CO2774
Machine Learning
Major Exam
Elliptical SVM: I Aug 31, 2020

D.J.

$h_0(x) = 1\left\{\sum_{j=1}^{n} \frac{x_j^2}{a_j^2} \leq 1\right\}$ ∈ ⇒ corresponds to elliptical decision boundary I

$\boxed{\sum x_j^2/a_j^2 = 1}$

$\phi(x_j) =$

Let $x_j^2 = t_j$ & $1/a_j^2 = w_j$ $(w_j \geq 0)$ — ①

$\phi(x)$ kernel fn

Then using the transformation defined in ① we have:-

$\sum_{j=1}^{n} w_j t_j = 1$ as the decision boundary.

Equivalently, we have

$\sum_{j=1}^{n} w_j t_j - 1 = 0$ ⇒ equation of a hyperplane

which can equivalently be written as:- $w' = \{w_j\}_{j=1}^{n}$

$\boxed{w_0'^T t + b = 0}$ [using the fact that t_j: $-w_j' = w_j * (-b)$ $(-b) > 0$]

There is an additional scale factor here that we can exploit

⇒ $1\{w_0'^T t + b \geq 0\}$ ⇒ ≡ $h_0(x)$.

& $1\{w_0'^T t + b < 0\}$ as −ve class [WLOG we can call above as +ve class & the signs of the classes]

with additional constraint that

$$\boxed{w' \geq 0}$$

Note:- Since data is separable using optimal boundary, to $\& \; x_i \geq 0$ we do not need to explicitly enforce the constraint $-b \geq 0$ (think why!).

$\Rightarrow$ converting this into SVM form we get

$$\min_{w,b} \; \tfrac{1}{2} w'^T w'$$

$$y^{(i)}\left[w^T t^{(i)} + b\right] \geq 1 \quad \forall i \quad y^{(i)} \in \{1, -1\}$$

→ Langrange multipliers α.

$w' \geq 0$ — Langrange multipliers: $-\beta \; (\beta_1 \cdots \beta_n)$

Therefore, Langrangian ~~multiplier~~ can be written as:-

$$\mathcal{L}(w, b, \alpha, \beta) = \tfrac{1}{2} w^T w' + \sum_{i=1}^{m} \alpha_i \left[1 - y^{(i)}\left[w^T t^{(i)} + b\right]\right] + (-1)\beta^T w'$$

Now, to get the dual we will equate gradient wrt primal variables to 0.

$$\nabla_w \mathcal{L}(w, b, \alpha, \beta) = w' + \sum_{i=1}^{m} -\alpha_i y^{(i)} t^{(i)}$$

$+(-1)\beta$ =

Equating this to zero, we get

$$\boxed{w' = \sum_{i=1}^{m} \alpha_i y^{(i)} t^{(i)} + \beta}$$

similarly, $\nabla_b \mathcal{L}(w, b, \alpha, \beta) = \sum_{i=1}^{m} \alpha_i y^{(i)} (-1) = 0$

$$\Rightarrow \boxed{\sum \alpha_i y^{(i)} = 0}$$

② substituting the value of w in the Lagrangian, we get:-

$$L(w,b,\alpha,\beta) = \frac{1}{2}\left[\sum_{i=1}^{m}\alpha_i y^{(i)} x^{(i)} + \beta\right]^T\left[\sum_{i=1}^{m}\alpha_i y^{(i)} x^{(i)} + \beta\right] + \sum_{i=1}^{m}\alpha_i - \sum_{i=1}^{m}\alpha_i y^{(i)}\left[\sum_{i=1}^{m}\alpha_i y^{(i)} x^{(i)} + \beta\right]^T x^{(i)} + b\Big] - \beta^T\left[\sum_{i=1}^{m}\alpha_i y^{(i)} x^{(i)} + \beta\right]$$

$$= \frac{1}{2}\left[\sum_{i,j=1}^{m} \overset{①}{\alpha_i\alpha_j y^{(i)}y^{(j)} x^{(i)T} x^{(j)}}\right] + \frac{1}{2}\beta^T\beta + \frac{2}{2}\beta^T\sum_{i=1}^{m}\alpha_i y^{(i)} x^{(i)} + \sum_{i=1}^{m}\alpha_i - \sum_{i=1}^{m}\alpha_i y^{(i)}\left[\sum_{j=1}^{m} \overset{②}{\alpha_j y^{(j)} x^{(j)}} + \beta\right]^T x^{(i)} + b\Big] - \beta^T\sum_{i=1}^{m}\alpha_i y^{(i)} x^{(i)} + (-1)\beta^T\beta$$

$$= \sum_{i=1}^{m}\alpha_i + -\frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y^{(i)}y^{(j)} x^{(i)T} x^{(j)} - \frac{1}{2}\beta^T\beta - \sum_{i=1}^{m}\left[\beta^T\alpha_i y^{(i)} x^{(i)}\right] - b\sum_{i=1}^{m}\alpha_i y^{(i)}$$

