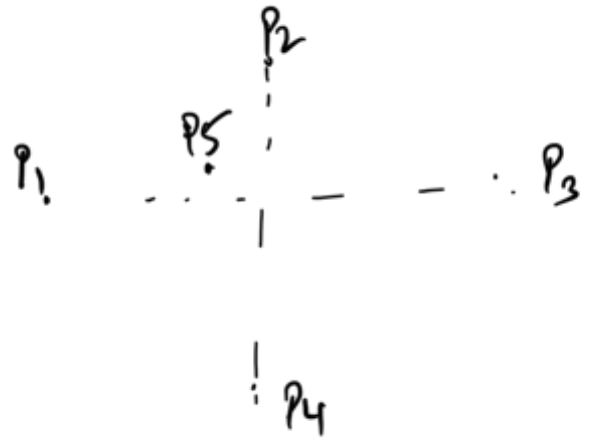
COL774
Major Solutions

Q.1. First we will show that VC-Dim($\mathcal{H}$) ≥ 5

Here $\mathcal{H} = \mathcal{H}^+ \cup \mathcal{H}^-$

Consider the following set of 5 points. (in 2-D)



There 32 possible different labelings & we need to show that all of those are achievable using $h \in \mathcal{H}$.

① Clearly all +ve & all -ve can be achieved. :- we can draw a rectangle $h \in \mathcal{H}^+$ enclosing all points. Similarly, all -ve is achievable since we can draw a rectangle $h \in \mathcal{H}^-$ enclosing all points.

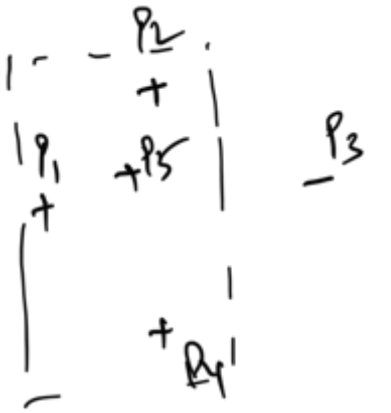
2 possibilities

② Consider where 4 of the point are +ve & 1 -ve.

↳ @ if the point P5 is -ve then we can draw a rectangle

1 possibility

5 pos



$\exists$ $h \in H^-$ around it & separately p_1, p_2, p_3, p_4.

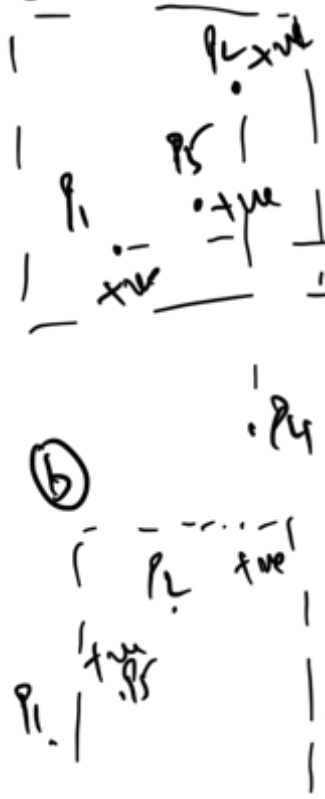
(b) if any other point is -ve example then by symmetry we can easily draw a rectangle like shown to the left for each such combination.

5 possibilities

can when 4 points are -ve & 1 +ve is identical to above with roles of H^+ & H^- reversed.

last:- Consider the can when 3 points are +ve & 2 -ve.

↳ 10 such possibility



First let us consider when p_5 is one of +ve points.

$\Rightarrow$ 2 of p_1 p_2 p_3 p_4 are +ve & 2 -ve

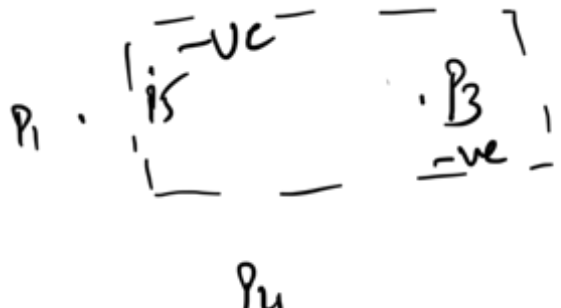
6 possibilities

see figure to left. We can find a $h \in H^+$ to achieve this labeling. Similarly, see the labeling to left. $\Rightarrow \exists h \in H^+$ to achieve the labeling

By similar argument, all
such possibilities can be
achieved [4 like fig ⓐ & 2
like fig ⓑ]

Consider:- 3+ve & 2-ve where
P5 is one of -ve points

⇒ 4 such possibilities



⇒ we can easily draw a
rectangular hyperbola
RECT enclosing P5 &
the other -ve point
like shown on left.
Similar argument holds for
any chosen negative point
among P1, P2, P3, P4.

⇒ All 16 possible labelings
when 3+ve & 2-ve can
be achieved by RECT

Finally:- Case when 3-ve & 2+ve
is identical to 3+ve & 2-ve
case with roles of R+ & R-
reversed

⇒ All 32 labelings can
be found by RECT.

be achieved by $\mathcal{H}$.

$\Rightarrow$ 3 set of points in 2-D P_1, P_2, P_3, P_4 & P_5 which can be shattered by $\mathcal{H}$.

$\Rightarrow$ VC-D($\mathcal{H}$) ≥ 5. — (A)

Next, we will show/argue that

VC-D($\mathcal{H}$) < 6

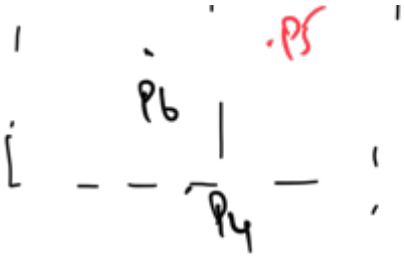
We will show that $\nexists \{P_1 -- P_6\}$ which $\in \mathbb{R}^2$ can be shattered by $\mathcal{H}$.

Argument :-

Consider some set of 6 points. We can safely assume that points are distinct. if not, ∴ clearly we can assign opposite labels to 2 points which not distinct & the points can not be shattered.

We will explain our argument with the help of following example

↳ Chose $P_1 P_3 P_4 P_2$ such that they correspond min-max value of x- coordinate & y-coordinate respectively. If the same



points has ... value along x or y direction, we will choose the next best points along that direction such that P_1 P_2 P_3 P_4 are distinct.

