COL774
Machine Learning
Major Exam

Elliptical SVM: Aug 31, 2020

Q.1.

$h_\theta(x) = 1\left\{ \sum_{j=1}^{n} \frac{x_j^2}{a_j^2} \le 1 \right\}$ $\Rightarrow$ corresponds to (elliptical) decision boundary: $\boxed{\sum x_j^2 / a_j^2 = 1}$

$\phi(x_j) =$

$\phi(x)$ Kernel fn: Let $x_j^2 = t_j$ & $1/a_j^2 = w_j$ $(w_j \ge 0)$ — ①

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

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

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

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

There is an additional scale factor here that we can exploit

$\Rightarrow 1\{w'^T t + b \ge 0\}$ $\equiv$ ~~the~~ ~~class~~ $h_\theta(x)$.

(wlog we can call above as +ve class)

& $1\{w'^T t + b < 0\}$ as -ve class. [we are flipping the signs of 2 classes] with additional constraint that

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

Note:- Since data is separable using elliptical boundary, w_0 & $t_{ij} \equiv x_i x_j \geq 0$

we do not need to explicitly enforce the constraint $-b \geq 0$ (think why).

$\Rightarrow$ converting this into SVM form we get

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

$\to$ Langrange multipliers α.

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

$\boxed{w' \geq 0}$ — Langrange multipliers: $-\beta$ $(\beta_1 \ldots \beta_n)$

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

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

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

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

extra compared to notes

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

dual is unconstrained, main idea

Equating this to zero, we get

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

Similarly,

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

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

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

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

$$+ \sum_{i=1}^{m} \alpha_i - \sum_{i=1}^{m} \alpha_i y^{(i)} \left[\left[\sum_{i=1}^{m} \alpha_i y^{(i)} t^{(i)} + \beta\right]^T t^{(i)} + b\right]$$

$$- \beta^T \left[\sum_{i=1}^{m} \alpha_i y^{(i)} t^{(i)} + \beta\right]$$

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

$$+ \cancel{2\times\frac{1}{2}}\, \beta^T \cancel{\sum_{i=1}^{m} \alpha_i y^{(i)} t^{(i)}} + \sum_{i=1}^{m} \alpha_i$$

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

$$- \beta^T \cancel{\sum_{i=1}^{m} \alpha_i y^{(i)} t^{(i)}} + (-1)\beta^T \beta$$

$$= \sum_{i=1}^{m} \alpha_i \cancel{+} - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} t^{(i)T} t^{(j)} - \frac{1}{2} \beta^T \beta$$

$$- \sum_{i=1}^{m} \left[\beta^T \alpha_i y^{(i)} t^{(i)}\right] - b \cancel{\sum_{i=1}^{m} \alpha_i y^{(i)}}^{\,0}$$

$D(\alpha,\beta)$

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

$$- \beta^T \sum_{i=1}^{m} \alpha_i y^{(i)} t^{(i)}$$

Dual: -

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

$$\text{s.t.} \quad \alpha \geq 0, \quad \beta \geq 0$$

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

Q.2. PCA (generalized):

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

(a) $$\max_{u_\ell^\phi} \quad u_\ell^{\phi T} \underbrace{\sum_{i=1}^{m} \left[\frac{\phi(x^{(i)}) \phi(x^{(i)})^T}{m} \right]}_{\equiv \Sigma^\phi} u_\ell^\phi$$

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

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

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

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

substituting this in the above equation, we get

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

(3)

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

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

$\Rightarrow$ Desired coefficients are given as:-

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

$$u_l \phi = \sum_{i=1}^{m} a_l^i \phi(x^{(i)})$$

(b) We have:-

$$u_l \phi = \cancel{\frac{1}{\lambda m}} \sum_{i=1}^{m} a_l^i \phi(x^{(i)})$$

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

$$\underbrace{\phi(x^{(j)})^T u_l \phi}_{[a_l^j \cdot \lambda m]} = \cancel{\frac{1}{\lambda m}} \phi(x^{(j)})^T \sum_{i=1}^{m} a_l^i \phi(x^{(i)})$$

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

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

This equation gets rid of individual $\phi(x^{(i)})$ terms

(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 $\left[\frac{K^M}{m}\right]$ & λ corresponding eigenvalue.

