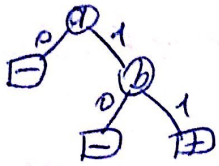
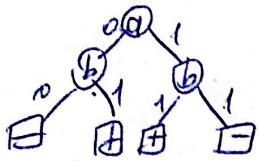
α $f(A,B) = A \vee B$

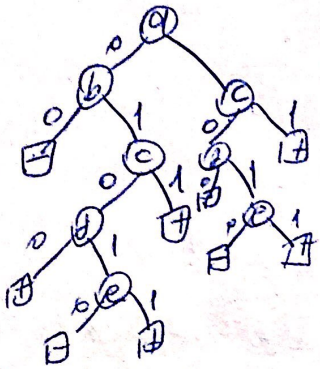
$f(A,B) = A \wedge B$



$f(A,B) = A \text{ xor } B$



$f(A,B,C,D,E) = (A \vee B) \wedge (C \vee \neg D \vee E)$



α TDIDT algo ?

Gain (S, outlook) = $H(\frac{6}{11}, \frac{5}{11}) - \left[\frac{4}{11} H(\frac{2}{4}, \frac{2}{4}) + \frac{3}{11} H(\frac{2}{3}, \frac{1}{3}) + \frac{4}{11} H(\frac{1}{4}, \frac{3}{4})\right]$

sunny — overcast — rainy

$\left[\frac{6}{11} \cdot \log_2 \frac{1}{\frac{6}{11}} + \frac{5}{11} \cdot \log_2 \frac{1}{\frac{5}{11}}\right] - \left[\frac{4}{11}\left(\frac{2}{4}\log_2 \frac{1}{\frac{2}{4}} + \frac{2}{4}\log_2 \frac{1}{\frac{2}{4}}\right) + \frac{3}{11} + \cdots\right] = 0.936$

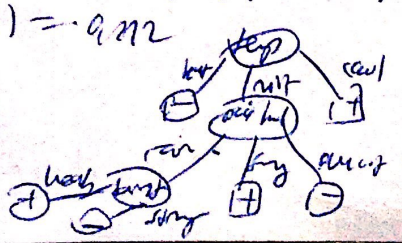
Gain (S, temp.) = 0.6289

Gain (S, wind) = 0.0278

calculer gain de temperature et wind ?

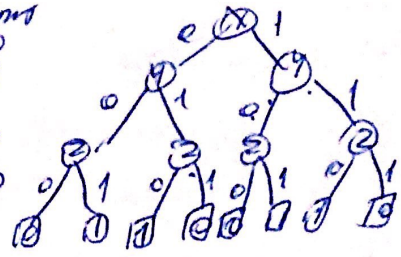
G(S, outlook) = $H(\frac{2}{5}, \frac{3}{5}) - \left[\frac{4}{5} H(0,1) + \frac{1}{5} \cdots\right] = 0.97$

G(S, wind) = 0.952



α exactly TDIDT guarantee not produce smallest decision tree, binary attributes ?

X	Y	Z	output
0	0	0	0
0	0	1	1
0	1	0	1
0	1	1	0
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	0



α X takes real values $v_1, \dots, v_n$ with $p_1, \dots, p_n$ such $\sum p_i = 1$

Shannon entropy $H(X) = H(p_1, \dots, p_n) = \sum_{i=1}^{n} -p_i \log_2 p_i$ [$\log_2 p$]

By def, $p \log_2 p = 0$ for $p \geq 0$

function $x \mapsto \log_2 x$ is concave as second derivative is $-ve \Rightarrow E[f(x)] \leq f(E(x))$

$H(x) = E_{P(x)}[-\log_2 P(x)]$ where P_X is prob. mass function.

$= \sum_{i=1}^{n} (-\log_2 p_i) \cdot p_i$

$= \sum_{i=1}^{n} (-\log_2 P_X(v_i)) \cdot P_X(v_i)$

$= E_{P_X}[-\log_2 P_X(X)]$

By Jensen Inequality

$= E_{P_X}\left[\log_2 \frac{1}{P_X(X)}\right] \leq \log_2 \left[E_{P_X}\left[\frac{1}{P_X(v_i)}\right]\right]$

$= \log_2 \sum_{i=1}^{n} P_X(v_i) \cdot \frac{1}{P_X(v_i)}$

$= \log_2 n$

α prove information gain is always non-negative ?

show that $Gain(X,Y) \geq 0 \Leftrightarrow -Gain(X,Y) \leq 0$

$Gain(X,Y) = \sum_i P(X=i) \cdot \log P(X=i) - \sum_v P(Y=v) \cdot \sum_i P(X=i \mid Y=v) \cdot \log P(X=i \mid Y=v)$

$= \sum_i \sum_v P(X=i, Y=v) \cdot \log P(X=i) - \sum_i \sum_v \underbrace{P(X=i \mid Y=v) P(Y=v)}_{P(X=i, Y=v)} \cdot \log P(X=i \mid Y=v)$

$= \sum_i \sum_v P(X=i \mid Y=v) \cdot P(Y=v) \left(\log \frac{P(X=i)}{P(X=i \mid Y=v)}\right)$

By Jensen Inequality

$\leq \sum_v P(Y=v) \cdot \log \frac{\sum_i P(X=i \mid Y=v) \cdot P(X=i)}{P(X=i \mid Y=v)}$

$\log 1$

$= 0$

a) $h_1 \leq h_2$ iff $(\forall x \in X)[(h_2(x)=1) \Rightarrow (h_1(x)=1)]$

Show $\leq$ is partial order and find bottom element?

→ reflexivity: $\forall h \in H: h \leq h$

let $h \in H$, $x \in X$ with $h(x)=1$ then $h(x)=1$

$h \leq h$, it is reflexive

→ antisymmetry: $\forall h_1, h_2 \in H$: if $h_1 \leq h_2$ and $h_2 \leq h_1$ then $h_1 = h_2$

let $h_1, h_2 \in H$ s.t. $\forall x \in X$: $h_1 \leq h_2$, $h_2 \leq h_1$

if $h_2(x)=1$, then $h_1(x)=1$

if $h_1(x)=1$ then $h_2(x)=1$

$\Rightarrow h_1(x)=1$ iff $h_2(x)=1$, $h_1 = h_2$ is antisymmetric

→ transitivity: if $h_1 \leq h_2$ and $h_2 \leq h_3$ then $h_1 \leq h_3$

let $h_1, h_2, h_3 \in H$, $x \in X$ s.t. $h_1 \leq h_2$, $h_2 \leq h_3$

if $h_2(x)=1$ then $h_1(x)=1$

if $h_3(x)=1$ then $h_2(x)=1$

then $h_3(x)=1 \Rightarrow h_1(x)=1$ for x

---

a) X be $\{0,1\}^n$, $n \in \mathbb{N}$, H_n of $x_1 \to x_n$

for $h \in H_n := \bigvee_{l \in L}$ l.c. for some $L \subseteq \{x_1, \neg x_1, x_2, \neg x_2, \dots, x_n, \neg x_n\}$

Represent this input E if coding $E \subseteq X_n$

then hypo. can be represented?

let E^+ least set of ex. $c(\vec{x}_i)=1$

E^- least set of ex. $c(\vec{x}_i)=0$

input $E = \{e_1, \dots, e_m\}$

$e_i = (\vec{x}_i, c(\vec{x}_i))$ where $x_i \in \{0,1\}$

$h = x_1 \vee x_2 \vee \dots \vee x_n \vee \bar{x}_1 \vee \dots \vee \bar{x}_n$

for $e = (\vec{x}, c(\vec{x})) \in E^-$ → program correctness

if $h(\vec{x}) = 1$

remove all literals in $\vec{x}$ from h

for $e = (\vec{x}, c(\vec{x})) \in E^+$ → program correctness

if $h(\vec{x}) = 0$

then "no".