$P_1.x \le P_i.x$ $\forall i$
$P_3.x \ge P_i.x$ $\forall i$
$P_4.y \le P_i.x$ $\forall i \in [1,2,3]$
$P_2.y \ge P_i.y$ $\forall i \in [1,3]$
P_1 P_2 P_3 P_4 are distinct.

By construction:-
P_5 & P_6 lie inside the rectangle enclosing P_1, P_2, P_3, P_4.

Next we will construct a labeling as follows- our labeling will have 3 +ve points, 3 -ve points.

WLOG:- Assume.

$P_5.x \ge P_1.x$
$P_5.y \ge P_6.y$

if it was the case if $P_5.x \ge P_1.x$ $P_5.y \le P_1.y$ then in the following step we will choose points with max x value (R) & max y value i.e. P_5, P_1, P_2 +ve, P_6, P_3, P_4 -ve

then Assign:-

P_5, P_1, P_4 :- +ve
P_6, P_2, P_3 :- -ve

Clearly:-
Rectangle enclosing P_1, P_4, P_5 also contains P_6 :- & next which achieves desired labeling.

Similarly, rectangle enclosing

$P_6, P_2, \& P_3$ also contains 1s.
$\Rightarrow$ & $h \in \mathcal{H}$ which achieves desired labeling
$\Rightarrow$ $\exists$ $h \in \mathcal{H}$ which can achieve labeling constructed in the above manner.

∴ Since, $P_1, P_2, P_3, P_4, P_5, P_6$ were chosen generically, $\exists$ set of six points which can be shattered by $\mathcal{H}$. $\Rightarrow$ VC-Dim $(\mathcal{H}) \geq 6$ — (B)

Combining (A) & (B) we get

$$\boxed{\text{VC-Dim}(\mathcal{H}) = 5}$$

Q.2. This question is similar to question 2 in Minor examination.

(a) Logistic Regression loss:-

$$L(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left[y^{(i)} - h_\theta(\phi(x^{(i)})) \right]^2$$

$$h_\theta(\phi(x^{(i)})) = \frac{1}{1+e^{-\theta^T \cdot \phi(x^{(i)})}}$$ ← Sigmoid function applied to $\theta^T \phi(x^{(i)})$

(b) First let us look at the gradient:-

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

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

Note: If $\nabla_z g(z) = g(z)(1-g(z))$

$$= -\frac{1}{m} \sum_{i=1}^{m} \left[y^{(i)} - h_\theta(\phi(x^{(i)})) \right] g(\theta^T \phi(x^{(i)})) \left(1 - g(\theta^T \phi(x^{(i)}))\right) \cdot \nabla_\theta [\theta^T \phi(x^{(i)})]$$

g: sigmoid where $g(\theta^T \phi(x^{(i)})) = \frac{1}{1+e^{-\theta^T \phi(x^{(i)})}}$

Sigmoid function

$$\nabla_\theta \theta^T \phi(x^{(i)}) = \phi(x^{(i)})$$

$$\Rightarrow \nabla_\theta L(\theta) = \frac{-1}{m} \sum_{i=1}^{m} \left[y^{(i)} - h_\theta(\phi(x^{(i)})) \right] \left[h_\theta(\phi(x^{(i)})) \left[1 - h_\theta(\phi(x^{(i)}))\right] \phi(x^{(i)}) \right]$$

Now, $h_\theta(\phi(x^{(i)})) = \frac{1}{1+e^{-\theta^T \cdot \phi(x^{(i)})}}$

Let us assume that θ can be represented as coefficients $\alpha_1 \cdots \alpha_m$ of data points $x^{(1)}, x^{(2)} \cdots x^{(m)}$

then $\theta = \sum_{i=1}^{m} \alpha_i \cdot \phi(x^{(i)})$ [And we will show that this assumption is

satisfied during the entire learning process] → see below

At iteration t of learning

$$\theta^{(t)} = \sum_{i=1}^{m} \alpha_i^{(t)} \phi(x^{(i)})$$

Clearly At $t=0$, $\alpha_1 \dots \alpha_m = 0$

Since $\theta^{(0)} = \vec{0}$

Next, we will show that this holds using an induction argument.

Learning update:-

$t=0$

$\theta^{(t)} \leftarrow$ init ();

do {

$\theta^{(t+1)} \leftarrow \theta^{(t)} - \eta \cdot \nabla_\theta LL(\theta)\big|_{\theta^{(t)}}$

$t \leftarrow t+1$;

} (until ! converged)

$\nabla_\theta LL(\theta)$

$= \frac{1}{m} \sum_{i=1}^{m} [y^{(i)} - h_\theta(\phi(x^{(i)}))]$

$\cdot h_\theta(\phi(x^{(i)}))$

$(1 - h_\theta(\phi(x^{(i)})))\, \phi(x^{(i)})$

Assume our assumption holds at iteration t: Then, at $t+1$:-

(it clearly holds at $t=0$)

$$\theta^{(t+1)} = \theta^{(t)} + \eta \cdot \sum_{i=1}^{m} \left[y^{(i)} - h_\theta(\phi(x^{(i)})) \right] h_\theta(\phi(x^{(i)})) \left[1 - h_\theta(\phi(x^{(i)})) \right] \phi(x^{(i)})$$

By inductive assumption:-

$$\theta^{(t)} = \sum_{i=1}^{m} \alpha_i^t \, \phi(x^{(i)})$$

using a different variable

$$\theta^{(t)T} \cdot \phi(x^{(i)}) = \left[\sum_{e=1}^{m} \alpha_e^{(t)} \phi(x^{(e)}) \right]^T \cdot \phi(x^{(i)})$$

$$= \sum_{e=1}^{m} \alpha_e^{(t)} \left[\phi(x^{(e)})^T \phi(x^{(i)}) \right]$$

can be computed using only dot products

$$= \sum_{e=1}^{m} \alpha_e^t \, K(x^{(e)}, x^{(i)}) \equiv \beta_i^t$$

[Another name for this expression]

$$\Rightarrow \theta^{(t+1)} \leftarrow \theta^t + \eta \sum_{i=1}^{m} \left[y^{(i)} - \frac{1}{1+e^{-\beta_i t}} \right] \left[\frac{1}{1+e^{-\beta_i t}} \right] \left[\frac{1}{1+e^{\beta_i t}} \right] \cdot \phi(x^{(i)})$$

$$\gamma_i^t \equiv \left[y^{(i)} - \frac{1}{1+e^{-\beta_i t}} \right] \left[\frac{1}{1+e^{-\beta_i t}} \right] \left[\frac{1}{1+e^{\beta_i t}} \right]$$

$$\Rightarrow \theta^{(t+1)} \leftarrow \theta^{(t)} + \eta \cdot \frac{1}{m} \sum_{i=1}^{m} \gamma_i^t \, \phi(x^{(i)})$$

