|\alpha+m) = \frac{\Gamma(\alpha_0+N)}{\Gamma(\alpha_1+m_1)\ldots\Gamma(\alpha_K+m_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k+m_k}$$

$m = (m_1, \ldots m_K)$

- Topic Modelling & LDA

- Given a corpus of do... we consider that each wor...

Ch-1 SUMS

Page No.
Date

① 2 opaque bag → each containing 2 balls
one bag → 2 black balls
other → a black & a white

pick a bag at random & pick one of the balls

when you look at the balls it's black & picked second ball from same bag
prob that ball also black?

→ B1: picked ball black
B2: second picked ball black.

$$P(B_2/B_1) = \frac{P(B_2 \cap B_1)}{P(B_1)}$$

$$P(B_1) = P(B_1 | \text{Bag 1})\, P(\text{Bag 1}) + P(B_1/\text{Bag 2})\, P(\text{Bag 2})$$

$$= 1 \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2}$$

$$= \frac{3}{4}$$

→ $P(B_1 \cap B_2)$
$= P(B_1 \cap B_2/\text{Bag 1})\, P(\text{Bag 1}) + P(B_1 \cap B_2/\text{Bag 2})\, P(\text{Bag 2})$

$$1 \cdot \frac{1}{2} + 0 \cdot \frac{1}{2} = \frac{1}{2}$$

$$P(B2/B1) = \frac{1/2}{3/4} = \frac{2}{3}$$

(2) consider perceptron in 2D $h(x) = \text{sign}(w^T x)$ where $w = [w_0\ w_1\ w_2]^T$ and $x = [1\ x_1\ x_2]^T$

(a) show that regions on plane where $h(x) = +1$, $h(x) = -1$ are separated by a line
if we express this line by $x_2 = a x_1 + b$
what are slope 'a' & intercept b in terms of w_0, w_1, w_2.

(b) case $w = [1, 2, 3]^T$, $w = -[1, 2, 3]^T$

→ Soln :- $h(x) = \text{sign}(w^T x)$.

(a) $h(x) = +1$, $w^T x > 0$
↳ so sepration beth 2 regions is line whose eq. is $w^T x = 0$

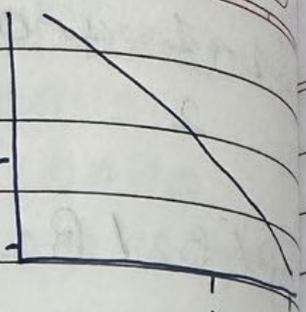
$$w_0 + w_1 x_1 + w_2 x_2 = 0.$$

↳ also can write $x_2 = a x_1 + b$

so $a = \frac{-w_1}{w_2}$, $b = \frac{-w_0}{w_2}$

(b) $w = (1, 2, 3)^T$

$w = -(1, 2, 3)^T$



(Identical line both case)

(b) prove that PLA eventually conv to linear separator for seprabl data.

- w^* be an optimal set of value weights.
- proof shows PLA weights $w(t)$ get more aligned with w^* with each iteration. $w(0) = 0$

(a) Let $\rho = \min_{1 \le n \le N} y_n (w^{*T} x_n)$ show that $\rho > 0$

↳ as every x_n well classified by w^* for all $= 1, \ldots N$

$y_n = \text{sign}(w^{*T} x_n)$

↳ $y_n (w^{*T} x_n) > 0$

— for all $1 \ldots N$

$\rho = \min_n y_n (w^{*T} x_n) > 0$

(b) $w^T(t) w$

conclude

↳ $w^T(t) w^*$

$= w^T(t$

$\geq w^T(t$

↳ Indu

obvious

(b) $w^T(t)w^* \geq w^T(t-1)w^* + \rho$

conclude that $w^T(t)w^* \geq t\rho$

↳ $w^T(t)w^* = [w^T(t-1) + y(t-1)x^T(t-1)]w^*$

$= w^T(t-1)w^* + y(t-1)w^{*T}x(t-1)$

$\geq w^T(t-1)w^* + \rho$

↳ Induction, If $t=0$ we obviously get $0 \cdot w^* \geq 0$

(6) Sample of 10 m... from bin that holds red & ...
- prob of red mar... μ
for $\mu = 0.05$, $\mu = 0.5$, ... $= 0.8$
compute prob of getting no red marbles ($v = 0$) in following case