$D(\alpha,\beta)$

$$\equiv \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\beta^T\beta - \frac{1}{2}\sum\alpha_i\alpha_j y^{(i)}y^{(j)} x^{(i)T} x^{(j)} - \beta^T\sum_{i=1}^{m}\alpha_i y^{(i)} x^{(i)}$$

Dual:-

$$\max_{\alpha,\beta} \quad D(\alpha,\beta)$$

~~α,β~~ $\alpha \geq 0, \quad \beta \geq 0$

$$\sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

Q.2. PCA (generalized):

From standard PCA, we know that the PCA components can be obtained by solving

(a) $$\max_{u_\phi} \quad u_\phi^T \sum_{i=1}^{m} \left[\phi(x^{(i)}) \phi(x^{(i)})^T\right] u_\phi$$

$$u \equiv \Sigma^\phi$$

in the transformed space (one component at a time).

This can be done using (finding)/solving the following eigen problem: value

$$\Sigma^\phi u_\phi = \lambda u_\phi$$

with ~~subset~~ where $$\Sigma^\phi = \sum_{i=1}^{m} \frac{\phi(x^{(i)}) \phi(x^{(i)})^T}{u}$$

substituting this in the above equation, we get

$$\sum_{i=1}^{m} \frac{\phi(x^{(i)}) \phi(x^{(i)})^T}{u} u_\phi = \lambda u_\phi$$

(3)

$$\sum_{i=1}^{m} \phi(x^{(i)}) \underbrace{\left[\phi(x^{(i)})^T u_p \phi\right]}_{u} = \lambda u_p \phi$$

$$\Rightarrow u_p \phi = \frac{1}{\lambda m} \left[\sum_{i=1}^{m} \phi(x^{(i)}) \left[\phi(x^{(i)})^T u_p \phi\right] \right]$$

$\Rightarrow$ Desired coefficients are given as:-

$$\boxed{a_p^i = \frac{\phi(x^{(i)})^T u_p \phi}{\lambda m}} \quad \text{s.t}$$

$$u_p \phi = \sum_{i=1}^{m} a_p^i \phi(x^{(i)})$$

(b) We have:-

$$u_p \phi = \frac{1}{\lambda m} \sum_{i=1}^{m} a_p^i \phi(x^{(i)})^T$$

Pre-Multiplying both sides by $\phi(x^{(j)})^T$ for some j, we get

$$\underbrace{\phi(x^{(j)})^T u_p \phi}_{\left[a_p^j \cdot \lambda m\right]} = \frac{1}{\lambda m} \cdot \phi(x^{(j)})^T \sum_{i=1}^{m} a_p^i \phi(x^{(i)})$$

$$\Rightarrow a_p^j \lambda m = \frac{1}{\lambda m} \sum_{i=1}^{m} a_p^i \underbrace{\phi(x^{(j)})^T \phi(x^{(i)})}_{K(x^{(i)}, x^{(j)})}$$

$$\boxed{\lambda a_p^j = \frac{1}{m} \sum_{i=1}^{m} a_p^i K(x^{(i)}, x^{(j)})}$$

This equation get rid of individual $\phi(x^{(i)})$ term

(c) Continuing further we get (using simple matrix algebra)

$$\boxed{\lambda a_\ell = \frac{1}{m} K^M a_\ell}$$ [λ is some scalar quantity]

$\Rightarrow$ a_ℓ is eigenvector of the matrix

$\boxed{\frac{K^M}{m}}$ & λ corresponding eigenvalue.

$K^M = \begin{bmatrix} & | & \\ \text{---} & K(x^{(i)}, x^{(j)}) & \text{---} \\ & | & \end{bmatrix}$ is the kernel matrix.