Using inductive assumption

$$\theta^{(t)} = \sum_{i=1}^{m} \alpha_i^t \phi(x^{(i)})$$

$$\Rightarrow \theta^{(t+1)} \leftarrow \sum_{i=1}^{m} (\alpha_i^t + \eta \cdot \gamma_i^t)\phi(x^{(i)})$$

α_i^{t+1} define it this way.

$$= \sum_{i=1}^{m} \alpha_i^{t+1} \phi(x^{(i)})$$

Hence, $\theta^{(t+1)}$ can be expressed as a linear combination

Proof of induction step

$$\sum_{i=1}^{m} \alpha_i^{(t+1)} \phi(x^{(i)})$$

where $\alpha_i^{(t+1)} = \alpha_i^{(t)} + \frac{\eta}{m} \gamma_i^t$

where γ_i^t is as defined earlier & can be computed purely in terms of dot products

~~$\Rightarrow$ Entire learning algorithm~~

Since we only needed to compute $K(\cdot,\cdot)$ during the entire process without explicitly involving $\phi(x^{(i)})$'s in

computation ⇒ entire learning algorithm can be implemented in terms of kernel operations.

(c) Practically:- Assuming $k(x,z)$ can be computed efficiently ($O(n)$ time)

⇒ entire learning update can be done in $O(n)$ time as opposed to $O(N)$ if we were to represent $\phi(x)$'s explicitly to perform computations over them directly.

Thus can be a huge saving if $N >> n$.

Q3. (a) f is convex if

$\forall \; x^1, x^2 \in R^n$

$$f(\underbrace{\alpha x^1 + (1-\alpha) x^2}_{\text{where } \alpha \in [0,1]}) \leq \alpha f(x^1) + (1-\alpha) x^2$$

convex combination of x^1 & x^2.

(b) We will show the required inequality holds using induction over k.

If $k=2$, this holds (using part (a))

[Also it is true for $k=1$
since $f(1 \cdot x^1) = 1 \cdot f(x^1)$]

Assume it holds for $(k-1)$ consider (k)

LHS:-

$$\sum_{i=1}^{k} \alpha_i f(x^{(i)}) = \sum_{i=1}^{(k-1)} \alpha_i f(x^{(i)}) + \alpha_k f(x^{(k)})$$

$$= \left[\sum_{i=1}^{(k-1)} \alpha_i f(x^{(i)}) \right] \cdot \frac{1 \cdot \left[\sum_{i=1}^{k-1} \alpha_i \right]}{\sum_{i=1}^{k-1} \alpha_i} + \alpha_k f(x^{(k)})$$

Multiply & devide by this quantity

$\equiv$ ~~Let~~ Term 1

$$\sum_{i=1}^{k-1} \alpha_i \left[\sum_{i=1}^{k-1} \frac{\alpha_i}{\left[\sum_{i=1}^{k-1} \alpha_i\right]} f(x^{(i)}) \right] + \alpha_k f(x^{(k)})$$

Let $\frac{\alpha_i}{\sum_{i=1}^{k-1} \alpha_i} = \beta_i \Rightarrow \sum_{i=1}^{k-1} \beta_i = 1$

$\Rightarrow$ Applying Jensen's inequality in Term 1

$$\left[\sum_{i=1}^{k-1} \alpha_i \right] \sum_{i=1}^{k-1} \beta_i f(x^{(i)}) + \alpha_k f(x^{(k)})$$

$$\geq \left[\sum_{i=1}^{k-1} \alpha_i \right] f\left(\sum_{i=1}^{k-1} \beta_i x^{(i)} \right) + \alpha_k f(x^{(k)})$$

$$\text{RHS} = \left[\sum_{i=1}^{k-1} \alpha_i\right] f\left\{\frac{\sum_{i=1}^{k-1} [\alpha_i f(x^{(i)})]}{\left[\sum_{i=1}^{k-1} \alpha_i\right]}\right\} + \alpha_k f(x^{(k)})$$

Let $\sum_{i=1}^{k-1} \beta_i x^{(i)} = \underline{x'}$ where $\beta_i = \frac{\alpha_i}{\sum_{i=1}^{k-1} \alpha_i}$

& $\sum_{i=1}^{k-1} \alpha_i = \alpha'$

Note that $\boxed{\alpha' + \alpha_k = 1}$ $4\alpha\alpha' \leq 1$

$0 \leq \alpha_k \leq 1$

$$\text{RHS} = \alpha' f(x') + \alpha_k f(x^{(k)})$$

Applying Jensen's inequality again

$$\text{RHS} \geq f(x'\alpha' + \alpha_k x^{(k)})$$

$$= f\left(\sum_{i=1}^{k-1} \frac{x^{(i)} \alpha_i}{\sum_{i=1}^{(k-1)} \alpha_i} \cdot \sum_{i=1}^{(k-1)} \alpha_i + \alpha_k x^{(k)}\right)$$

$$= f\left(\sum_{i=1}^{k-1} x^{(i)} \alpha_i + \alpha_k x^{(k)}\right)$$

$$= f\left(\sum_{i=1}^{k} \alpha_i x^{(i)}\right)$$

$\Rightarrow$ ∎

$$\text{LHS} = \sum_{i=1}^{k} \alpha_i f(x^{(i)}) \geq f\left(\sum_{i=1}^{k} \alpha_i x^{(i)}\right) = \text{RHS}$$

Hence the original

equation)

Hence, proved

(c) Jensen's inequality:- f:- convex

$$f\left(E_p[x]\right) \leq E_p[f(x)]$$

under a given distribution $P(X=x)$

We will first prove this when $x \in R^1$.

Let $p(x=x) \geq 0$ when $x \in (a,b)$

we will eventually take the limit $a \to -\infty$, $b \to \infty$.