---

a) X be $\{0,1\}^6$, H_6 be set conjunctions over x_1 to x_6

for all $h \in H_6$, $h = \bigwedge_{l \in L}$ l.c. $(\forall x_1 \forall x_2 \forall x_3 \to \forall x_4)$

$E = E^+ \cup E^-$ with $E^+ = \{(1,1,1,1,1,1), (1,1,0,1,1,1), (1,1,0,1,0,0)\}$, $E^- = \{(0,0,0,1,1,0)\}$

d) Using candidate elimination algorithm, S boundary $VS_{H,E}$?

$G_0 = \{\top\}$

$S_0 = \{x_1 \wedge x_2 \wedge \neg x_3 \wedge x_4 \wedge \neg x_5 \wedge \neg x_6\}$

$e_1 = (1,1,1,1,1,1)$

$G_1 = \{\top\}$

$S_1 = \{x_1 \wedge x_2 \wedge x_3 \wedge \neg x_4 \wedge x_5 \wedge \neg x_6\}$

$e_2 = (1,1,0,1,1,1)$

$G_2 = \{\top\}$

$S_2 = \{x_1 \wedge x_2 \wedge x_4 \wedge x_5 \wedge \neg x_6\}$

$e_3 = (1,1,0,1,0,0)$

$G_3 = \{\top\}$

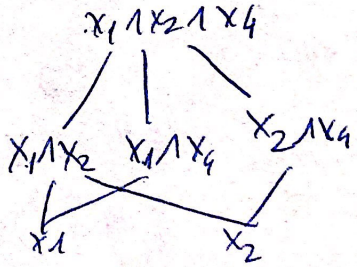
$S_3 = \{x_1 \wedge x_2 \wedge x_4\}$

$e_4 = (0,0,0,1,1,0)$

$G_4 = \{x_1, x_2\}$

e) Give all hypo. between S, G?

$VS_{H,E} = \{h \in H \mid \exists g \in G, s \in S: g \leq h \leq s\}$



$VS_{H,E} = \{x_1, x_2, x_1 \wedge x_2, x_1 \wedge x_4, x_2 \wedge x_4, x_1 \wedge x_2 \wedge x_4\}$

α Prove $H \subseteq 2^X$; X, $E^+, E^- \subseteq X$; $c \in H$, cand-elim. algo is correct, it returns general and specific boundaries at S_k and G_k?

Let $E = \{e_1, \dots, e_n\}$

$E_k = \{e_1, \dots, e_k\}$ → elements seen after k step

S_k, G_k denote boundaries after seen E_k

S_k^*, G_k^* denote actual boundaries.

Show $S_k = S_k^*$, $G_k = G_k^*$ for $k \leq m$

observation: $1 \leq k \leq m-1$, if h is consistent with E_{k+1}, it is consistent with E_k.

induction on k

if $k = 0$, $E_0 = \emptyset$, trivially holds as G_0, S_0 are initialized.

if $k+1$, $G_k = G_k^*$, $S_k = S_k^*$

assume e_{k+1} is positive,

$G_{k+1}^* = G_{k+1}$,

claim: $G_{k+1}^* \subseteq G_{k+1}$

let $g^* \in G_{k+1}^*$, by def of general boundary.

g^* is consistent with E_{k+1},

any $g < g^*$ is inconsistent with e_{k+1},

e_{k+1} is positive so any $g < g^*$ is inconsistent with e_k.

by observation g^* is consistent with e_k

then $g^* \in G_k^*$.

by induction $G_k^* = G_k$, then $g^* \in G_k$

$g^* \in G_{k+1}^*$ covers e_{k+1} then $g^* \in G_{k+1}$

claim: $G_{k+1} \subseteq G_{k+1}^*$

let $g \in G_{k+1}$, by induction e_{k+1} is positive

then $g \in G_k = G_k^*$ and

g is consistent with e_{k+1},

any $g' < g$ is inconsistent with e_k

by observation g' is inconsistent with e_{k+1}

then $g \in G_{k+1}^*$

it implies $G_{k+1} \subseteq G_{k+1}^*$

α Prove on the learning of conjunctive concepts done at most $n+1$ prediction mistakes?

Let have x_1 to x_n, set E of m examples

consider FIND-S algo, start with $h = x_1 \land \dots \land x_n \land \bar{x}_1 \land \dots \land \bar{x}_n$

if $m \leq n+1$, the claim is done

if $m > n+1$,

if first prediction mistake is made, it can delete all wrong literals in example from h

so only n literals remain

for each wrongly classified example, there is at least one literal gets deleted from ~~h~~

so there at least one literal that there may be at most n more mistakes

then in total at most $n+1$ mistakes.

α for C and H, (concept class → C, hypo. class → H) if draw $m \geq \frac{1}{\varepsilon}\left(\ln|H| + \ln\frac{1}{\delta}\right)$ random samples, find consistent hypo from H; can we PAC-learn C using H?

claim: if draw $m \geq \frac{1}{\varepsilon}\left(\ln|H| + \ln\frac{1}{\delta}\right)$ examples,

we can find consistent hypothesis,

we can PAC learn C that output h ~~[...]~~

with $\Pr[\text{error}(h) \leq \varepsilon] \geq 1-\delta$

$\Leftrightarrow \Pr[\text{error}(h) > \varepsilon] < \delta$

Let $H_B = \{h_1, \dots, h_k\} \subseteq H$. is set of hypo h_i with $\text{error}(h_i) > \varepsilon$ and it can be an output if it is consistent with all examples

Let $E = \{e_1, \dots, e_m\}$ ~~[...]~~ set of examples, $h \in H_B$

then $\Pr[h \text{ is consistent with } e \in E] \leq 1-\varepsilon$

$\Rightarrow \Pr[h \text{ is consistent with } E] \leq (1-\varepsilon)^m$

$\Leftrightarrow \Pr[\exists h \in H_B \text{ consistent with } E] \leq |H_B| \cdot (1-\varepsilon)^m \leq |H| \cdot (1-\varepsilon)^m$

$(1-\varepsilon) \leq e^{-\varepsilon}$ for $0 \leq \varepsilon \leq 1$

$\Leftrightarrow |H| \cdot e^{-\varepsilon m} \overset{!}{\leq} \delta$

probability of bad hypothesis being consistent with m examples are bounded by $|H| \cdot e^{-\varepsilon m}$ and should be at most δ

$\Leftrightarrow -\varepsilon m \leq \ln\left(\frac{\delta}{|H|}\right)$

$m \geq \frac{1}{\varepsilon}\left(\ln|H| - \ln\delta\right)$

$\Leftrightarrow m \geq \frac{1}{\varepsilon}\left(\ln|H| + \ln\frac{1}{\delta}\right)$

Let k-CNF$_n$ be set of all CNFs of each clause contains at most k literals from $\{x_1, \bar{x}_1, \dots, x_n, \bar{x}_n\}$

Then k-CNF$_n$ is efficiently PAC-learnable?

claim: it is PAC-learnable

Bound $|H|$, each clause has at most k literals and total there are $2n$ literals, $\binom{2n}{i}$ clauses with i literals.

total num of different clauses are bounded by

$$\sum_{i=1}^{k} \binom{2n}{i} \le \binom{2n}{k} \cdot \frac{2n-(k-1)}{2n-(2k-1)}$$

$$\le (2n)^k \cdot \frac{2n-(k-1)}{2n-(2k-1)} = \text{poly}(n)$$

each clause is part of certain k-CNF or not,

total number of k-CNF is bounded by $|H| \le 2^{\text{poly}(n)}$

And consider hypothesis

$$m \ge \frac{1}{\varepsilon}\left(\ln|H| + \ln\frac{1}{\delta}\right)$$

$$\ge \frac{1}{\varepsilon}\left(\ln 2^{\text{poly}(n)} + \ln\frac{1}{\delta}\right)$$

$$\ge \frac{1}{\varepsilon}\left(\text{poly}(n)\ln 2 + \ln\frac{1}{\delta}\right) \text{ examples}$$

thus is polynomial in $\frac{1}{\varepsilon}, \frac{1}{\delta}, n$.

so there are $\le$ poly(n) different clauses, ~~define~~ Consider hypothesis, let boolean variables $x_1', \dots, x_m'$ where $m \le$ poly(n) is num of different clauses.