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

(d) $\phi(x^{(i)})^T u_\ell^\phi$
$= \phi(x^{(i)})^T \sum_{j=1}^{m} a_\ell^j \phi(x^{(j)})$
$= \sum_{j=1}^{m} a_\ell^j \underbrace{\phi(x^{(i)})^T \phi(x^{(j)})}_{K(x^{(i)}, x^{(j)})}$

a_ℓ^j can be computed in terms of Ker function

$\Rightarrow \phi(x^{(i)})^T u_\ell^\phi$ can be computed in terms of K fn.

Finally we have to ensure that

$u_\ell^{\phi T} u_\ell^\phi = 1$

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

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

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

Putting this in equation

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

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

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

(IV)

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

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

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

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

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

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

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

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

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

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

Thus is ~~+ve~~ ≤ -ve semi-definite

since:-

~~$u^T K \Lambda$~~ +ve number

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

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

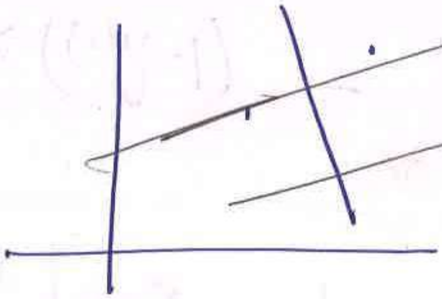
$\Rightarrow LL(\theta)$ is concave.

$-LL(\theta)$ is ~~concave~~ convex

Holds:- ~~for m points~~

$\Rightarrow$ 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)~~

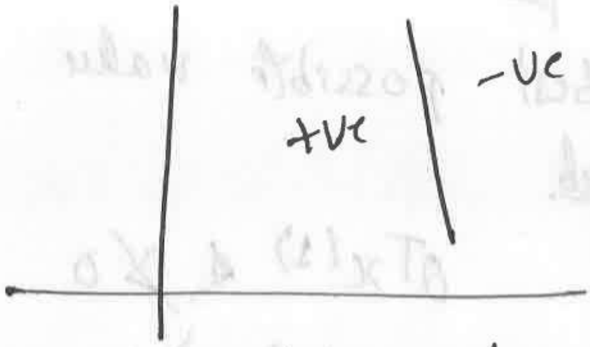
~~$LL(\theta) = \log P_1 + \log P_2$~~

~~$\Rightarrow$ Dist :- d_1~~

~~Dist :- d_2:~~

(5)

Q.3 (b)



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

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

$$LL(\theta) = \underbrace{\log\left[\frac{1}{1+e^{-\theta^T x^{(1)}}}\right]}_{\text{likelihood for 1st point}} + \underbrace{\log\left[\frac{e^{-\theta^T x^{(2)}}}{1+e^{-\theta^T x^{(2)}}}\right]}_{\text{likelihood for 2nd point}}$$

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

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 $k \to \infty$

$LL(\theta) \to \cancel{1+} \frac{1}{\cancel{e}}$ $1+1=2$

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 $\frac{1}{1+e^{-\theta^T x^{(1)}}}$

is $1/2 = 0.5$

$\Rightarrow$ clear $LL(\theta) \leq \cancel{2}\ 1.5$ in this case

$\Rightarrow \theta$ can not be optimal.

~~(c) For Using part(b) optimal deus~~

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

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

$$\underline{\overline{k \to \infty}}$$

But for any finite value of k

$$\boxed{LL(\theta) < 2}$$

But that is not an issue since, during practical implementation, we stop learning when

$$|LL(\theta^{(t+1)}) - LL(\theta^{(t)})| < \varepsilon_0$$

Alternatively, we can use: $\nabla_\theta LL(\theta) \approx 0$

for some ε_0. This will 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

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

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

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

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

This 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 ~~a~~ 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 $\frac{1}{1+e^{-\theta^T x^{(1)}}}$