Finally we have to ensure that $\phi_\ell^T u_\ell \phi_\ell = 1$

(a) $\phi(x^{(i)})^T u_\ell \phi$
$= \phi(x^{(i)})^T \sum_{i=1}^{m} a_\ell^i \phi(x^{(i)})$
$= \sum_{i=1}^{m} a_\ell^i \phi(x^{(i)})^T \phi(x^{(j)})$
$K(x^{(i)}, x^{(j)})$
a_ℓ^i can be computed in terms of Kernel fun.
$\Rightarrow \phi(x^{(i)})^T u_\ell \phi$ can be computed in terms of K fn.

$\Rightarrow \sum_{i=1}^{m} a_\ell^i \phi(x^{(i)})^T \sum_{i=1}^{m} a_\ell^i \phi(x^{(i)}) = 1$

$\Rightarrow \sum_{i,j=1}^{m} a_\ell^i \phi(x^{(i)})^T \phi(x^{(j)}) a_\ell^j = 1$

$\Rightarrow$ Hence $\boxed{a_\ell^T K^M a_\ell = 1}$

Putting this in equation

$a_\ell = \frac{1}{\lambda m} K^M a_\ell$ [Pre-multiply by a_ℓ^T]

$\hookrightarrow a_\ell^T a_\ell = \frac{1}{\lambda m} a_\ell^T K^M a_\ell$

we get $\boxed{a_\ell^T a_\ell = \frac{1}{\lambda m} \cdot 1}$ $\square$

(24)

Q3. (a) In general. For m points. $LL(\theta)$ = for logistic regression

$$\nabla_\theta LL(\theta) = \nabla_\theta \sum_{i=1}^{m} y^{(i)} \log \frac{1}{[1+e^{-\theta^T x^{(i)}}]} + [1-y^{(i)}] \log\left(\frac{e^{-\theta^T x^{(i)}}}{1+e^{-\theta^T x^{(i)}}}\right)$$

$$LL(\theta) = \sum_{i=1}^{m} y^{(i)} \left[\log \frac{1}{1+e^{-\theta^T x^{(i)}}}\right] + \sum_{i=1}^{m} (1-y^{(i)}) \log\left[\frac{e^{-\theta^T x^{(i)}}}{1+e^{-\theta^T x^{(i)}}}\right] \Rightarrow$$

$$= -\nabla_\theta \left[\sum_{i=1}^{m} y^{(i)} \log[1+e^{-\theta^T x^{(i)}}] + [1-y^{(i)}][\theta^T x^{(i)}] + [1-y^{(i)}] \log[1+e^{-\theta^T x^{(i)}}]\right]$$

$$= -\nabla_\theta \left[\sum_{i=1}^{m} \log(1+e^{-\theta^T x^{(i)}}) + \sum_{i=1}^{m} (1-y^{(i)}) \theta^T x^{(i)}\right]$$

$$= \sum_{i=1}^{m} \frac{1}{[1+e^{-\theta^T x^{(i)}}]} [e^{-\theta^T x^{(i)}}][-x^{(i)}](-1) + \nabla_\theta \sum_{i=1}^{m} \theta^T x^{(i)} (1-y^{(i)}) (-1)$$

$$= \sum_{i=1}^{m} \left[\left[\frac{e^{-\theta^T x^{(i)}}}{1+e^{-\theta^T x^{(i)}}} x^{(i)}\right] - (1-y^{(i)}) x^{(i)}\right]$$

$$\boxed{\nabla_\theta LL(\theta) = \sum_{i=1}^{m} \left[y^{(i)} - 1/(1+e^{-\theta^T x^{(i)}})\right] x^{(i)}}$$

$$[\nabla_\theta^2 LL(\theta)]_{ij} = \sum_{i=1}^{m} \frac{1}{[1+e^{-\theta^T x^{(i)}}]}$$

$$\nabla_\theta^2 LL(\theta) = \sum_{i=1}^{m} -1 \left[\frac{1}{1+e^{-\theta^T x^{(i)}}}\right]^2 [x^{(i)} x^{(i)T}]$$

Thus is the ≤ -ve semi definite

since:-

~~$u^T K$~~ λ +ve number

$$-\sum_{i=1}^{m} u^T K^{(i)} x^{(i)} x^{(i)T} u$$

$$= -\sum_{i=1}^{m} K^{(i)} \left[u^T x^{(i)}\right]^2 \le 0$$

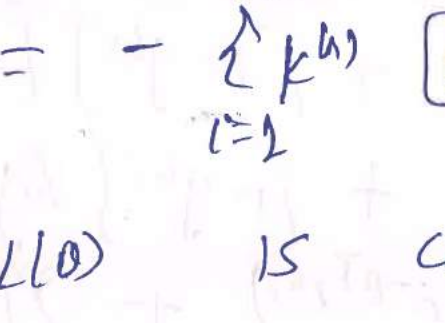
⇒ $LL(\theta)$ is concave.