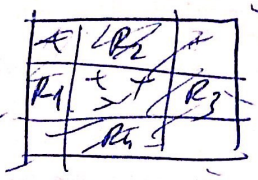
x_i' is 1 iff clause is satisfied

any k-CNF formula over $x_1, \dots, x_n$ is equivalent to conjunctions over $x_1', \dots, x_m'$,

consistent hypothesis can be found by FIND-S algo.

by applying algo for conjunctive concepts,

Let $\mathcal{Y} = \mathbb{R}^2$, concepts C are set of axis parallel aligned rectangles, is C poly. PAC learnable?



the first consistent hypothesis, define smallest rectangle contains all positive examples called R_E

area of possible misclassification: $R \setminus R_E$

probability example drawn from $R \setminus R_E$ is error(R_E)

Guarantee $Pr(\text{error}(R_E) \le \varepsilon) \ge 1-\delta$

if $P(R) < \varepsilon$, then $P(R \setminus R_E) \le P(R) < \varepsilon$

if $P(R) > \varepsilon$, then consider shrinking R_1 to R_4 of $R \setminus R_E$ s.t. $P(R_i) = \frac{\varepsilon}{4}$

choose enough examples with higher prob. that there is a point in each R_i, then R_E intersects all regions and error is at most ε.

$$R \setminus R_E \subseteq \bigcup_{i=1}^{4} R_i.$$

$$P(R \setminus R_E) \le P\left(\bigcup_{i=1}^{4} R_i\right) \le \sum_{i=1}^{4} P(R_i)$$

$$\le \sum_{i=1}^{4} \frac{\varepsilon}{4}$$

$$\le \varepsilon$$

And lower bound,

$P(R_i \text{ is not hit}, i \ge 1 \text{ to } 4)$

$\Leftrightarrow P\left(\bigcup_{i=1}^{4} R_i \text{ is not hit}\right)$

$$\le \sum_{i=1}^{4} P(R_i \text{ is not hit})$$

$$= \sum_{i=1}^{4} (1 - P(R_i))^m$$

$$> \sum_{i=1}^{4} \left(1 - \frac{\varepsilon}{4}\right)^m$$

$$= 4 \cdot \left(1 - \frac{\varepsilon}{4}\right)^m$$

$$\le 4 \cdot e^{-\frac{\varepsilon}{4}m} \overset{!}{\le} \delta$$

since $1 - \frac{\varepsilon}{4} \le e^{-\frac{\varepsilon}{4}}$

$\Leftrightarrow -\frac{\varepsilon}{4} m \ln 4 \le \ln\delta$

$$m \ge \frac{4}{\varepsilon}\ln\left(\frac{4}{\delta}\right)$$

if we draw at least m examples, we output hypothesis with error at most ε, with probability ~~ε~~ at least $1-\delta$

α Can X, $C \subseteq P(X)$, C shatters $S \subseteq X$ for $A \subseteq S$, there a $C' \subseteq C$, C' covers every A - $VC(C)$ size of largest finite subset of X can be shattered by C?

a) Circles in $\mathbb{R}^2$?

claim = $VC(C) = 3$

case 0 - draw 3 points, set of points is shattered by C

why not 4,

case 1 = some point x in interior of convex hull of others, we can't use circle to cover exactly points in convex hull

case 2 = every point in convex hull, we have 4 points either any circle contains A also contains B or any circle contains B also contains points in A.

c) convex hyp?

claim = $VC(C) = 2d+1$

take $2d+1$ points on circle. show can be shattered by d half spaces.

induction $d=1$, 3 points by 1 half-space

$d \geq 3$; intersection of these half spaces is convex hyp.

why not $2d+2$?

case 1 = some points are in interior of convex hull of others, we can't separate convex hull from these points with any convex shape.

case 2 = all points on convex hull, subset contains every other points in convex hull, we can't separate subset from its complement with d-hyps because we need at least $d+1$ half spaces

b) by C is a $\cdot$ for VC(triangles)

α a) $|C| < \infty$, show $VC(C) \leq \log_2 |C|$?

By def, $VC(C)$ is largest subset s.t.
A is shattered by C
then $\forall B \subseteq A: \exists c \in C: c \cap A = B$
$2^{VC(C)} = 2^{|A|} \leq |C|$
$\Leftrightarrow VC(C) \leq \log_2 |C|$

b) $C_1, C_2 \subseteq 2^X$ are concept class over X, show $VC(C_1) \leq VC(C_2)$?

A is largest set shattered by C_1 and for each $c \in C_1 : c \in C_2$ where $C_1 \subseteq C_2$,
$VC(C_1) = |A|$ then A is also shattered by C_2
then $VC(C_2) \geq |A| = VC(C_1)$

c) $C \subseteq 2^X$ over X, $\bar{C} = \{X \setminus c : c \in C\}$, show $VC(C) = VC(\bar{C})$?

Yes, let $A \subseteq X$ s.t $VC(C) = |A|$,
claim: A is also shattered by $\bar{C}$
A is shattered by C,
for each $B \subseteq A$, there is $c \in C$ s.t. $c \cap A = A \setminus B$

$X \setminus c \in \bar{C}$
$A \cap (X \setminus c) = A \setminus c = A \setminus (A \cap c) = A \setminus (A \setminus B) = B$
so A is shattered by $\bar{C}$
then $VC(\bar{C}) \geq VC(C)$

α boolean $f = \{0,1\}^n \to \{0,1\}$ is symmetric if $f(x_1, \dots, x_n) = f(x_{\pi(1)}, \dots, x_{\pi(n)})$ for π on $\{0,1\}^n$, corollary: show C_n is poly. PAC-learnable?

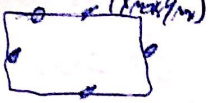
claim = $VC(C) = n+1$

size of largest $A \subseteq \{0,1\}^n$ s.t. any subset Y of A is given by symmetric $\forall x \in X (x \in Y \Leftrightarrow f_Y(x) = 1)$
if it satisfies then elements of A must be independent
s.t. $\exists x \in X$ and have s and s'
consider equivalent classes $s \sim s'$ iff s and s' are permutations of each other, then any A that is shattered by C must contain at most one representative of each equivalence class,
there are $n+1$ such classes, then largest A size is $n+1$
then $VC(C) = n+1 = \text{poly}(n)$
we can find $h \in C$ consistent with S in linear time with $S \subseteq X$ and $c \in C$
then BEHW C is poly. PAC-learnable.

α Show C consisting of all axis parallel rectangle is poly. PAC learnable by using Corollary 1?

Consistent hypothesis can be found in poly. time by computing smallest rectangle containing all positive examples.

Claim: VCdim(C) = 4



Consider all subsets can be realized by some axis-aligned rectangle each set of A with 5 points. If 3 of them are on line, this cannot be shattered.

Assume none are colinear, let $x_{max}, x_{min}, y_{max}, y_{min}$ denote max/min x/y-coordinates of points in A, consider bounding box BB(A) is given by points $(x_{max}, y_{max}), (x_{max}, y_{min}), (x_{min}, y_{min}), (x_{min}, y_{max})$.

if at least one point lies inside BB(A), then it cannot be shattered.

if all lie on bounding box, then there exists labeling that 4 positive and 1 negative that cannot be realized by C

then VCdim(C) < 5 so VCdim(C) = 4.