is $1/2 = 0.5$

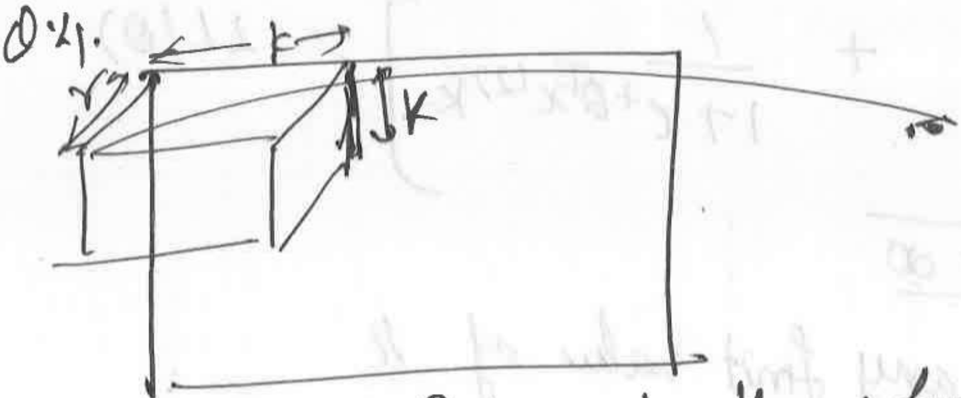
$\Rightarrow$ clear $LL(\theta) \leq \log[0.5]$ in this case (other term can be at most 0)

$\Rightarrow$ θ can not be optimal.

~~c) For Using part (b) optimal class~~

Therefore, for the optimal boundary, it must be somewhere between the +ve & -ve points, specifically the boundary, with equal distance from +ve & -ve points will be optimal. (one of the) gradient goes to zero. Similarly for the second term

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



Let us first write the relevant equations:-

Let the kernel be denoted by K.

i, j vary over width / height of the image

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

$1 \le \ell \le r$

p:- Kernel index $\underline{1 \le p \le r}$

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

Feature map before the pooling layer

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

$1 \le i,j \le W$ $\quad$ $\overline{1 \le \ell \le r}$

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

(7) Average Pooling:-

$$x^{(t+1)}_{i,j,p} = \frac{\sum_{\alpha,\beta=1}^{2} x'^{(t)}_{i+\alpha,j+\beta,p}}{4}$$

Then:- we have: (Gradient wrt ~~poo~~ feature map going into Pooling layer)

$$\frac{\partial L}{\partial x'^{(t)}_{i,j,p}} = \sum_{\alpha,\beta=1}^{2} \underbrace{\frac{\partial L}{\partial x^{(t+1)}_{i+\alpha,j+\beta,p}}}_{\text{From } (5)} \cdot \underbrace{\left[\frac{\partial x^{(t+1)}_{i+\alpha,j+\beta,p}}{\partial x'^{(t)}_{i,j,p}}\right]}_{1/4}$$

(1)

$$= \frac{1}{4} \sum_{\alpha,\beta=1}^{2} \frac{\partial L}{\partial x^{(t+1)}_{i+\alpha,j+\beta,p}} \quad \text{--- (7)}$$

~~dL~~ Gradient wrt the feature map going into ~~in the layer before~~ convolution ~~operation~~ layer.

$$\frac{\partial L}{\partial x^{(t)}_{i,j,p}} = \sum_{u,v=1}^{k} \sum_{p=1}^{r} \underbrace{\left[\frac{\partial L}{\partial x'^{(t)}_{i,j,p}}\right]}_{\text{From Eq (1)}} \left[\frac{\partial x'^{(t)}_{i,j,p}}{\partial x^{(t)}_{i+u,j+v,p}}\right]$$

~~From Eq (1)~~

(2)

when ~~t=1~~ $t = d$ we know that

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

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

Finally,

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

From (1)

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

(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)

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

Need to double check the computation.