Given: $x^{(1)} \ldots x^{(k)}$ equally spaced

Then, let $\Delta x = \frac{b-a}{k}$

Let $x^{(0)} = a$ $\quad x^{(i)} = a + i\Delta x$

$$\alpha_i = P(x^{(i-1)} \leq X \leq x^{(i)})$$

Then using Jensen's inequality

$$\sum_{i=1}^{k} \alpha_i f(x^{(i)}) \geq \sum_{i=1}^{k} f(\alpha_i x^{(i)})$$

where α_i as given above

$x^{(i-1)} + \Delta x = x^{(i)}$ $\forall i$

$$\sum_{i=1}^{k} P(x^{(i-1)} \leq x \leq x^{(i+1)} + \Delta x)\, f(x^{(i)})$$

$$\geq f\left(\sum_{i=1}^{k} P(x^{(i-1)} \leq x \leq x^{(i-1)} + \Delta x \cdot x^{(0)}\right.$$

Take $\lim k \to \infty$,

$$\Delta x = \frac{b-a}{k} \to 0$$

In the limit $P(x^{(i-1)} \le x \le x^{(i-1)} + \Delta x)$ can be written as:-

$$P_d(x^{(i)})\,\Delta x$$

↓ density at $x^{(i)}$

Sum becomes integral

$$\int_a^b P_d(x) f(x)\,dx \ge f \int_a^b P_d(x)\,dx$$

$$\Rightarrow \boxed{E_{P(x)}[f(x)] \ge f\left(E_{P(x)}[x]\right)}$$

This holds for any given value of a & b

In particular we can keep

$$a \to -\infty, \quad b \to \infty.$$

Finally when $x \in R^n$

Then intervals have to be defined along each dimension:-

Δx corresponds to the interval between

$x_j^{(i+1)}$ & $x_j^{(i)}$ in each dimension

Jensen's inequality becomes:- We assume k^n such intervals (one Δx along each dimension)

$$\sum_{i=1}^{k} \cdots \sum_{j=1}^{k} \left[\prod_{j=1}^{n} P\left(x_j^{(i_j-1)} \le x_j \le x_j^{(i_j-1)} + \Delta x_j\right) \right] f\left(x_1^{(i_1)}, \ldots, x_n^{(i_n)}\right)$$

$$\geq f\left(\sum_{i_1=1}^{k} \cdots \sum_{i_n=1}^{k} \left(\prod_{j=1}^{n} P\left(x_j^{(i_j-1)} \leq x_j \leq x_j^{(i_j+\Delta x_j)}\right)\right)\right.$$

$$\left.\left(x_1^{(i_1)} \cdots x_n^{(i_n)}\right)\right)$$

$$\Rightarrow \sum_{i_1=1}^{k} \cdots \sum_{i_n=1}^{k} \prod_{j=1}^{n} P\left(x_j^{(i_j-1)} \leq x_j \leq x_j^{(i_j-1)} + \Delta x_j\right) f(x^{(i)})$$

Becomes integral in the limit

$$\geq f\left(\int_{i_1=1}^{k} \cdots \int_{i_n=1}^{k} \prod_{j=1}^{n} P\left(x_j^{(i_j-1)} \leq x_j \leq x_j^{(i_j-1)} + \Delta x_j\right) x^{(i)}\right)$$

$$\Rightarrow \int_{x_1=a_1}^{b_1} \cdots \int_{x_n=a_n}^{b_n} P_d(x_1 \cdots x_n) \cdot dx_1 \cdots dx_n \Big] f(x^{(i)})$$

$$\geq f\left(\int_{x_1=a_1}^{b_1} \int_{x_2=a_2}^{b} \cdots \int_{x_n=a_n}^{b_n} P_d(x_1 \cdots x_n) \, dx_1 \cdots dx_n\right)$$

$$\Rightarrow E_{P(x)}[f(x)] \geq f_{P(x)}\left(E[x]\right)$$

We can take $\lim a_1 \cdots a_n \to -\infty$ & $b_1 \cdots b_n \to \infty$

To argue for the general case.

Q.4. Unconstrained SVMs:-

$$\min_{w,b,\xi} \frac{1}{2} w^T w + c \sum_{i=1}^{m} \xi_i$$

$$y^{(i)}\left(w^T x^{(i)} + b\right) \geq 1 - \xi_i \quad \forall i \quad -①$$

$$\xi_i \geq 0 \quad \forall i \quad -②$$

set of equations in ① we get

from sl

$$\xi_i \geq 1 - y^{(i)}(w^T x^{(i)} + b)$$

& (2) tells us that

$$\xi_i \geq 0$$

Together these equations tell us that

$$\xi_i \geq \max(0, 1 - y^{(i)}(w^T x^{(i)} + b))$$

Also, since we are optimizing wrt ξ_i s.t.

$$\min_{w,b,\xi_i} \frac{1}{2} w^T w + C \sum_{i=1}^{m} \xi_i$$

it must be the case that

$$\xi_i = \max(0, 1 - y^{(i)}(w^T x^{(i)} + b)) \quad \forall i$$

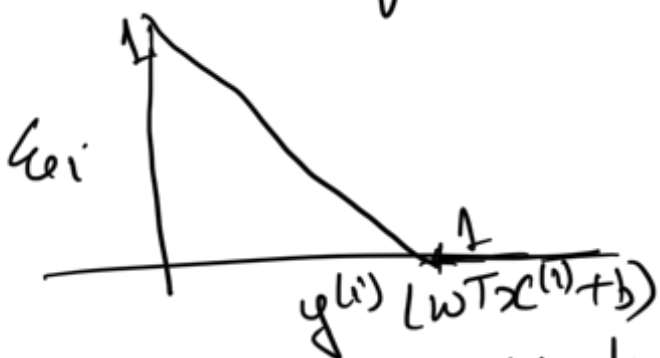
(since we need to minimize wrt ξ_i's)

$$\Rightarrow \xi_i = 1 - y^{(i)}(w^T x^{(i)} + b) \text{ if } y^{(i)}(w^T x^{(i)} + b) \leq 1$$

$$0 \quad \text{o.w.}$$

$\Rightarrow$ Hinge loss wrt $y^{(i)}(w^T x^{(i)} + b)$

$$L_H(y^{(i)}(w^T x^{(i)} + b))$$



$\Rightarrow$ unconstrained objective can be written as-

(obtained by substituting the value for ξ_i)