α Consider lab result: positive returns 98% of disease present, correct negative result 97% of disease not present, 0.008 of some pop. have cancer: 0.008

$$P(C|+) = \frac{P(+|C) \cdot P(C)}{P(+)} = \frac{0.00078}{P(+)}$$

$$P(\neg C|+) = \frac{P(+|\neg C) \cdot P(\neg C)}{P(+)} = \frac{0.0298}{P(+)}$$

(0.03, 0.992)

we know $P(C|+) + P(\neg C|+) = 1$

$$P(C|+) = \frac{0.00078}{0.00078 + 0.0298} = 0.21$$

Suppose we decide to take second test for some person, suppose second test responds as well. What are person prob. of cancer and same following those independent assumptions?

prob of two tests given person has cancer $P(+_1, +_2|C) = P(+|C) \cdot P(+|C)$

$P(+|C) = 0.98$, $P(C) = 0.008$

$P(-|\neg C) = 0.97$, $P(\neg C) = 0.992$

$P(+|\neg C) = 0.03$ and $P(C|+) = 0.21$

$$P(C|+,+) \overset{\text{Bayes}}{=} \frac{P(+,+|C) \cdot P(C)}{P(+,+)} \overset{\text{conditional independence}}{=} \frac{P(+|C) \cdot P(+|C) \cdot P(C)}{P(+,+)} = 0.87$$

(0.98 · 0.98 · 0.008)

$$P(+,+) = P(+,+|C) \cdot P(C) + P(+,+|\neg C) \cdot P(\neg C)$$

(0.00817)

α unfair coin, estimate bias of coin compare with known prob θ of head. Is ML estimation $\hat{\theta}$ of θ for sequences of n coin flips?

We observed h occurrences of heads during n tosses, then prob of dataset is $L(\theta) = \theta^h \cdot (1-\theta)^{n-h}$ (h heads, n−h tails)

log likelihood:

$$\ell(\theta) = h \cdot \log\theta + (n-h) \cdot \log(1-\theta)$$

to maximize $\hat{\theta}$, take first derivative and set $= 0$

$$\frac{d\,\ell(\theta)}{d\theta} = \frac{h}{\theta} - \frac{(n-h)}{1-\theta} \overset{!}{=} 0$$

$$\Leftrightarrow \frac{h}{\theta} = \frac{n-h}{1-\theta}$$

$$\Leftrightarrow \theta = \frac{h}{n} \text{ so } \hat{\theta} = \frac{h}{n} \text{ maximizes likelihood}$$

α $D \subseteq [0,1] \times \{0,1\}$ of training examples with $D = \{\langle(1,1,1),1\rangle, \langle(1,0,1),0\rangle, \langle(0,1,0),0\rangle, \langle(0,1,1),0\rangle\}$

$X = (1,1,0)$ and $H = \{h_1, h_2, h_3\}$ with

if $a_1 = 1$ then h_1 returns "1" with $\frac{4}{5}$, else returns "1" with $\frac{1}{5}$

if $a_2 = 1$, h_2 returns "1" with $\frac{4}{5}$, else h_2 returns 1 with $\frac{1}{5}$

if $a_3 = 1$, h_3 returns "1" with $\frac{2}{3}$, else returns "1" with $\frac{1}{3}$

assume hypothesis h_1, h_2, h_3 are uniformly distributed.

1) What is map hypo. $\arg\max_{h \in H} P(h|D)$?

$$h_{MAP} = \arg\max_{h \in H} \frac{P(D|h) \cdot P(h)}{P(D)}$$

$$= \arg\max_{h \in H} P(D|h) \cdot P(h)$$

by def $P(h_i) = \frac{1}{3}, i = 1..3$

$$= \arg\max_{h \in H} P(D|h)$$

$$h_1 = P(D|h_1) = \frac{4}{5} \cdot \frac{1}{5} \cdot \frac{4}{5} \cdot \frac{4}{5} = 0.1024$$

$$h_2 = P(D|h_2) = \frac{4}{5} \cdot \frac{4}{5} \cdot \frac{1}{5} \cdot \frac{4}{5} = 0.1024$$

$$h_3 = P(D|h_3) = \frac{2}{3} \cdot \frac{2}{3} \cdot \frac{2}{3} \cdot \frac{2}{3} = 0.1975 = h_{MAP}$$

↪

b) What is bayes optimal classification of x?

argmax $\sum_{v_j \in V} \sum_{h_i \in H} P(v_j|h_i) \cdot P(h_i|D)$, we don't need $P(x)$

$P(h_1|D) = \frac{0,1024}{0,1024+0,1024+0,1875} \approx 0,2545$

$P(h_2|D) \approx 0,2545$

$P(h_3|D) = \frac{0,1875}{\dots} \approx 0,491$

$X = (1,1,0)$

for '0': $\sum_{h_i \in H} P(0|h_i) \cdot P(h_i|D)$

$= P(0|h_1) \cdot P(h_1|D) + P(0|h_2) \cdot P(h_2|D) + P(0|h_3) \cdot P(h_3|D)$

$\frac{1}{5} \cdot 0,2545 + \frac{1}{5} \cdot 0,2545 + \frac{2}{5} \cdot 0,491 \approx 0,3964$

for '1': $\frac{4}{5} \cdot 0,2545 + \frac{4}{5} \cdot 0,2545 + \frac{1}{5} \cdot 0,491 \approx 0,6036$

c) What is naive bayes classifier of x?

$v_{NB} = \operatorname{argmax}_{v_j \in V} P(v_j) \cdot \prod_i P(x_i|v_j)$

$\Rightarrow P(v=1) \cdot \prod_{i=1}^{3} P(x_i|v=1) = \frac{1}{4} \cdot 1 \cdot 1 \cdot 0 = 0$

$P(v=0) \cdot \prod_{i=1}^{3} P(x_i|v=0) = \frac{3}{4} \cdot \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{1}{3} = \frac{1}{36}$

naive bayes classifies to '0'

Let $P \subseteq \mathbb{R}^2$, $|P| = n$, each all centers are pairwise different

Let be Voronoi cell containing p holds $d(p,q) < d(p,q')$ for all $q' \in P \setminus \{p\}$

a) $n \geq 3$ exist $P \subset \mathbb{R}^2$ of n points s.t. all cells of Voronoi diagram of P are unbounded?

take n points on circle,
voronoi diagrams edges correspond to perpendicular bisectors of secants of neighbour points
then no bounded cell in diagram

b) $P \subset \mathbb{R}^2$ and $|P| = n$ for $n \geq 3$, show number of vertices of voronoi diagram of P is atmost $2n-5$ and number of its edges is atmost $3n-6$ where points in plane $\mathbb{R}^2$ is vertex of voronoi diagram of P, point in plane lies on edge?

Construct $G = (V(G), E(G))$ as $V(G)$ has all vertices of diagram + ∞, $E(G)$ has all edges + bounded edges

V denote num of vertices, e denote num of edges

observations: G is planar, $|V(G)| = v+1$, $|E(G)| = e$,

num of faces of $G = n$ (each face corresponds to Voronoi cell)

any vertex of voronoi diagram must have at least 3 edges

$|\delta(v)| \geq 3$, $v \in V(G)$

$2|E(G)| = \sum_{v \in V(G)} |\delta(v)| \geq 3|V(G)|$

$\Leftrightarrow v+1 \leq \frac{2}{3}e$

from euler, $|V(G)| - |E(G)| + f = 2$

$(v+1) - e + n = 2$

$\frac{2e}{3} - e + n \geq 2$

$e \leq 3n-6$

and $(v+1) - e + n = 2$

$v+1 - (3n-6) + n \leq 2$

$v \leq 2n-5$

a 100 rand. examples, h classifies 87 correctly,

a) What is 95% confidence interval for true error of h?