(a) we draw only one such sample that compute prob $v = 0$
(b) we draw 1000 independent sample, compute prob that one of sample has $v = 0$
(c) repeat for 100000 ind. samples

→ Soln:-

(a) for one sample (10 marble)
$P(v = 0) = (1 - \mu)^{10}$

$\mu = 0.05 \rightarrow P(v = 0) = 0.59873$
$\mu = 0.5 \rightarrow P(v = 0) = 9.7656 \times 10^{-4}$
$\mu = 0.8 \rightarrow P(v = 0) = 1.024 \times 10^{-7}$

(b) for 1000 independent samples

P(Aleast one sample has $v = 0$)

$= 1 - P(v_i > 0 \ \forall i)$

$= 1 - \prod_{i=1}^{1000} [P(v_i > 0)]$

$= 1 -$ ...

$= 1 -$ ...

$= 1 -$ ... $\mu = 0.0$

(c) for 100

P(At lea... sam...

(7) sample of ... tossing ... Assume ... generate ...

→ for coin ... prob of ... tosses g...

P[k/N...

training ...

(a) assume sample size (N) = 10
if all coins have $\mu = 0.05$
compute prob (at least one coin will have $\nu = 0$ for case of 1 coin 1000 coin, 1000000 coins.
repeat for $\mu = 0.8$

↓

Soln → $\mu = 0.05$

P(at least one coin has $\nu = 0$)
$= (1 - 0.05)^{10}$
$= 0.05987$

→ for 1000 coins ⇒
$1 - [1 - (1 - 0.05)^{10}]^{1000} = 1;$

→ for 1000000 coins
$1 - [1 - [1 - 0.05)^{10}]^{1000000}$

(b) $N = 6$, 2 coins with both coins plot prob $P[\max_i |\nu_i - \mu_i| > \epsilon]$
L G ϵ [0, 1]
hoeffding bound

↓ $P[|\nu - \mu| > \epsilon] \leq 2e^{-2N\epsilon^2}$

Soln:- $N = 6$, $\mu = 0.5$ → Hoeffding inequality
2 coins

$P(\max_i |\nu_i - \mu_i| > \epsilon)$

$$P(|V_1 - \mu_1| > \epsilon \text{ or } |V_2 - \mu_2| > \epsilon)$$

$$= P(|V_1 - \mu_1| > \epsilon) + P(|V_2 - \mu_2| > \epsilon)$$

$$- P(|V_1 - \mu_1| > \epsilon \text{ and } |V_2 - \mu_2| > \epsilon)$$

$$= P(|V_1 - \mu_1| > \epsilon) + P(|V_2 - \mu_2| > \epsilon)$$

$$- P(|V_1 - \mu_1| > \epsilon) P(|V_2 - \mu_2| > \epsilon)$$

from bin

μ

1 = 0.8

g no red

wing cas

h sample

ndent

at one

samples

narble)

0.59873

656 × 10^-4

024 × 10^-7

amples

has v=0)

$$= 1 - \prod_{i=1}^{1000} [1 - P(v_i = 0)]$$

$$= 1 - \prod_{i=1}^{1000} [1 - (1-\mu)^{10}]^{1000}$$

$$= 1 - [1 - (1-\mu)^{10}]^{1000}$$

$\mu = 0.05 \rightarrow$ ans = 1

(c) for 100000 independant sample.

$$P(\text{At least one sample has } v=0) = 1 - (1 - (1-\mu)^{20})^{100000}$$

(7) sample of head & tail created by tossing coin a no. of times independtly Assume we have no. of coins that generate diff. samples independently

→ for coin → prob of head = μ
prob of obtaining k heads in N tosses given by binomial dist.

$$P[k/N, \mu] = \binom{N}{k} \mu^k (1-\mu)^{N-k}$$

training error $v = k/N$

Page No.
Date

2.1 eq: $E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{1}{2N} \ln \frac{2M}{\delta}}$

in this eq. set $\delta = 0.09$

let $\epsilon(M, N, \delta) = \sqrt{\frac{1}{2N} \ln \frac{2M}{\delta}}$

(a) for $M = 1$, how many examples do we need to make $\epsilon \leq 0.05$?