$-LL(\theta)$ is convex

Holds:- for m points

⇒ holds for m = 2 points.

(b)



~~Let :-~~

~~P_1:- prob. of classifying point 1 as +ve~~

~~P_2:- prob. of classifying point 2 as -ve (correct)~~

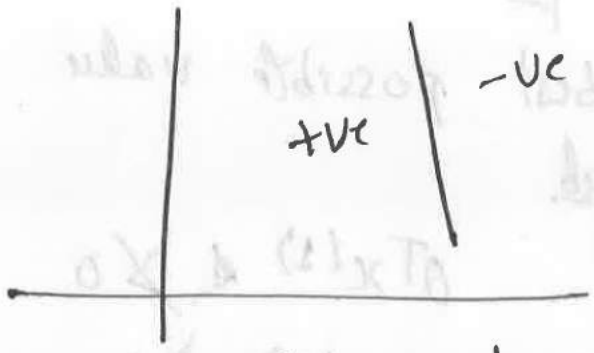
~~$LL(\theta) = \log P_1 + \log P_2$~~

~~⇒ Dist :- d_1~~

~~Dist :- d_2:~~

(5)

Q.3 (b)



Two points in R^2 for visualization (only).

First, we will show that optimal decision boundary must classify the 2 points correctly.

→ likelihood for 1st point, likelihood for 2nd point

$$LL(\theta) = \log\left[\frac{1}{1+e^{-\theta^T x^{(1)}}}\right] + \log\left[\frac{e^{-\theta^T x^{(2)}}}{1+e^{-\theta^T x^{(2)}}}\right]$$

$$= \log\left[\frac{1}{1+e^{-\theta^T x^{(1)}}}\right] + \log\left[\frac{1}{1+e^{+\theta^T x^{(2)}}}\right]$$

To maximize this, we would need to maximize both terms

clearly if $\theta^T x^{(1)} > 0$ & $\theta^T x^{(2)} < 0$ i.e. decision boundary is between the two points

we can make $\theta^T x^{(1)} \to \infty$ & $+[\theta^T x^{(2)}] \to -\infty$ by scaling (θ) arbitrarily & not changing the decision surface. Replace θ by $k\theta$ $\underline{k>0}$

as $k \to \infty$

$LL(\theta) \to 1 + \frac{2}{\infty} \quad 1+1=2$

Thus is the best possible value that can be achieved.

Note that if $\theta^T x^{(1)} \not< 0$

or $\theta^T x^{(2)} \not> 0$

i.e the points are to separated by the boundary (with +ve & -ve on correct sides):-

Then at least one of the terms will not be able to reach 1

i.e. if $\theta^T x^{(i)} < 0$ then

maximum value of $\frac{1}{1+e^{-\theta^T x^{(i)}}}$

is $1/2 = 0.5$

$\Rightarrow$ (Hence $LL(\theta) \leq 1.5$ in this case

$\Rightarrow$ θ can not be optimal.

~~(c) for using part (b) optimal theta~~

Therefore, for the optimal boundary, it must be somewhere between the +ve & -ve points, specifically the boundary, with equal distance from the +ve & -ve points will be optimal. (one of the

(c) Maximum value of $LL(\theta)$ in this setting as shown in part (b) above is 2.

But it is only in the limit

$$\left[\frac{1}{1+e^{-\theta^T x^{(1)}}k} + \frac{1}{1+e^{+\theta^T x^{(2)}}k}\right] = LL(\theta)$$

$k \to \infty$

But for any finite value of k

$LL(\theta)$ $LL(\theta) < 0$

But that is not an issue since, during practical implementation, we stop learning when $|LL(\theta^{(t+1)}) - LL(\theta^{(t)})| < \epsilon$ for some ϵ. Thus with Actually, we can write $\nabla_\theta LL(\theta) \neq 0$

happen since $LL(\theta) \to 2$ in the limit $\Rightarrow \exists \theta^{(t)}$ & $\theta^{(t+1)}$ as we take gradient steps such that $LL(\theta^{(t)})$ is as close to 2 as possible

$$\nabla_\theta LL(\theta) = \sum_{i=1}^{2} \left[y^{(i)} - h_\theta(x^{(i)})\right] x^{(i)}$$

$h_\theta(x^{(i)}) \to 1$ or 0 as $h_\theta(x^{(i)}) = \left[\frac{1}{1+e^{-\theta^T x^{(i)}}}\right]$ the coefficient would be as $\theta \to k \to \infty$ $h_\theta(x^{(i)}) \to 1$ &

as $k \to \infty$ $\log(1) + \log(1) = 0 + 0 = 0$

$LL(\theta) \to 1 + \frac{2}{\infty}$ ~~$1+1=2$~~

Thus is the best possible value that can be achieved.