(8)

Q.5:- Generative Process:-

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

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

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

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

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

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

Parameters of the Model:

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

$\Theta \Rightarrow$ $\sum \Phi_l = 1$ $\quad$ k parameters (one redundant)

ϕ:- prior class probability. (one parameter)

$\{\theta_{j|y=t, z=l}\}_{j=1, t=0, z=1}^{n, 1, k}$ $\quad$ $n \times 2 \times k = 2nk$ parameters

Log - likelihood:

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

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

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

[since $y^{(i)}$ is independent of $z^{(i)}$.]

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

can be expressed in terms of parameters

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

→ 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)

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

⑨

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

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

(b) E+M based estimation:

E-step: →

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

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

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

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

We know how to compute this given set of parameters.

M-step: - Given $P(z^{(i)}=\ell)$, we want to estimate the parameters of the Naive Bayes models in the mixture

$$\phi = \frac{\sum_{i=1}^{m} 1\{y^{(i)}=1\}}{m}$$ → independent of $z^{(i)}$'s

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

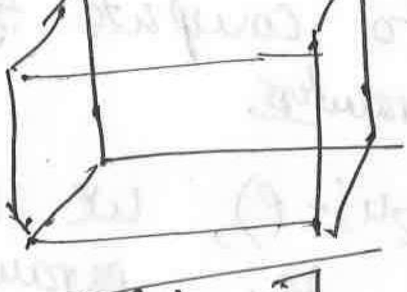
$$\theta_{j|y=t,z^{(i)}=l} = \frac{\sum_{i=1}^{m} P(z^{(i)}=l)\, 1\{y^{(i)}=t\}\, 1\{x^{(i)}=1\}}{\sum_{i=1}^{m} 1\{y^{(i)}=t\}\, P(z^{(i)}=l)}$$

softer version of Naive Bayes parameter estimation

[can do Laplace smoothing] by adding 1 to the numerator & 2 to the denominator

---

Q8. $H^3_{\|}$:-



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

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

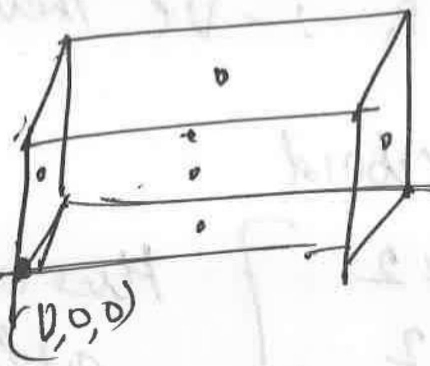
(10) First, we will show that $VC(H^3_{11}) = 6$

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

$\boxed{0 \le x \le 2, \quad 0 \le y \le 2, \quad 0 \le z \le 2}$

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

~~(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)~~



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)

Then these six points can be easily shattered by moving the plane at which the point lies by $\pm \varepsilon$ ($\varepsilon_e > 0$) to give the concerned point a label of +ve or -ve (or vice-versa) where ε_e is very small

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

the in if P_1 : +ve then

in the cuboid,

$-\varepsilon_e \le x \le 2$

$0 \le y \le 2$

$0 \le z \le 2$

Similarly if P_1 : -ve then

in the cuboid

$$\left.\begin{aligned} \varepsilon_e &\le x \le 2 \\ 0 &\le y \le 2 \\ 0 &\le z \le 2 \end{aligned}\right]$$ this does not affect the label of other points

$\Rightarrow$ All 2^6 labelings can be achieved by doing $\pm \varepsilon_e$ movement of 6 faces independently (corresponding to label of each point).

(11)

Next, we will show that $VC(H^3_{11}) = 6$

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

Assume contrary:-

$\exists\, P_1 \ldots P_7$ s.t. H^3_{11} can shatter them. We will show that $\exists$ 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}.

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

Assign p :- $-ve$ label

All other points $+ve$ label

Let $h \in H^3_{11}$ that achieves this labeling.

Clearly:- $P_x^{min}[x] \le P[x] \le P_x^{max}[x]$

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

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

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 ~~without~~ lying. 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 +ve label contradiction.

(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 considering 2 points on two faces $\parallel$ to each of n ~~axis~~ axis.

& then making the same argument as earlier

$2n+1$ points can not be shattered by:- assigning +ve label to those points which maximum value along any dimension & assigning -ve label to remaining point.