$error_S(h) \mp z_{95\%} \cdot \sqrt{\frac{error_S(h)(1-error_S(h))}{n}}$

$= 0,13 \mp 1,96 \cdot \sqrt{\frac{0,13 \cdot 0,87}{100}}$

interval is $[0,0661; 0,1939]$

b) What is the meaning of this confidence interval?

with at least 95% confidence probability, this interval contains value of true error.

a Derive equivalence for F_β since and

$F_\alpha = \frac{1}{\alpha \cdot \frac{1}{prec} + (1-\alpha) \cdot \frac{1}{recall}}$ for $\alpha \in [0,1]$

where $F_\beta = (1+\beta^2) \cdot \frac{P \cdot R}{\beta^2 \cdot P + R}$?

observations F_α is weighted mean of harmonic mean of P and R; $\frac{2 \cdot P \cdot R}{P+R}$

claim = if we choose $\alpha = \frac{1}{\beta^2+1}$, then $F_\alpha = F_\beta$

proof = $F_\alpha = \frac{1}{\frac{1}{1+\beta^2} \cdot \frac{1}{P} + (1 - \frac{1}{\beta+\beta^2}) \cdot \frac{1}{R}}$

$= \frac{1}{\frac{1}{1+\beta^2} \cdot (\frac{1}{P} + \frac{\beta^2}{R})}$

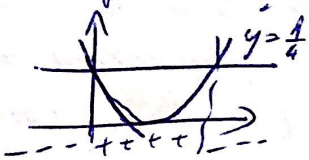
$= (1+\beta^2) \cdot \frac{P \cdot R}{\beta^2 \cdot P + R} = F_\beta$

α 1) $S^+ \subset [0,1]$ and $S^- \subset \mathbb{R} \setminus [0,1]$ be finite sets of positive and negative reals. Give fun. $\Phi: \mathbb{R} \to \mathbb{R}^N$ for some N s.t. sets $\{\Phi(x) \cdot w + b : S^+\}$ and $\{\Phi(x) \cdot w + b : S^-\}$ become linearly separable?

$\Phi: \mathbb{R} \to \mathbb{R}^2$, $N=2$

$x \to (x, (x - \frac{1}{2})^2)$

so we define the points



2) $S^+ \subset B$ and $S^- \subset \mathbb{R}^3 \setminus B$ be finite sets with $B = \{(x,y,z) : x^2+y^2+z^2 \le 1\}$, $\Phi: \mathbb{R}^3 \to \mathbb{R}$ for some N s.t. $\{\Phi(x) \cdot w + b : S^+\}$ and $\{\Phi(x) \cdot w + b : S^-\}$ become linearly separable?

$\Phi: \mathbb{R}^3 \to \mathbb{R}$

$(x,y,z) \to x^2+y^2+z^2$

observe if $(x,y,z) \in S^+ \subset B$, $x^2+y^2+z^2 \le 1$

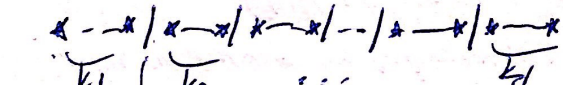
if $(x,y,z) \in S^- \subset \mathbb{R} \setminus B$, $x^2+y^2+z^2 > 1$

so plane 1 is separating hyperplane

α $X_1, \dots, X_d$ real valued variables, monomial of degree k over $X_1, \dots, X_d$ is $\prod_{i=1}^{d} X_i^{k_i}$ with nonneg integers $k_1, \dots, k_d$ s.t. $\sum_{i=1}^{d} k_i = k$

a) Let D be num. of monomials of degree k over $X_1, \dots, X_d$ show that $D = \binom{d+k-1}{d-1} = \binom{d+k-1}{k}$?

encode monomials $X_1^{k_1}$ to $X_d^{k_d}$ as sequences of bars and stars,



$d-1$ bars for d variables

$$\frac{(d+k-1)!}{(d-1)! \cdot k!} = \binom{d+k-1}{d-1} = \binom{d+k-1}{k}$$

b) $\Phi: \mathbb{R}^d \to \mathbb{R}^D$ with $\Phi: \vec{x} = (x_1, \dots, x_d) \to (M_1(\vec{x}), \dots, M_D(\vec{x}))$ for $\vec{x} \in \mathbb{R}^d$, $\{M_1, \dots, M_D\}$ is set of all monomials of degree k over $X_1, \dots, X_d$. $S_1^- \subset \mathbb{R}^d$ with $|S_1^-| = n$, $S^- = \{\Phi(s) : s \in S_1^-\}$, (S^+, S^-) be dichotomy of S^- chosen uniformly at rand. Give prob. $P = P(n,d,k)$ of event that (S^+, S^-) is linearly separable?

probability that dichotomy randomly selected uniformly from general positions of $\mathbb{R}^N$ is linearly separable by $P = \frac{1}{2^n} \left(\sum_{k=0}^{N} \binom{n-1}{k} \right)$

there are $D = \binom{d+k-1}{d-1}$ monomials of degree k

choosing $N = D$,

$$P = \frac{1}{2^n} \left(\sum_{k=0}^{D} \binom{n-1}{k} \right)$$

α 1) Prove $S = \{(\vec{x}_1, y_1), \dots, (\vec{x}_n, y_n)\} \subset \mathbb{R}^d \times \mathbb{R}$, and vector $\vec{w} \in \mathbb{R}^d$ s.t. $f(\vec{x}) = \langle \vec{x}, \vec{w} \rangle$ best interpolates meaning minimizes empirical error under square loss, minimize $R_{emp}[f] = \frac{1}{n} \sum_{i=1}^{n} (f(\vec{x}_i) - y_i)^2$?

claim = empirical risk wrt. square loss function is $R_{emp}[f] = \frac{1}{n} \| X\vec{w} - \vec{y} \|^2$

proof = $\vec{y} = (y_1, \dots, y_n)$ and X be $n \times d$ matrix that rows are vectors $\vec{x}_1, \dots, \vec{x}_n$.

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1d} \\ x_{n1} & \cdots & x_{nd} \end{pmatrix}_{n \times d}$$

$$R_{emp}[f] \overset{def}{=} \frac{1}{n} \sum_{i=1}^{n} (f(\vec{x}_i) - y_i)^2$$

$$= \frac{1}{n} \left(\sum_{i=1}^{n} (\langle \vec{x}_i, \vec{w} \rangle - y_i)^2 \right)$$

$$\overset{def\ of\ euclidean\ form}{=} \frac{1}{n} \| X\vec{w} - \vec{y} \|^2 \quad \text{where}$$

$$\|\vec{a}\| = (a_1^2 + \cdots + a_d^2)^{\frac{1}{2}}, \ \vec{a} \in \mathbb{R}^d$$

2) Prove convex min prob.

$$\min_{\vec{w}} R_\lambda(\vec{w}, S) = \min_{\vec{w}} \| X\vec{w} - \vec{y} \|^2 + \lambda \|\vec{w}\|^2.$$

claim = $\nabla_{\vec{w}} R_\lambda(\vec{w}, S) = 2X^T X\vec{w} - 2X^T\vec{y} + 2\lambda\vec{w}$

proof = $\nabla_{\vec{w}} (\langle X\vec{w} - \vec{y}, X\vec{w} - \vec{y} \rangle + \lambda \langle \vec{w}, \vec{w} \rangle)$

$= \nabla_{\vec{w}} (\vec{w}^T X^T X\vec{w} - 2\vec{y}^T X\vec{w} + \vec{y}^T\vec{y} + \lambda \vec{w}^T\vec{w})$

(by $\|\vec{a}\|^2 = \langle \vec{a}, \vec{a} \rangle = \vec{a}^T\vec{a}$) (1) (2) (3)

1 $= \nabla_{\vec{w}} (\vec{w}^T X^T X\vec{w}) = \nabla_{\vec{w}} (\vec{w}^T X^T) \cdot X\vec{w} + \vec{w} X^T \cdot \nabla_{\vec{w}} (X\vec{w})$

$= X^T X \cdot \vec{w} + \vec{w}^T \cdot X^T X = 2X^T X\vec{w}$

2 $= \nabla_{\vec{w}} (-2\vec{y}^T X\vec{w}) = -2 \cdot \nabla_{\vec{w}} ((\vec{y})_1 w_1 + \cdots + (\vec{y})_d w_d)$

