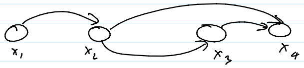
Representations

- Belief Networks

V = Set of variables
$G(V,E)$: directed acyclic graph



p_{x_1} $\quad p_{x_2|x_1}$ $\quad p_{x_3|x_2}$ $\quad p_{x_4|x_3 x_2}$

- easy to compute prob.
- easy to sample

- Markov Networks

$$p(x_1 \ldots x_n) = \frac{\prod_{c \in G} \phi_c(x|_c)}{Z}$$

$$Z = \sum_{x} \prod_{c \in G} \phi_c(x|_c)$$

Inference Algorithms

$P(x_1 | x_3 \, x_4 \ldots x_n)$

Graph is like a tree

- Variable Elimination

- Tail Bounds
  - Markov, Chebyshev, Chernoff
- Markov Chain Monte Carlo
  - Construct Markov Chain s.t. the stationary dist. is same as the dist. we want to sample from
- Ising Model
  - Simulated Annealing

Learning Theory

- PAC Learning Model
- $\mathcal{X}$ input space $\mathcal{Y}$ output space
  $\mathcal{D}$ on $\mathcal{X}$ (unknown)
  $S = \{(x_i, y_i)\}_{i=1}^m$
  x_i - IID rv's from $\mathcal{D}$
- Hypothesis class $\mathcal{H}$
- Realizable Case
  $y_i = f(x_i)$ $\quad f \in \mathcal{H}$
- Agnostic Case
  $(x_i, y_i) \sim \mathcal{D}$

- Gen Error: (realizable case)
  $L_D(h) = \Pr_{x \sim D}[h(x) \neq f(x)]$
  $\qquad\quad P_{x,y \sim D}[h(x) \neq y]$

Constant value, we cannot compute

- Sample Error (Empirical)
  $L_S(h) = \frac{|\{i : h(x_i) \neq y_i\}|}{|S|}$
  a random variable, we can compute
- ERM: Output $h \in \mathcal{H}$ that minimizes sample error.

PAC guarantee

$\Pr_S[L_D(h_S) - L_D(h^*) > \varepsilon] < \delta$

(ε, δ) - PAC guarantee

Sample Complexity

How large should S be for achieving (ε,δ)-PAC guarantee?

For Finite Hypothesis Class

$|S| \approx \frac{1}{\varepsilon^2} \log(|\mathcal{H}|/\delta)$

- Rectangles
  - Output smallest rectangle fitting positive examples

VC dimension

- Shattering: A set S of points is shattered by $\mathcal{H}$ if $|\mathcal{H}_{|S}| = 2^{|S|}$ $\quad \mathcal{Y} = \{0,1\}$
- VCdim$(\mathcal{H}) = d$
  - if there a set of size d that is shattered
  - no set of size $d+1$ is shattered

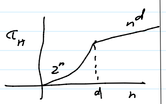
$S = \{x_1 \; x_2 \; x_3 \; x_4\}$

2^4

- Rect - 4
- Hyperplanes $n+1$
- Sauer-Shelah Lemma

$$T_H(n) \le \sum_{i=1}^{d} \binom{n}{i} \equiv n^d$$

$T_H(n) := \max_{S:|S|=n} |\mathcal{H}_{|S}|$



Learning Conjunctions

$\mathcal{X} = \{0,1\}^n$ $\quad \mathcal{Y} = \{0,1\}$

$\mathcal{H} \equiv$ Conjunctions (and of literals)

literal: x_i or $\neg x_i$

$x_1 \wedge x_2 \wedge \neg x_4 \wedge x_{10} \wedge \neg x_3$

n - variables $\qquad f : \mathcal{X} \to \mathcal{Y}$

$|\mathcal{H}| = 3^n + 1$ $\qquad f : \{0,1\}^n \to \{0,1\}$

Algorithm:

- Start $x_1 \wedge \neg x_1 \wedge x_2 \wedge \neg x_2 \ldots x_n \wedge \neg x_n = h$

- for each +ve example
  - find all literals that eval. to 0
  - remove all such literals from h

x	y
1 1 1 1 1 1 1	1
1 1 1 0 0 0 0	0
1 1 1 1 1 0 1 1	1
1 0 1 1 0 0 1 1	0

Note: Suppose $g \in \mathcal{H}$ is the correct conjunction. and h is output of algorithm.

- literals in $h \supseteq$ literals in g
- $L_D(h) = \Pr_{a \sim D}[h(a) \neq g(a)]$
  $= \Pr_{a \sim D}[h(a) = 0 \text{ and } g(a) = 1]$
  $= \Pr[\exists l \text{ in } h \text{ s.t } l(a) = 0 \text{ and } g(a) = 1]$

$\le \sum_{l \in h} \Pr_a[l(a) = 0 \text{ and } g(a) = 1]$

Suppose for all the literals
$\Pr_a[l(a) = 0 \text{ and } g(a) = 1] \le \frac{\varepsilon}{2n}$
$\le 2n \times \frac{\varepsilon}{2n} \le \varepsilon$

Suppose there is literal l in h s.t. $\Pr_a[l(a) = 0 \text{ and } g(a) = 1] > \frac{\varepsilon}{2n}$

What is the probability that $\Pr[l \text{ in } h] \le (1 - \frac{\varepsilon}{2n})^m$

m = sample size

$\Pr\{L_D(h) > \varepsilon\}$
$\le 2n \, (1 - \frac{\varepsilon}{2n})^m = \delta$

$2n \, e^{-\frac{\varepsilon m}{2n}} = \delta$

$\log\left(\frac{2n}{\delta}\right) = \frac{\varepsilon m}{2n}$

$m = \frac{1}{\varepsilon} \, n \log\left(\frac{2n}{\delta}\right)$

Running time $= m$ $\qquad 3^n \cdot m$ ERM algo

"Comp Learning Theory."