$M = 1, \delta = 0.03, \epsilon \leq 0.05$

$$N \geq \frac{1}{2\epsilon^2} \ln \frac{2 \cdot M}{\delta}$$

$$\geq \frac{1}{2 \cdot (0.05)^2} \ln \frac{2}{0.03}$$

$$= 839.94$$

(2) Show that for learning model for +ve rectangles

$m_H(4) = 2^4$, $m_H(5) < 2^5$

Hence give a bound for $m_H(N$

↳ for $N = 4$, if we consider 4 non-aligned (all dichotomies generated) H shatters these points

So we have $m_H(4) = 2^4$

→ for $N = 5$, no matter how you place 5 points, some points inside rectangle

- not able to generate all dichotomies so $m_H(5) < 2^5$

↳ for rectangle $d_{VC} = 4$

$$m_H(N) = N^4 + 1$$

③ compute maximum no. of dichotomies $m_H(N)$ for these learning models. and consequently compute d_{VC}, the VC dim.

↳ (a) positive or negative ray
H contains func. which are +1 on $[a, \infty)$ for some a. together with those that are +1 on $(-\infty, a]$

Soln:- we already know that growth function for +ve rays is equal to $N+1$

- if we enumerate dichotomies added by negative rays, we get $N-1$ new dichotomies.

- you got opposite of ones from positive rays and you have to substract 2 dichotomies where all points are +1 & -1.

so $m_H(N) = 2N$

Page No.
Date

- As the largest value of N for which $m_H(N) = 2^N$
(as $m_H(3) = 6$)
so $d_{VC} = 2$

(b) positive / neg. interval
- H contains func. which are +1 on an interval [a, b]
-1 elsewhere on an interval [a, b], +1 elsewhere

→ growth func for +ve interval
$= N^2/2 + N/2 + 1$

If we add new dichotomies generated by -ve interval we get N-2 new ones.

$N = 1$ → already generate 2 dichotomies with +ve interval

$$m_H(N) = \frac{N^2}{2} + \frac{3N}{2} - 1 \quad N \geq 1$$

& $m_H(N) = 2$ if $N = 1$

largest value of N for which $m_H(N) = 2^N$ → (3)
$d_{VC} = 3$

Page No.
Date

(c) 2 concentric spheres :-

↳ H contains func. which are +1 for $a \leq \sqrt{x_1^2 + \dots x_d^2} \leq b$

↳ growth func of concentric circle map.

$R^d \rightarrow [0, +\infty]$

$\phi(x_1, \dots x_d) \rightarrow r = \sqrt{x_1^2 + \dots x_d^2}$

same as the interval.

$m_H(N) = \frac{N^2}{2} + \frac{N}{2} + 1$.

independent of d.

As the largest value of N for which $m_H(N) = 2^N$.

↳ $d_{VC} = 2$

* Show that $m_H(2N) \leq m_H(N)^2$. hence obtain generalization bound. which involves $m_H(N)$

→ ex: we have 3 ways to dichotomize 2 points x_1, x_2 → for 2 ways to dichotomize another 2 points here we have $3 \times 2 = 6$ ways to dichotomize all 4 points.

Page No.
Date

let us say $m_H(N) = p$
now if we partition any set of 2N points into set of N points each, each of these 2 partition produce p dichotomies at most.

If we combine → max. dicho. would be prod. of p by p
$m_H(2N) = m_H(N)^2$

* Suppose $m_H(N) = N+1$ so $d_{VC} = 1$ you have 100 training ex. use generalization bound to give a bound for Eout with 90% confidence, Repeat for N=1

$N = 100$

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{100} \ln \frac{4 \cdot (2 \cdot 100+1)}{0.1}}$$

$$= E_{in}(g) + 0.8481596.$$

→ with prob at least 90%. N=10000

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{10000} \ln 4\left(\frac{2 \cdot 10000+1}{0.1}\right.}$$

* let H = {
prove th
$d_{VC}(H) = d$
$m_H(d) = 2$
$2^d = m_H(d$

Page No.
Date

* let $H = \{h_1, \ldots h_m\}$ with finite M. prove that $dvc(H) \leq \log_2 M$

$dvc(H) = d$. by definition we have $m_H(d) = 2^d$

$$2^d = m_H(d) = \max_{x_1, \ldots x_d} |H(x_1, \ldots x_d)|$$

$$= \max_{x_1, \ldots x_d} |\{(h(x_1) \ldots h(x_d)) : h \in H\}|$$

$$\leq |H| = M$$

so $d \leq \log_2(M)$

* for hypothesis set $H_1, \ldots H_K$ with finite vc dim $dvc(H_K)$ derive and prove tightest upper & lower bound that you can get on dvc