Note that if $\theta^T x^{(1)} \not< 0$

or $\theta^T x^{(2)} \not> 0$

i.e the points are to separated by the boundary (with +ve & -ve on correct sides):-

Then at least one of the terms will not be able to reach 1

i.e. if $\theta^T x^{(i)} \geq 0$ then

maximum value of $\frac{1}{1+e^{-\theta^T x^{(i)}}}$

is $1/2 = 0.5$

$\Rightarrow$ (then $LL(\theta) \leq$ ~~2~~ 0.5] in this case (other term can be at most 0)

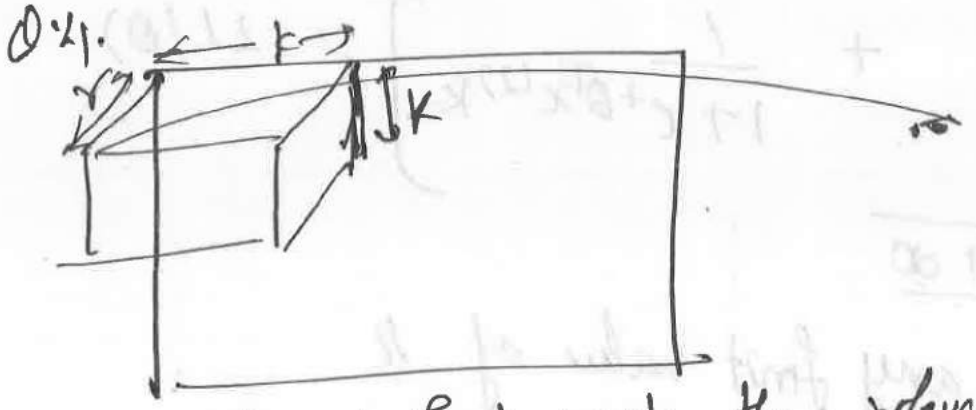
$\Rightarrow$ θ can not be optimal.

~~(c) For (b) optimal theta~~

Therefore, for the optimal boundary, it must be somewhere between the +ve & -ve points, specifically the boundary, with equal distance points from the +ve & -ve will be optimal. (one of the)

gradient goes to zero. Similarly for the second term

$$\text{Ex: } \left[h_\theta(x^{(i)}) = \frac{1}{1+e^{-\theta^T x^{(i)}}} \right]$$



Let us first write the relevant equations:-

Let the kernel be denoted by K.

i, j vary over width/height of the image

Let $x^{(t)}_{i,j,l}$ denote the feature map at (i,j) position & depth l $(1 \le i,j \le w)$

$1 \le l \le r$

p:- Kernel index $\quad 1 \le p \le r$

x':- Feature after convolution then (before average pooling)

Feature map before the pooling layer

$$\left[\sum_{l=1}^{r} K^{p}_{u,v,l} \cdot x_{i+u, j+v, l} + b^{p} \right] = x'^{(t)}_{i,j,p}$$

$1 \le i,j \le w \quad$ & $\quad 1 \le l \le r$

$\boxed{1 \le p \le r}$

(1)

Average Pooling :-

$$x^{(l+1)}_{i,j,p} = \sum_{\alpha,\beta=1}^{2} \frac{x'^{(l)}_{i+\alpha,j+\beta,p}}{4}$$

Then :- We have :- (Gradient wrt feature map going into Pooling layer)

$$\frac{\partial L}{\partial x'^{(l)}_{i,j,p}} = \sum_{\alpha,\beta=1}^{2} \frac{\partial L}{\partial x^{(l+1)}_{i+\alpha,j+\beta,p}} \cdot \left[\frac{\partial x^{(l+1)}_{i+\alpha,j+\beta,p}}{\partial x'^{(l)}_{i,j,p}}\right]$$

(1) From (2) ⇓ 1/4

$$= \frac{1}{4} \sum_{\alpha,\beta=1}^{2} \frac{\partial L}{\partial x^{(l+1)}_{i+\alpha,j+\beta,p}} \quad \text{--- (1)}$$