Lobjective

$$\min_{w,b} \quad \tfrac{1}{2} w^T w + C \sum_{i=1}^{m} L_H(z_i)$$

$$z_i = y^{(i)}(w^T x^{(i)} + b)$$

unconstrained loss

$$L_H(z_i) = \begin{cases} 1 - z_i & \text{if } z_i \le 1 \\ 0 & \text{o.w.} \end{cases}$$

$L_u(w,b)$

$$\Rightarrow \min_{w,b} \quad \tfrac{1}{2} w^T w + C \sum_{i=1}^{m} \max(0, 1 - y^{(i)}(w^T x^{(i)} + b)) \quad \text{— (3)}$$

Intuitively, the loss term is non zero if point is inside the (corresponding) margin boundary

(b) Let us differentiate wrt w

$$\nabla_w \; \tfrac{1}{2} w^T w + C \sum_{i=1}^{m} \max(0, 1 - y^{(i)}(w^T x^{(i)} + b))$$

$$= w + C \sum_{i=1}^{m} \nabla_w \max(0, 1 - y^{(i)}(w^T x^{(i)} + b))$$

if $y^{(i)}(w^T x^{(i)} + b) > 1$: 0

if $y^{(i)}(w^T x^{(i)} + b) = 1$: $C[-y^{(i)} x^{(i)}]$ (using right derivative at the point of discontinuity)

if $y^{(i)}(w^T x^{(i)} + b) < 1$: $-y^{(i)} x^{(i)}$

$$\nabla_b \left[\tfrac{1}{2} w^T w\right] + C \sum_{i=1}^{m} \nabla_b \max(0, 1 - y^{(i)}(w^T x^{(i)} + b))$$

$$= 0 + C \sum_{i=1}^{m} \nabla_b \left[\max(0, 1 - y^{(i)}(w^T x^{(i)} + b))\right]$$

$y^{(i)}(w^T x^{(i)} + b) > 1$	$y^{(i)}(w^T x^{(i)} + b) = 1$	$y^{(i)}(w^T x^{(i)} + b) < 1$
$\forall i :-$ 0	0 (wrong Right derivative)	$C[-y^{(i)}]$

Gradient descent update:-

$$\begin{cases} w^{(t+1)} \leftarrow w^{(t)} - \eta \cdot \nabla_w L_u(w, b) \\ b^{(t+1)} \leftarrow b^{(t)} - \eta \cdot \nabla_b L_u(w, b) \end{cases}$$

where gradients are as defined above.

Note :- loss is -ve of $LL(\theta)$. We will show $LL(\theta)$ is concave $\Rightarrow$ loss is convex.

Q.5:

Logistic Regression:-

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}} \equiv P(y = 1 | x; \theta)$$

$$LL(\theta) = \sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log (1 - h_\theta(x^{(i)}))$$

$$\nabla_\theta LL(\theta) = \sum_{i=1}^{m} \frac{y^{(i)}}{h_\theta(x^{(i)})} \cdot \nabla_\theta h_\theta(x^{(i)}) + \frac{(1 - y^{(i)})}{1 - h_\theta(x^{(i)})} \nabla_\theta (-h_\theta(x^{(i)}))$$

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

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

$$= \sum_{i=1}^{m} x^{(i)} (y^{(i)} - h_\theta(x^{(i)}))$$

$$\nabla^2_\theta L(\theta) = \sum_{i=1}^{m} h_\theta(x^{(i)})\left(1-h_\theta(x^{(i)})\right)\left[x^{(i)}x^{(i)T}\right] \quad (-1)$$

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

$\equiv$ Hessian matrix

For concavity, we need to show that H is -ve semi-definite $\Rightarrow$

$$z^T H z \le 0 \qquad z \in \mathbb{R}^{(n+1)} \left[\theta \in \mathbb{R}^{(n+1)}\right]$$

$$\Rightarrow z^T\left[-\sum_{i=1}^{m}\left[h_\theta(x^{(i)})\right]\left[1-h_\theta(x^{(i)})\right] x^{(i)} x^{(i)T}\right] z$$

$$\Rightarrow -\sum_{i=1}^{m} h_\theta(x^{(i)})\left(1-h_\theta(x^{(i)})\right) z^T x^{(i)} x^{(i)T} z$$

$$= -\sum_{i=1}^{m} h_\theta(x^{(i)})\left(1-h_\theta(x^{(i)})\right)\underbrace{\left[(x^{(i)T}z)^T(x^{(i)T}z)\right]}_{\le 0}$$

$$\Rightarrow -\sum_{i=1}^{m} \underbrace{h_\theta(x^{(i)})}_{\ge 0}\underbrace{\left(1-h_\theta(x^{(i)})\right)}_{\ge 0}\underbrace{(x^{(i)T}z)^T(x^{(i)T}z)}_{\ge 0} \le 0$$

$\Rightarrow$ Hessian matrix is -ve semi-definite

$\Rightarrow$ $L(\theta)$ is concave. $\Rightarrow$ Loss function $= -L(\theta)$ is convex.

$\nabla_\theta L(\theta)$

Q.6. (a)

$$P(X) = \frac{1}{Z} \prod_k \phi_k(X_{C_k})$$

$X_i \in \{0,1\}$ $\forall i$. (Nodes represent Boolean valued variables in the given setting)

⇒ Since probabilities should add up to 1

$$\sum_{X \in \{0,1\}^n} \frac{1}{Z} \prod_k \phi_k(X_{C_k}) = 1$$

$X_1 \in \{0,1\}, X_2 \in \{0,1\}$ -- $X_n \in \{0,1\}$

$$Z = \sum_{x \in \{0,1\}^n} \left[\prod_k \phi_k(X_{C_k} = x_{C_k}) \right]$$

Here x_{C_k} denote the value of x restricted to variables in X_{C_k}

(b) Recall that the distribution defined by a Bayesian network is given as:-

$$P(X) = \prod_{i=1}^{n} P(X_i \mid Pa(X_i))$$

Now, given a Bayesian network $\mathcal{B}$, let us define a Markov corresponding network where the factors