$= -2X^T\vec{y}$

3 $= \lambda \nabla_{\vec{w}} (\vec{w}^T\vec{w}) = \lambda \nabla_{\vec{w}} (w_1^2 + \cdots + w_d^2) = \lambda (2w_1 + \cdots + 2w_d)$

$= 2\lambda\vec{w}$

$= 2X^T X\vec{w} - 2X^T\vec{y} + 2\lambda\vec{w}$

α $S \subseteq \mathbb{R}^d \times \{+1, -1\}$ finite set of +and − examples (assume at least one positive, one negative examples)

Let $c^+ = \frac{1}{|S^+|} \sum_{(x,+1) \in S} x$ and $c^- = \frac{1}{|S^-|} \sum_{(x,-1) \in S} x$ be means of + and − examples, assign new value $x \in \mathbb{R}^d$ to class whose mean is closer to x, give corresponding decision function. Was modification of x only through inner products?

assign each new point to class which minimizes distance to its mean

$f(x) = \operatorname{sgn}(\|c^- - x\|^2 - \|c^+ - x\|^2)$

$= \operatorname{sgn}(\langle c^- - x, c^- - x\rangle - \langle c^+ - x, c^+ - x\rangle)$

we have $f(x) := \begin{cases} 1 & \text{if } x \text{ is closer to } c^+ \\ -1 & \text{if } x \text{ is closer to } c^- \end{cases}$

α

1) $k: X \times X \to \mathbb{R}$ be kernel funct. : its structure of underlying feature space embedding funct. Φ (ex for which $k(x,y) = \langle \Phi(x), \Phi(y)\rangle$) is unique, why or not?

No, Φ be embedding with $k(x,y) = \langle \Phi(x), \Phi(y)\rangle$, then we have $\langle -\Phi(x), -\Phi(y)\rangle = \langle \Phi(x), \Phi(y)\rangle$
$= k(x,y)$

so $-\Phi$ is also feature space embedding

2) $k: \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ be funct. defined by $k(\vec{x}, \vec{y}) = (\langle \vec{x}, \vec{y}\rangle + 1)^2$ for all $\vec{x}, \vec{y} \in \mathbb{R}^2$, is k kernel funct. if so give funct. Φ for which $k(\vec{x}, \vec{y}) = \langle \Phi(x), \Phi(y)\rangle$?

Yes, $k(\vec{x}, \vec{y}) = (\langle \vec{x}, \vec{y}\rangle + 1)^2$

$= \langle \vec{x}, \vec{y}\rangle^2 + 2\langle \vec{x}, \vec{y}\rangle + 1$

$= (x_1 y_1 + x_2 y_2)^2 + 2(x_1 y_1 + x_2 y_2) + 1$

$= x_1^2 y_1^2 + 2 x_1 y_1 x_2 y_2 + x_2^2 y_2^2 + 2 x_1 y_1 + 2 x_2 y_2 + 1$

choose $\Phi(x_1, x_2) = (x_1^2, \sqrt{2} x_1 x_2, x_2^2, \sqrt{2} x_1, \sqrt{2} x_2, 1)$

for all $x = (x_1, x_2)$ and $y = (y_1, y_2)$

we have $k(x,y) = \langle \Phi(x), \Phi(y)\rangle$

3) $k: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be given by $k(\vec{x}, \vec{y}) = \sum_{i=1}^{d} (x_i + y_i)$ for all $\vec{x}, \vec{y} \in \mathbb{R}^d$, is k kernel?

No, inner product of vector with i-th value is always nonnegative,

consider $x = (-1, 0, \dots) \in \mathbb{R}^d$

$y = (-1, 0, \dots) \in \mathbb{R}^d$

$\Rightarrow k(\vec{x}, \vec{y}) = -2$

so there cannot be any Φ s.t. $k(\vec{x}, \vec{y}) = \langle \Phi(\vec{x}), \Phi(\vec{y})\rangle$

4) $k: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be funct. defined by $k(x,y) = \prod_{i=1}^{d} h\left(\frac{x_i + y_i}{b}\right) h\left(\frac{x_i - y_i}{b}\right)$

for all $x, y \in \mathbb{R}^d$ and $h: \mathbb{R} \to \mathbb{R}$ with $b \in \mathbb{R}$, is k kernel funct.?

Yes, $k(x,y) = f(x) \cdot f(y)$ is kernel over X for all functions

$f: X \to \mathbb{R}$ set $f(x) = \prod_{i=1}^{d} h\left(\frac{x_i + q}{b}\right)$

then $k(\vec{x}, \vec{y}) = f(x) \cdot f(y)$, k is kernel

5) $k: \mathbb{R}^d \setminus \{0\} \times \mathbb{R}^d \setminus \{0\} \to \mathbb{R}$ be funct. defined by $k(\vec{x}, \vec{y}) = -\frac{\langle \vec{x}, \vec{y}\rangle}{\|\vec{x}\| \cdot \|\vec{y}\|}$, is k kernel funct?

No, consider $d = 1$, $k(\vec{x}, \vec{x}) = -\frac{x \cdot x}{|x| \cdot |x|} = -1 < 0$

but $\langle \Phi(x), \Phi(y)\rangle \geq 0$ should be for any inner product

so k is not kernel

α U be finite set and $k: 2^U \times 2^U \to \mathbb{R}$ by $k(S_1, S_2) = |S_1 \cap S_2|$ for all $S_1, S_2 \subseteq U$, is k kernel funct?

Yes, let $U = \{u_1, \dots, u_m\}$

define $\Phi: 2^U \to \{0,1\}^m$

$S \to x^S$ where $x_i^S = \begin{cases} 1 & \text{if } u_i \in S \\ 0 & \text{else} \end{cases}$

for $S_1, S_2 \subseteq U$,

$\langle \Phi(S_1), \Phi(S_2)\rangle = \sum_{i=1}^{m} \underbrace{x_i^{S_1} \cdot x_i^{S_2}}_{1 \text{ iff } u_i \in S_1 \cap S_2}$

$= |S_1 \cap S_2|$

α $k: X \times X \to \mathbb{R}$ over $X = \{x_1, \dots, x_t\}$ for $n \in \mathbb{N}$ and $K_{x_i x_k}$, Gram (or kernel) matrix of X wrt k.

1) for any nonnegative real number c, define $n \times n$ matrix $K_{x_i x_k}^c$ by $(K_{x_i x_k}^c)_{ij} = (K_{x_i x_k})_{ij} + c$ for all $i, j = 1, \dots, n$. Show $K_{x_i x_k}^c$ is Gram matrix of X wrt k' for some kernel $k': X \times X \to \mathbb{R}$ on X?

k is kernel funct. there is a feature map $\Phi: X \to F$ with $k(x,y) = \langle \Phi(x), \Phi(y)\rangle$

define Φ', $X \to (F \times \mathbb{R})$

$x \to (\Phi(x), \sqrt{c})$

define $\langle (x, \hat{x}), (y, \hat{y})\rangle_{F \times \mathbb{R}} = \langle x, y\rangle + \hat{x} \cdot \hat{y}$

for $\forall x, y \in F$, $\hat{x}, \hat{y} \in \mathbb{R}$

define $k'(x,y) = k(x,y) + c$

$\langle \Phi'(x), \Phi'(y)\rangle = \langle (\Phi(x), \sqrt{c}), (\Phi(y), \sqrt{c})\rangle_{F \times \mathbb{R}}$

$= \langle \Phi(x), \Phi(y)\rangle + c$

$= k(x,y) + c$

$= k'(x,y)$

so k' is kernel with $K_{x_i x_k}$ as its gram matrix

2) for any nonnegative real number c, define $n \times n$ matrix $K_{X,k}^{c}$ by $(K_{X,k}^{c})_{ij} = \begin{cases} (K_{X,k})_{ij} + c & \text{if } i = j \\ (K_{X,k})_{ij} & \text{otw} \end{cases}$ for all $x_i, x_j = 1 \to n$.