Gradient wrt feature map going into convolution Layer.

$$\frac{\partial L}{\partial x^{(l)}_{i,j,p}} = \sum_{u,v=1}^{K} \sum_{p=1}^{\gamma} \left[\frac{\partial L}{\partial x'^{(l)}_{i,j,p}}\right] \left[\frac{\partial x'^{(l)}_{i,j,p}}{\partial x^{(l)}_{i+u,j+v,p}}\right]$$

(2)

⇓ From Eq (1)

⇓ From Eq (1)

When ~~t=1=t=d~~ we know that

$\left[\frac{\partial L}{\partial x(t)}\right]$ is given to us

$$\frac{\partial x'^{(t)}_{i,j,p}}{\partial x^{(t)}_{i+u,j+v,p}} = \sum_{l=1}^{p} K^{p}_{u,v,l}$$

Finally,

(3) $$\frac{\partial L}{\partial K^{p}_{u,v,l}} = \sum_{t=1}^{d}\left[\sum_{i,j=1}^{w}\left[\frac{\partial L}{\partial x'^{(t)}_{i,j,p}}\right]\cdot\left[\frac{\partial x'^{(t)}_{i,j,p}}{\partial K^{p}_{u,v,l}}\right]\right]$$

From (1)

$$\frac{\partial x'^{(t)}_{i,j,p}}{\partial K^{p}_{u,v,l}} = x_{i+u,j+v,l}$$

(4) $$\frac{\partial L}{\partial b^{p}} = \sum_{t=1}^{d}\left[\sum_{i,j=1}^{w}\frac{\partial L}{\partial x'^{(t)}_{i,j,p}}\cdot\frac{\partial x'^{(t)}_{i,j,p}}{\partial b^{p}}\right]$$

From (1)

$$\frac{\partial x'^{(t)}_{i,j,p}}{\partial b^{p}} = 1$$

Need to double check the computation.

0.5 - Generative Process -

$z^{(i)} \sim$ ~~Bernoulli~~ MultiNom(Φ)

$z^{(i)} \in \{1 \dots k\}$

$y^{(i)} \sim$ Bernoulli(ϕ) :- does not depend on $z^{(i)}$ (as given in the question).

$\forall j$:- $x_j^{(i)} | y^{(i)}=t, z^{(i)}=l$ $\quad l \in \{1 \dots k\}$, $t \in \{1, 0\}$

$\equiv$ Bernoulli($\theta_{j|y=y^{(i)}, z=z^{(i)}}$)

$\equiv$ Bernoulli($\theta_{j|y=t, z=l}$)

Parameters of the Model:

Φ :- $(\Phi_1 \dots \Phi_k)$:- latent mixture parameters

k parameters (one redundant)

$\Theta \Rightarrow \sum \Phi_l = 1$

ϕ :- prior class probability. (one parameter)

$\{\theta_{j|y=t, z=l}\}_{j=1, t=0, z=1}^{n, 1, k}$

$n \times 2 * k$

$= 2nk$ parameters

Log - Likelihood

$$\sum_{i=1}^{m} \log\left[p(x^{(i)}, y^{(i)}; \theta)\right]$$

$$= \sum_{i=1}^{m} \log\left[p(x^{(i)} | y^{(i)}; \theta)\; p(y^{(i)}; \phi)\right]$$

$$= \sum_{i=1}^{m} \log\left[\sum_{z^{(i)}} \left[p(x^{(i)} | y^{(i)}, z^{(i)}; \theta)\right] \cdot p(z^{(i)}; \Phi)\right] \cdot p(y^{(i)}; \phi)\Big]$$

(since $y^{(i)}$ are independent of $z^{(i)}$)

$$= \sum_{i=1}^{m} \log\left[\sum_{z^{(i)}} \left[p(x^{(i)} | y^{(i)}, z^{(i)}; \theta)\right] p(z^{(i)}; \Phi)\right] + \log p(y^{(i)}; \phi)\Big]$$

can be expressed in terms of params

$$= \sum_{i=1}^{m} \log \sum_{z^{(i)}} \left[\prod_{j=1}^{n} p(x_j^{(i)} | y^{(i)}, z^{(i)}; \theta)\right] p(z^{(i)}; \Phi)\Big] + \log p(y^{(i)}; \phi)\Big]$$

can be expressed in terms of parameters of the model

⇓

can be expressed in terms of params.