Markov ... involve a variable and ... & there are as many factors as there are variables then

under equivalent Markov network

$$P(X) = \frac{1}{Z} \prod_{j=1}^{n} \phi_j(X_j, Pa(X_j))$$

where $Z = 1$

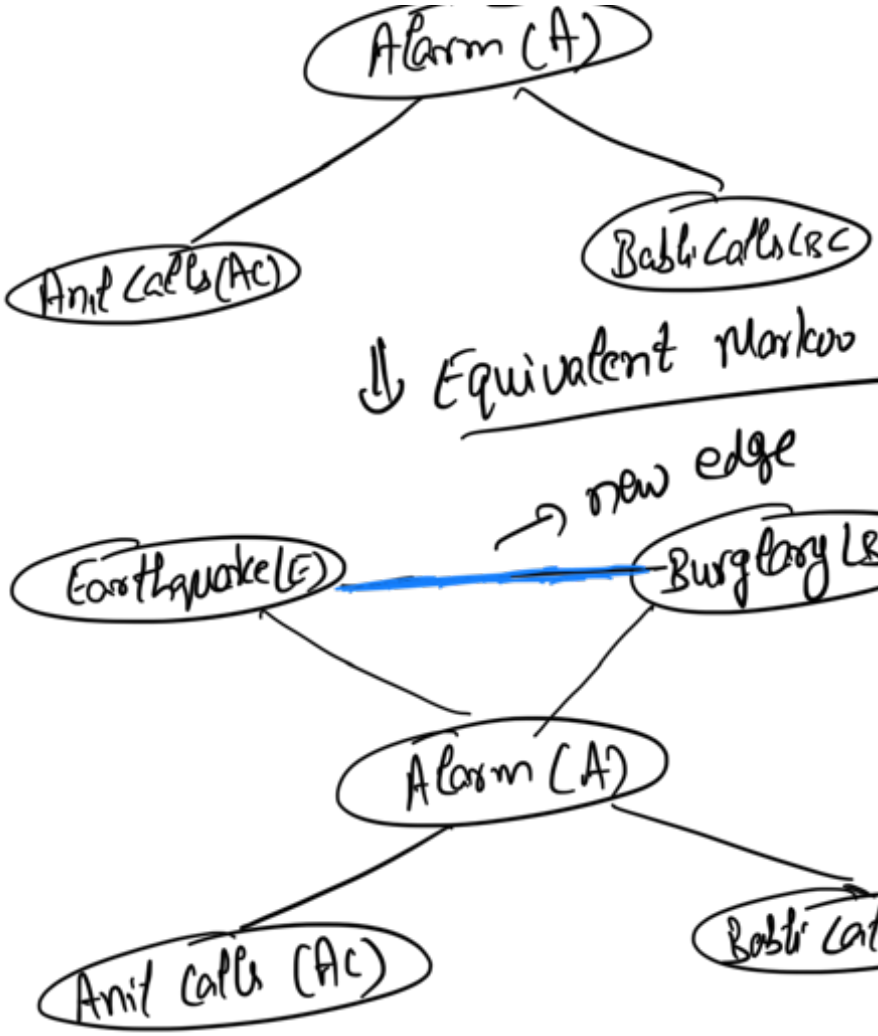
$$\phi_j(X_j, Pa(X_j)) = P(X_j | Pa(X_j))$$

comes from the parameters of CPTs in the Bayesian network

In particular, the equivalent Markov network graph is obtained by connecting all the parents of each node (directionality) with each other & dropping the arrows in Bayesian network arcs.

[since factors are defined over cliques & above procedure results in desired set of cliques.

Example (for illustration only):-



Q.1. (a) $\min_{w,b} \frac{1}{2} w^T w + C \sum_{i=1}^{m} \xi_i$

$$y^{(i)} (w^T x^{(i)} + b) \geq 1 - \xi_i \quad \forall i$$

$$\xi_i \geq 0 \quad \forall i$$

In the limit $C \to \infty$, model strictly adhere to data, disallowing any form of noise. Any point outside the margin boundary is given a very high penalty (which goes to infinity

Therefore, in the limit $C \to \infty$

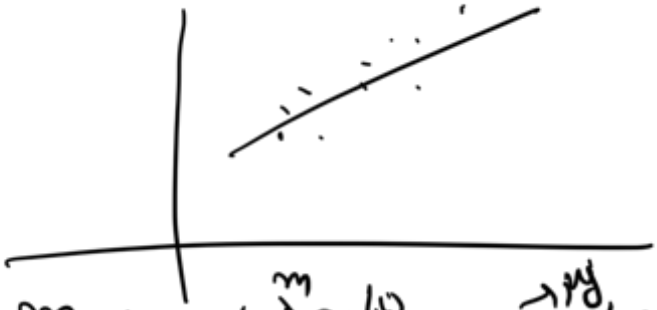
as C → ∞. ...
model is the more likely to overfit the data. (Since it really tries hard to fit every point on the "right" side of margin, even if it happened to be an outlier) For smaller values of C, this is not the case & model can softly deal with noise (i.e; by paying a finite penalty).

(b) Since both training & testing errors are low, this is the case of underfitting [the model is not able to fit even the patterns observable in training data as evident from high training error]

As a remedy to underfitting as described above, we would try a more sophisticated model first. Giving more training data is not going to help since the model may not simply be able to fit this additional amount of data. A more sophisticated model may ~~reduce~~ reduce the training errors by being able to ~~fit~~ fit better fit patterns seen in the training data,

(c) Decision trees are unstable learners since minor changes in the training distribution can lead to substantial changes in the learned tree by altering the choice of a node/attribute to split on at a particular level in the tree growing phase.
On the other hand, Naive Bayes is a relatively stable learner since minor changes in the training distribution will only result in minor changes in the learned probabilities (parameters) of the underlying model.

Q.8. Principal Component Analysis (PCA)



PCA tries to maximize the variance of the projected data

$\{x^{(i)}\}_{i=1}^{m}$ $\forall j$:- $\frac{1}{m}\sum_{i=1}^{m} x_j^{(i)} = 0 \rightarrow \mu_j$ (mean is zero)

$\forall j$:- $\frac{1}{m}\sum_{i=1}^{m}(x_j^{(i)} - 0)^2 = 1 \rightarrow \sigma_j^2$ (variance is 1) given

Let the projected dimensions be $u_1 \ldots u_k$.