Show that $K_{X,k}^{c}$ is Gram matrix of X wrt k' for some kernel $k': X \times X \to \mathbb{R}$ over X?

K is kernel that there is feature map $\Phi: X \to \mathbb{R}^n$ with

$k(x,y) = \langle \Phi(x), \Phi(y) \rangle$

define Φ', $X \to F \times \mathbb{R}^{|X|}$

$X \to (\Phi(x), \sqrt{c} \cdot e_x)$ where

$e_x \in \mathbb{R}^{|X|}$, $(e_x)_i = \begin{cases} 1 & \text{if } x = x_i \\ 0 & \text{otw} \end{cases}$

define $\langle (x, \tilde{x}), (y, \tilde{y}) \rangle_{F \times \mathbb{R}^{|X|}} = \langle x, y \rangle_F + \langle \tilde{x}, \tilde{y} \rangle_{\mathbb{R}^{|X|}}$

define $k'(x,y) = k(x,y) + c \cdot \langle e_x, e_y \rangle_{\mathbb{R}^{|X|}}$

$\langle \Phi'(x), \Phi'(y) \rangle = \langle (\Phi(x), \sqrt{c} e_x), (\Phi(y), \sqrt{c} e_y) \rangle$

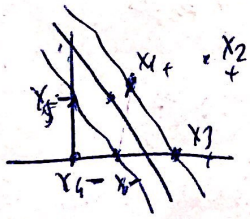
$= \langle \Phi(x), \Phi(y) \rangle_F + c \langle e_x, e_y \rangle_{\mathbb{R}^{|X|}}$

$= k(x,y) + c \langle e_x, e_y \rangle$

$= k'(x,y)$

so k' is kernel with $K_{X,k}$ as it is gram matrix

a



$(1,1)^T, +1)$, $((2,2)^T, +1)$, $((2,0)^T, +1)$,
$((0,0)^T, -1)$, $((1,0)^T, -1)$, $((0,1)^T, -1)$?

1) Give primal optimization problem for hard margin SVM?

$\min \frac{1}{2} \|\vec{w}\|^2$ s.t. $y_i(\langle \vec{w}, x_i \rangle + b) + 1 \le 0$

for each training data point $i = 1 \to n$

$\vec{w} = (w_1, w_2)$ and plugging in data points into constraints,

$\min \frac{1}{2}(w_1^2 + w_2^2)$ subject to

$-1((w_1 + w_2) + b) + 1 \le 0$

$-1((2w_1 + 2w_2) + b) + 1 \le 0$

$-1((2w_1) + b) + 1 \le 0$

$b + 1 \le 0$

using Lagrangian,

$\min_{\vec{w}, b} \max_{\vec{\alpha} \ge 0} \left(\frac{-1}{2} \|\vec{w}\|^2 + \sum_{i=1}^{6} \alpha_i y_i \langle \vec{w}, b \rangle \right)$

$\min_{w_1, w_2, b} \max_{\vec{\alpha} \ge 0} \left(\frac{w_1^2 + w_2^2}{2} + \sum_{i=1}^{6} \alpha_i y_i (w_1 + w_2 + b) \right)$

where y_i's are LHS of inequalities

2) Give dual optimization problem for hard margin SVM?

$L_D(\vec{\alpha}) = \max_{\vec{\alpha} \ge 0} \left(\sum_{i=1}^{6} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{6} \alpha_i \alpha_j y_i y_j (\vec{x}_i, \vec{x}_j) \right)$

s.t. $\sum_{i=1}^{6} \alpha_i y_i = \alpha_1 + \alpha_2 + \alpha_3 - \alpha_4 - \alpha_5 - \alpha_6 = 0$

3) What are support vectors?

$(0,1)^T, (1,0)^T, (1,1)^T, (2,0)$

4) How large is maximum margin?

$\gamma = \frac{2}{\|\vec{w}\|} = \frac{2}{\sqrt{2^2 + 2^2}} = \frac{\sqrt{2}}{2}$

a) Give derivation of dual optimization problem for soft margin SVM?

primal optimization for $C > 0$,

prob.

$\min_{\vec{w}, b, \vec{\xi}} C \sum_{i=1}^{n} \xi_i + \frac{1}{2} \|\vec{w}\|^2$ s.t. $y_i(\langle \vec{w}, \vec{x}_i \rangle + b) \ge 1 - \xi_i$

where $\xi_i \ge 0$, $i = 1 \to n$

lagrangian multiplier α_i, χ_i,

$L(w, b, \xi_i, \alpha, \chi) = C \cdot \sum_{i=1}^{n} \xi_i + \frac{1}{2} \|\vec{w}\|^2$

$- \sum_{i=1}^{n} \alpha_i (y_i (\langle \vec{w}, x_i \rangle + b) - (1 - \xi_i)) - \sum_{i=1}^{n} \chi_i \xi_i$

optimization problem will be,

$\max_{\alpha, \chi} \min_{w, b, \xi_i} L(w, b, \xi_i, \alpha, \chi)$ s.t. $\alpha, \chi \ge 0$

compare derivative and set it to 0,

$\frac{\partial L}{\partial \vec{w}} = \vec{w} - \sum_{i=1}^{n} \alpha_i y_i x_i \overset{!}{=} 0$

$w = \sum_{i=1}^{n} \alpha_i y_i x_i$

$\frac{\partial L}{\partial b} = - \sum_{i=1}^{n} \alpha_i y_i \overset{!}{=} 0$

$\sum_{i=1}^{n} \alpha_i y_i = 0$

$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \chi_i \overset{!}{=} 0$

$C = \alpha_i + \chi_i$ where $0 \le \alpha_i \le C$

then $\max_{\alpha, \chi} \min_{\vec{w}, b, \xi_i} L(w, b, \xi, \alpha, \chi)$

$\max_{\alpha, \chi} \min_{w, b, \xi_i} C \sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{ij} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle$

$- \sum \alpha_i \xi_i - \sum \chi_i \xi_i$

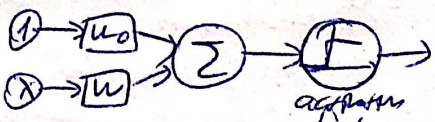
$\max_{\alpha, \chi} \min_{w, b, \xi_i} \sum_i (C - \alpha_i - \chi_i) \xi_i + \sum_i \alpha_i - \frac{1}{2} \sum_{ij} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle$

$\Rightarrow \max_{\alpha, \chi} \min_{w, b, \xi_i} \sum_i \alpha_i - \frac{1}{2} \sum_{ij} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle$ s.t. $0 \le \alpha_i \le C$, $\sum \alpha_i y_i = 0$

α Construct perceptrons

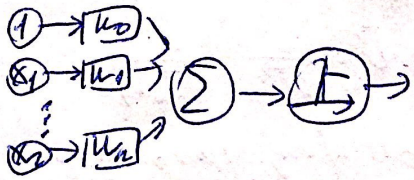
a) $f(x) = \neg x$?

$sgn(z) = \begin{cases} 1 & \text{if } z > 0 \\ 0 & \text{else} \end{cases}$



$w_0 = 1$, $w = -1$, output is 1 iff $x = 0$

b) $f(x_1, \dots, x_n) = \bigwedge_{i=1}^{n} x_i$?

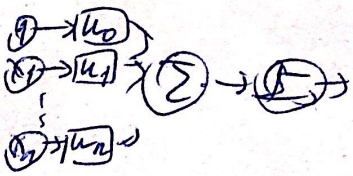


$w_0 = -n+1$, $w_i = 1$ for $i = 1$ to n

$\sum_{i=0}^{n} x_i w = \underbrace{x_0 w_0}_{-n+1} + \sum_{i=1}^{n} x_i w_i$

output is 1 iff all inputs are 1

c) $f(x_1, \dots, x_n) = \bigvee_{i=1}^{n} x_i$?