(skipping one step which substitute the parameters in place of corresponding expressions)

$$= \sum_{i=1}^{m} \log \sum_{z^{(i)}} 1\{z^{(i)}=c\} \prod_{j=1}^{n} 1\{y^{(i)}=1\}\left[1\{x_j^{(i)}=1\}\theta_{j|y=1,z=c} + 1\{x_j^{(i)}=0\}(1-\theta_{j|y=1,z=c})\right]$$

(a)

$$+ \; 1\{y^{(i)}=0\}\Big[1\{x_j^{(i)}=1\}\theta_{j|y=0,z^{(i)}=0} + 1\{x_j^{(i)}=0\}(1-\theta_{j|y=0,z^{(i)}=0})\Big]$$

$$+ \sum_{i=1}^{m} y^{(i)}\log\phi + (1-y^{(i)})\log(1-\phi)\Big]$$

(b) EM based estimation:-

E-step:→

$$P(z^{(i)} \mid x^{(i)}, y^{(i)}; \theta)$$

$$= \frac{P(z^{(i)}, x^{(i)}, y^{(i)}; \theta)}{P(x^{(i)}, y^{(i)}; \theta)}$$

$$= \frac{P(x^{(i)} \mid y^{(i)}, z^{(i)}; \theta)\; P(y^{(i)}; \phi)\; P(z^{(i)}; \Phi)}{\sum_{z^{(i)}} P(x^{(i)}, y^{(i)} \mid z^{(i)}; \theta)\; P(z^{(i)}; \phi)}$$

$$= \frac{\left[\prod_{j=1}^{n} P(x_j^{(i)} \mid y^{(i)}, z^{(i)}; \theta)\right] \cdot P(y^{(i)}; \phi)\; P(z^{(i)}; \Phi)}{\sum_{z^{(i)}} \left[P(x_j^{(i)} \mid y^{(i)}, z^{(i)}; \theta)\right] P(y^{(i)}; \phi)\; P(z^{(i)}; \Phi)\Big]}$$

We know how to compute θ this given set of parameters.

M-step:- Given $P(z^{(i)}=\rho)$, we want to estimate the parameters of the same Bayes model in the previous

$\phi = \rightarrow$ independent of $z^{(i)}$'s

$$\phi = \frac{\sum_{i=1}^{m} 1\{y^{(i)}=1\}}{m}$$ constant throughout

$$\phi = \frac{\sum_{i=1}^{m} P(z^{(i)}=l)}{m}$$ [softer version of $\frac{\sum_{i=1}^{m} 1\{z^{(i)}=l\}}{m}$]

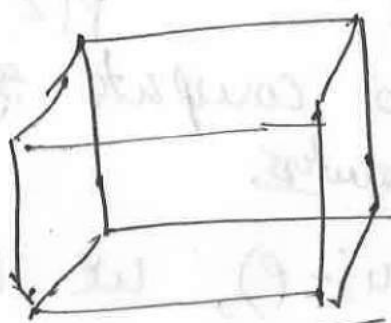
M-step

$$\theta_{j|y=t,z^{(i)}=l} = \frac{\sum_{i=1}^{m} P(z^{(i)}=l)\, 1\{y^{(i)}=t\}\, 1\{x^{(i)}=1\}}{\sum_{i=1}^{m} 1\{y^{(i)}=t\}\, P(z^{(i)}=l)}$$

Softer version of Naive Bayes parameter estimation [can do Laplace smoothing] by adding 1 to the numerator & 2 to the denominator