Then Projection of $x^{(i)}$ along the direction u_e is given as:- $x^{(i)T} u_e$

Note since the original data is 0 mean, projected data is also zero mean

$$\frac{1}{m}\sum_{i=1}^{m}[x^{(i)T}u_\ell] = \frac{1}{m}\sum_{i=1}^{m}x^{(i)T}u_\ell = 0^T u_\ell = \underline{0}$$

Total variance of the projected data is -

$$\sum_{u_1 \cdots u_k}\left[\frac{1}{m}\sum_{i=1}^{m}[x^{(i)T}u_\ell]^2\right] \quad \left[\text{i.e. sum of variances along each of the projected directions}\right].$$

$$= \frac{1}{m}\sum_{u_1 \cdots u_k}\sum_{i=1}^{m}\left[u_\ell^T x^{(i)}\right]\left[x^{(i)T}u_\ell\right]$$

$$= \frac{1}{m}\sum_{u_1 \cdots u_k}\sum_{i=1}^{m} u_\ell^T\left[x^{(i)}x^{(i)T}\right]u_\ell$$

$$= \sum_{u_1 \cdots u_k} u_\ell^T \underbrace{\frac{1}{m}\left[\sum_{i=1}^{m} x^{(i)}x^{(i)T}\right]}_{\hat{\Sigma}:\text{- Empirical co-variance matrix}} u_\ell$$

$$= \sum_{u_1 \cdots u_k}\left[u_\ell^T \hat{\Sigma} u_\ell\right] \quad :\text{- Projected variance.}$$

Therefore the goal in PCA is to find $u_1 \cdots u_k$ such that

$$\sum_{u_1 \cdots u_k} u_\ell^T \hat{\Sigma} u_\ell \quad \text{is maximized}$$

$$\arg\max\left(u_\ell^T \hat{\Sigma} u_\ell\right) \longrightarrow \text{Empirical co-variance matrix}$$

$u_1 - u_k$

subject to:- $\|u_\ell\|^2 = 1$ i.e. $u_\ell^T u_\ell = 1 \;\forall \ell$

& $u_{\ell_1}^T u_{\ell_2} = 0 \quad \ell_1 \neq \ell_2$

(orthonormality of $u_1 \cdots u_k$)

(b)

First let us solve for $k=1$

$$\max_{u_1} \; u_1^T \hat{\Sigma} u_1$$

$$u_1^T u_1 = 1$$

Lagrangian:-

$$L(u, \alpha_1) = u_1^T \hat{\Sigma} u_1 + \alpha_1 (1 - u_1^T u_1)$$

using KKT (necessary condition for optimality)

$$\nabla_u L(u, \alpha_1) = 0$$

$$\Rightarrow \quad 2\hat{\Sigma} u_1 - 2\alpha_1 u_1 = 0$$

$$\Rightarrow \quad \boxed{\hat{\Sigma} u_1 = \alpha_1 u_1}$$

$\Rightarrow$ u_1 is eigenvector of $\hat{\Sigma}$.

Also, since we want to maximize

$$u_1^T [\hat{\Sigma} u_1] = u_1^T \alpha_1 u_1 = \alpha_1$$

α_1: must correspond to highest eigenvalue

:- u_1 is eigenvector corresponding to highest eigenvalue

② $K=2$. We want to maximize

$$[\; u_2^T \hat{\Sigma} u_2 \;]$$

$$\max_{u_1, u_2} \; u_1^T \hat{\Sigma} u_1 + u_2^T \hat{\Sigma} u_2$$

subject to constraint $u_1^T u_1 = 1$

$u_2^T u_1 = 2$

$\Leftarrow$ $u_1^T u_2 = 0$

Aside

Empirical covariance between $u_1^T x$ & $u_2^T x$ is

$$\frac{1}{m} \sum_{i=1}^{m} u_2^T x^{(i)} x^{(i)T} u_1$$

$$= \frac{1}{m} \sum_{i=1}^{m} u_2^T \left(x^{(i)} x^{(i)T} u_1 \right)$$

$$= \sum_{i=1}^{m} u_2^T \hat{\Sigma} u_1$$

$$= \sum_{i=1}^{m} u_2^T u_1 = 0$$

will be used later

$$\mathcal{L}(u_1, u_2, \alpha_1, \alpha_2, \beta) = u_1^T \hat{\Sigma} u_1 + u_2^T \hat{\Sigma} u_2 + \alpha_1 (1 - u_1^T u_1) + \alpha_2 (1 - u_2^T u_2) + \beta u_1^T u_2$$

$$\nabla_{u_2} \mathcal{L}(u_1, u_2, \alpha_1, \alpha_2, \beta) = 0$$

[using KKT]

$$= 2 \hat{\Sigma} u_2 + 2\alpha_2 (-u_2) + \beta u_1 = 0$$

$\Rightarrow$ Pre-multiplying by u_1^T

$$\underbrace{2 u_1^T \hat{\Sigma} u_2}_{\text{Empirical covariance between projections along } u_1 \& u_2 = 0} + \underbrace{\alpha_2 u_1^T (-u_2)}_{0} + \underbrace{\beta (-u_1^T) u_1}_{-1} = 0$$

$$\Rightarrow \beta = 0$$

$$\Rightarrow 2 \hat{\Sigma} u_2 = 2\alpha_2 u_2 \Rightarrow \boxed{\hat{\Sigma} u_2 = \alpha_2 u_2}$$

u_2 corresponds to eigenvector of $\hat{\Sigma}$

$\Rightarrow u_2$ [illegible]

further we want to maximize

$$\max_{u_1, u_2} \left[u_1^T \Sigma u_1 + u_2^T \Sigma u_2 \right]$$

$$= \max_{u_1, u_2} \left[\underbrace{d_1 + d_2}_{\Downarrow} \right]$$

must correspond to highest eigenvalues of Σ_0.
Empirical co-variance matrix.

Q 9. Decision Tree pruning :-

Let

T_P :- Tree obtained by greedy pruning strategy

T_{best} :- best tree obtained any given pruning strategy. Let T_{best} be minimal s.t such tree i.e. no more nodes in T_{best} can be pruned.