$w_0 = -n+1$; $w_i = 1$ for $i = 1$ to n

$\sum_{i=0}^{n} x_i w = \underbrace{x_0 w_0}_{-n+1} + \sum_{i=1}^{n} x_i w_i$

output is 1 iff at least one input is 1

α Can you give perceptron for $(x_1 \wedge x_2 \wedge x_3) \vee (\neg x_1 \wedge \neg x_2 \wedge \neg x_3)$?

~~assume there is a perceptron~~

Claim: f.t. is not linearly separable.

proof: assume there is a perceptron satisfies this, w_0 to w_3 s.t. $\sum_{i=1}^{3} w_i x_i > 0$ iff formula is satisfied

input $(0,0,0)$ satisfies: $w_0 \geq 0$

input $(1,0,0)$, $(0,1,0)$, $(0,0,1)$ dissatisfy: $w_0 + w_1 \leq 0$
$w_0 + w_2 \leq 0$
$w_0 + w_3 \leq 0$

input $(1,1,1)$ satisfies the formula, $w_0 + w_1 + w_2 + w_3 > 0$

but this is contradiction because $w_0 \geq 0$ and $\exists$ sum $w_0 + w_1 + w_2 + w_3 \leq 0$

after $w_0 + w_1 + w_2 + w_3$ should be smaller than 0.

α Show every boolean fnct. can be represented by some 2-layer neural network?

every boolean fnct. $f(x_1, \dots, x_n)$ can be written in CNF:

$f = \bigwedge_{i \in I} c_i$ where $c_i = \bigvee_{j \in J} x_j$

find perceptron c_i for each c_i

$c_i = b = 1$

$w_j = \begin{cases} \frac{1}{|c_i|} & \text{if } x_j \in c_i \\ -\frac{1}{|c_i|} & \text{if } \bar{x}_j \in c_i \\ 0 & \text{else} \end{cases}$

so c_i implements c_i, join all c_i together in 2nd layer implementing conjunction over I

$\Lambda = \cdot b = -1$

$w_i = \frac{1}{|x|} + \frac{1}{|x|^2}$

α Suppose $(\vec{x}_1, y_1), \dots, (\vec{x}_n, y_n)$ of training examples with $\vec{x}_i \in \mathbb{R}^d$ and $y_i \in \{+1, -1\}$ s.t. examples
- have bounded Euclidean norm $\|\vec{x}_i\| \leq R$ for $x \in \mathbb{R}$,
- are linearly separable with some margin $\delta > 0$, i.e. exist vector $\vec{v} \in \mathbb{R}^m$ with $\|\vec{v}\| = 1$ s.t. $y_i(\vec{v} \cdot \vec{x}_i) \geq \delta$ for all $i = 1$ to n.

Let $\vec{w}_0 = \vec{0}$ and $\vec{w}_1, \dots, \vec{w}_k \in \mathbb{R}^d$ be vectors calculated iteratively by perceptron training rule. Prove that Block-Novikoff
- Euclidean norm of all examples in D is bounded by R
- positive and negative examples in D are linearly separable with some margin $\delta > 0$, then for the number k of updates in the set: $\vec{w}_{k+1} = \vec{w}_k + y \cdot \vec{x}$ (perceptron rule) of perceptron algo, we have $k \leq \left(\frac{R}{\delta}\right)^2$, proved?

perceptron algo updates weights if classification error was made so it is bounded by $\left(\frac{R}{\delta}\right)^2$. $\vec{w}_i$ denotes vector that was computed after i-th mistake was made

1st claim: $\vec{w}_k^T \vec{v} \geq k\delta$

let i_1 to i_k denote indices of examples where mistake was made.

$\vec{w}_k^T \vec{v} = \left(\sum_{j=1}^{k} x_{i_j} y_{i_j}\right)^T \vec{v} = \sum_{j=1}^{k} y_{i_j} (\vec{x}_{i_j}^T \vec{v}) \geq \sum_{j=1}^{k} \delta = k\delta$

$\vec{w}_k$ arises from $\vec{w}_{k-1}$ by adding $x_{k1} y_{k2} \dots$ and points are separable with margin δ

~~2nd claim~~

↪

- 2nd claim: $\|\vec{w}_k\|^2 \le \|\vec{w}_{k-1}\| + R^2$

Lets note $\|\vec{w}_k\|^2 = \vec{w}_k \vec{w}_k$

$= (\vec{w}_{k-1} + y_{ik}\vec{x}_{ik})(\vec{w}_{k-1} + y_{ik}\cdot\vec{x}_{ik})$

$= \vec{w}_{k-1}\vec{w}_{k-1} + 2y_{ik}\vec{w}_{k-1}\vec{x}_{ik} + y_{ik}\vec{x}_{ik}y_{ik}\vec{x}_{ik}$

observe $2y_{ik}\vec{w}_{k-1}\vec{x}_{ik} \le 0$ because mistake was made,

$y_{ik}y_{ik} = 1$ so we have

$\|\vec{w}_k\|^2 \le \vec{w}_{k-1}\vec{w}_{k-1} + \vec{x}_{ik}\vec{x}_{ik}$

$\le \|\vec{w}_{k-1}\|^2 + \|\vec{x}_{ik}\|^2$

$\le \|\vec{w}_{k-1}\|^2 + R^2$

now we can combine,

by first claim, $(kJ)^2 \le (\vec{w}_k\vec{v})^2 \le \|\vec{w}_k\|^2 \cdot \underbrace{\|\vec{v}\|^2}_{1}$

$(kJ)^2 \le \|\vec{w}_k\|^2$

by second claim $(kJ)^2 \le \|\vec{w}_{k-1}\|^2 + R^2$

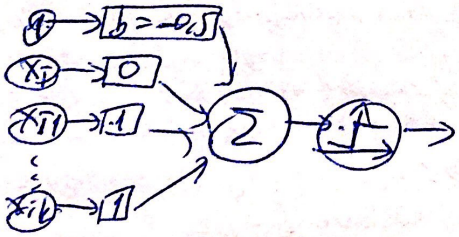
$\le \|\vec{w}_{k-2}\|^2 + 2R^2$

$\vdots$

$(kJ)^2 \le \|\vec{w}_0\|^2 + kR^2 = kR^2$

solving for k, J gives $k \le \frac{R^2}{J^2}$, it is proved.

α 1) Show that any k-literal monotone disjunction over $\{0,1\}^n$ can be represented by perceptron?



$w_i = \begin{cases} 1 & \text{if } i \in \{i_1, \dots, i_k\} \\ 0 & \text{otw} \end{cases}$

so disjunctions over literals can be realized by perceptron.

2) Give upper bound on number of prediction mistakes made by perceptron algo during learning of unknown k-literals monotone disjunctions over $\{0,1\}^n$, prove your upper bound?

Ensure to find separating hyperplane through origin.

Let $(\vec{x}_1, y_1), \dots, (\vec{x}_m, y_m)$ be training set,

transform each $\vec{x}_i = (x_{i1}, \dots, x_{in})$ to $\vec{u}_i = (1, u_{i1}, \dots, u_{in})$

so all are dimension of the input and all points get same cardinality in this new dimension.

The points $\vec{u}_1$ to $\vec{u}_m$ are separable by hyperplane through origin by $\vec{v}' = (-1, v_1', \dots, v_n')$ where

where $v_i' = \begin{cases} 2 & \text{if } i \in \{i_1, \dots, i_k\} \\ 0 & \text{otw} \end{cases}$

then, $\vec{v}'\cdot\vec{x} \ge 1$ if $\vec{x}$ satisfies formula,

$\vec{v}'\cdot\vec{v} \le -1$ otw

normalize $\vec{v}'$, then $\vec{v} = \frac{\vec{v}'}{\|\vec{v}'\|} = \frac{\vec{v}'}{\sqrt{4k+1}}$

apply theorem, radius of any input points is bounded by $\sqrt{n+1}$ and margin is lower bounded by

$J \ge \frac{1}{\sqrt{4k+1}}$

we get $k \le \frac{R^2}{J^2} = (n+1)\cdot(4k+1)$

—