---

Q8: $\mathcal{H}_{11}^3$:-

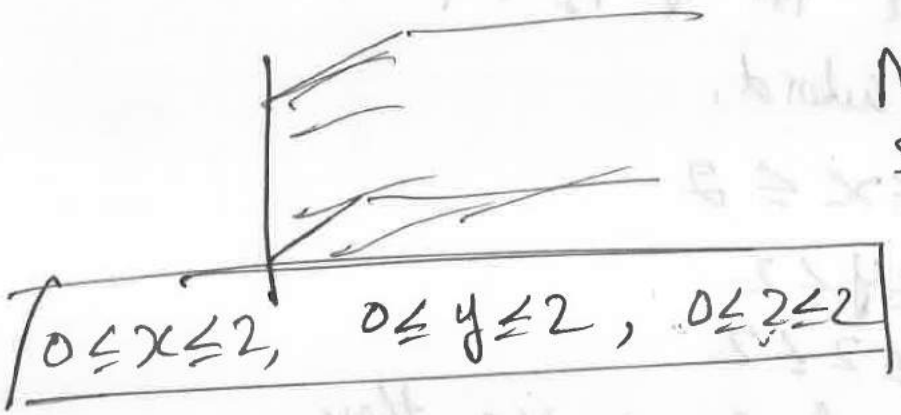


Recall:- $VC(\mathcal{H}_{11}^2) = 4$

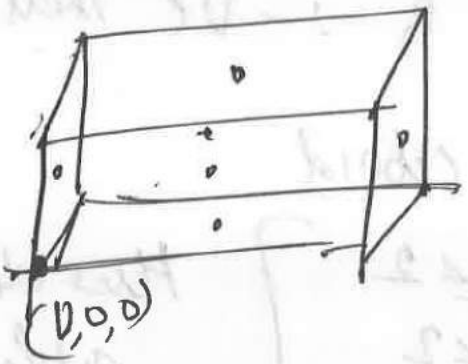
Claim:- VC-Dimension of class $\mathcal{H}_{11}^3$ (axis || cuboids) $= 6$

(10) First, we will show that $VC(H_1^3) = 6$

Consider a cuboid of length, width & height = 2 with bottom left inner corner at (0,0,0)



Now, let us place six points at the center of six faces of the cuboid i.e.,-



P_1 (1,1,0)] top & bottom face
P_2 (1,1,2)]

P_3 (1,0,1) Front & back face
P_4 (1,2,1)

P_5 (0,1,1) Left & Right Face
P_6 (2,1,1)

cut

~~(1/2, 1/2, 0)~~
~~(0, 1/2, 0)~~
~~(0, 0, 1/2)~~
(1/2, 1/2, 0)
~~(1/2, 0, 1/2)~~
~~(1/2, 1/2, 1/2)~~

Then these six points can be easily shattered by moving the plane

at which the point lies by $\pm\varepsilon$ ($\varepsilon>0$) to give the concerned point a label of +ve or −ve (or vice-versa) where ε is very small

e.g. Consider $P_1 :- (1,1,0)$

the x in if P_1 : + ve then

In the cuboid,

$-\varepsilon \le x \le 2$

$0 \le y \le 2$

$0 \le z \le 2$

Similarly if P_1 : - ve then

In the cuboid

$\varepsilon \le x \le 2$
$0 \le y \le 2$
$0 \le z \le 2$

] This does not affect the label of other points

⇒ All 96 labelings can be achieved by doing $\pm\varepsilon$ movement of 6 faces independently (corresponding to label of each point).

(11)

Next, we will show that $VC(H_{11}^2) = 6$

i.e. no set of 7 points can be shattered.

Assume Contrary:-

$\exists\ p_1 - - p_7$ s.t H_{11}^2 can shatter them. We will show that 3 labeling of these 7 points that can not be achieved.

Let p_x^{max} :- point with max x value

p_x^{min} :- point with min x value

Similarly define

p_y^{max} & p_z^{max}

p_y^{min} p_z^{min}.

F Let p :- remaining point (at least one such point must be there which is not covered above)

Assign p :- ve label

All other points +ve label

Clearly:-

$p_x^{min}[x] \le P[x] \le p_x^{max}[x]$

$p_y^{min}[y] \le P[y] \le p_y^{max}[y]$

$p_z^{min}[z] \le P[z] \le p_z^{max}[z]$

Let $h \in H_{11}^2$ that achieved this labeling.

Since $p_x^{min}, p_x^{max}, p_y^{min}, p_y^{max}, p_z^{min}, p_z^{max}$ are assigned +ve label, any point lying ~~without~~ they will be inside the cuboid formed by h

$p_x^{min} \le p' \le p_x^{max} \wedge p_y^{min} \le p' \le p_y^{max}$

$\wedge p_z^{min} \le p' \le p_z^{max}$ must also be inside. $\Rightarrow$ p must be inside $h \in H_{11}^3$

$\Rightarrow$ p must have a the label contradiction.

(b)(c) In general:- H_{11}^n :- $\boxed{VC\text{-}Dim(H_{11}^n) = 2n}$

A very similar argument follows

$2n$ points can be shattered by assigning n 2 points on two faces $\perp$ to each of n ~~axis~~ axis. & then making the $\oplus$ argument as earlier

$2n+1$ points can not be shattered by:- assigning the label to a there points which max/min value along any dimension & assigning -ve label to remaining point.