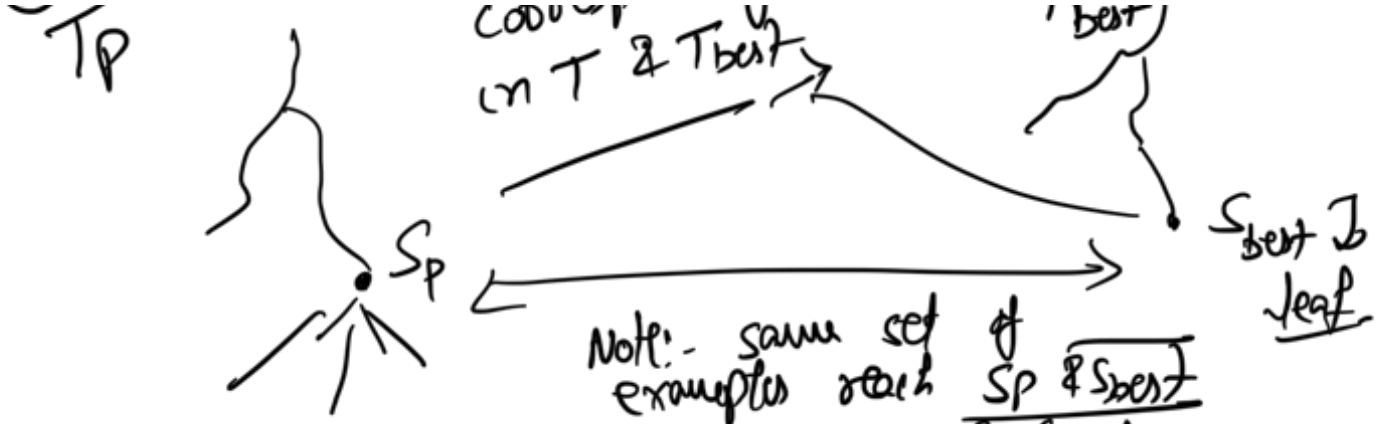
Let $\Rightarrow$ Acc (T_{best}) > Acc (T_P)

We will prove by contradiction that above is not possible

Consider some sub-tree in T_P which was pruned in T_{best} :-

corresponding sub-trees

(A)



Then let us consider Sp & Sbest

if Aval(Sbest) > Aval(Sp)

Then, clearly Tp is not final tree after following greedy strategy because we can prune tree below Sp & get more accuracy on validation set.

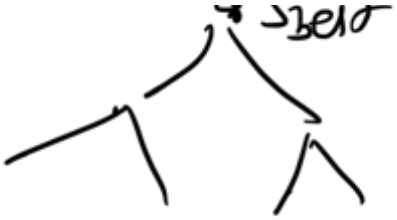
$\Rightarrow$ Aval(Sbest) $\leq$ Aval(Sp)

Since Sbest is ~~best~~ part of best possible & tree Tbest it must be the case that

Aval(Sbest) = Aval(Sp)

(B)

Next, let us consider subtree such that it is pruned in Tp but not in Tbest



Then let Aval (Sbest) > Aval (SP)

But this is not possible since

Claim:- If in greedy strategy if any subtree SP is pruned, then it must be the case that pruning any of its children would have also resulted in no loss in accuracy. If not :- ϕ

$$\Delta Sp = \sum_{l=1}^{k} \Delta Sl \cdot Wl$$

If $\Delta Sl < 0$ for some l:-

$\exists\, l'$ s.t

$\boxed{\Delta Sl > \Delta Sp}$

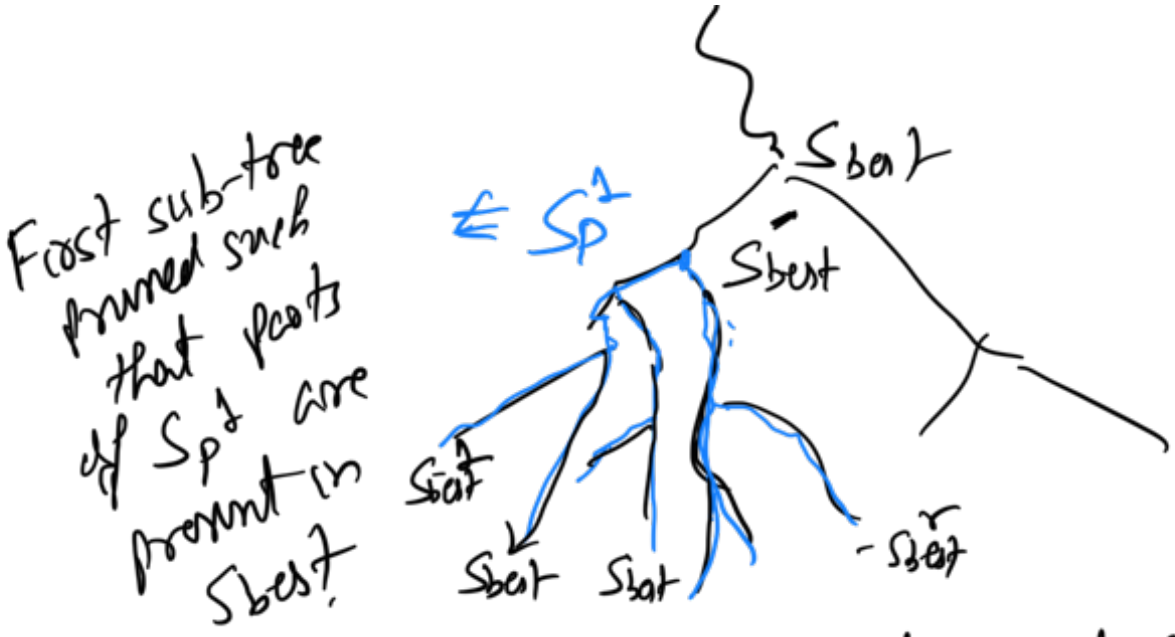
$0 \le Wl \le 1$

↓ weight of that branch

If any of its children result in loss in accuracy by pruning them, then clearly one of its other children must give more increase in accuracy by pruning it

This is true for every intermediate tree observed during pruning

Let us consider



Then:- clearly recursively speaking

Δ Tval(S_p^2) is at least as good as ΔTval(S_{best}^k) $1 \le k \le r$

if not:- we would have pruned that sub-tree fast $\Rightarrow$ contradiction.

$\Rightarrow$ Aval(S_p^2) $\geq$ Aval(BST')

but since S_{bad} is best possible

$\rightarrow$ Aval(S_p^2) = Aval(S_{best}')

$\Rightarrow$ we can get T_{best}' s.t.

tree below S_{best}' is pruned

& Aval(T_{best}) = Aval(T_{best}')

$\Rightarrow$ S_{best} is not minimal

⇒ contradiction.

∄ S_p & S_{best} s.t. S_p is pruned by S_{best} is not

⇒ All (differing) sub-trees of T_p are either identical to T_{best} or have same accuracy. contradiction

⇒ $Aval(T_p) = Aval(T_{best})$