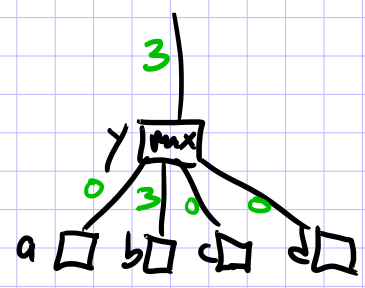
Lecture 22: 4/10/17

Max

$$y = \max(a, b, c, d)$$

$$\frac{dy}{da} = \begin{cases} 1 & \text{if } a \text{ is the max} \\ 0 & \text{otherwise} \end{cases}$$



Matrix Impl. of Conv. Layer

① Define new matrices $\hat{X}, \hat{\alpha}, \hat{y}$ [call this reshaping im2col]

$$\hat{X} = \begin{bmatrix} x_{11} & x_{12} & x_{21} & x_{22} \\ x_{12} & x_{13} & x_{22} & x_{23} \\ x_{21} & x_{22} & x_{31} & x_{32} \\ x_{22} & x_{23} & x_{32} & x_{33} \end{bmatrix}$$

each col is a patch

$$\hat{\alpha} = \begin{bmatrix} \alpha_{11}^{(1)} & \alpha_{11}^{(2)} \\ \alpha_{12}^{(1)} & \alpha_{12}^{(2)} \\ \alpha_{21}^{(1)} & \alpha_{21}^{(2)} \\ \alpha_{22}^{(1)} & \alpha_{22}^{(2)} \end{bmatrix} \qquad \hat{y} = \begin{bmatrix} y_{11}^{(1)} & y_{11}^{(2)} \\ \vdots & \vdots \\ y_{22}^{(1)} & y_{22}^{(2)} \end{bmatrix}$$

② Observe that $\hat{y} = \hat{X}^T \hat{\alpha}$

③ This means that a conv. layer is just concat. of several layers

① im2col

② $\hat{y} = \hat{X}^T \hat{\alpha}$

③ reverse of im2col

★ Can also use for max-pool

Background for Bayes Nets

Def: joint prob. $P(A,B) = P(A|B)P(B)$

Def: chain rule for r.v.s X_1, X_2, X_3, X_4 ← generalizes for arbitrary # of r.v.s!

$$P(X_1, X_2, X_3, X_4) = P(X_1 | X_2, X_3, X_4) P(X_2 | X_3, X_4) P(X_3 | X_4) P(X_4)$$

Def: r.v.s A, B are independent iff ← written $A \perp\!\!\!\perp B$

$$P(A,B) = P(A)P(B)$$

Def: r.v.s A, B are conditionally independent given C iff

$$P(A,B|C) = P(A|C)P(B|C)$$ ← written as $A \perp\!\!\!\perp B \mid C$

Joint Distributions

Ex: Tornado Alarm

Our r.v.s are binary, represent occurence of some event

- $T \Rightarrow$ a tornado
- $H \Rightarrow$ hackers compromised DWAS
- $A \Rightarrow$ 150+ tornado alarms go off
- $P \Rightarrow$ >100 911 Phone Calls
- $F \Rightarrow$ major fire in Dallas

Question: How to represent joint dist. $P(T,H,A,P,F)$

Idea #1: Independence

Assume $P(T,H,A,P,F) = P(T)P(H)P(A)P(P)P(F)$ (1, 1, 1, ...)

- Compact (small # parameters)
- Not expressive enough

#params: $5(2-1)$

Idea #2: Naive Bayes

Assume $P(T,H,A,P,F) = P(A)P(T|A)P(H|A)P(P|A)P(F|A)$

- Compact
- Lopsided: sensible for classification of A but unrealistic for F (e.g. $P(F|A)$)

Idea #3: Full Joint

$p(T,H,A,P,F|\theta)$

T	H	A	P	F	$P(\cdot\|\theta)$
0	0	0	0	0	θ_1
0	0	0	0	1	θ_2
0	0	0	1	0	θ_3
...					...

- Not compact
- Can encode both realistic and unrealistic distributions

#rows: 2^5

#params: $2^5 - 1$ (sum-to-one)

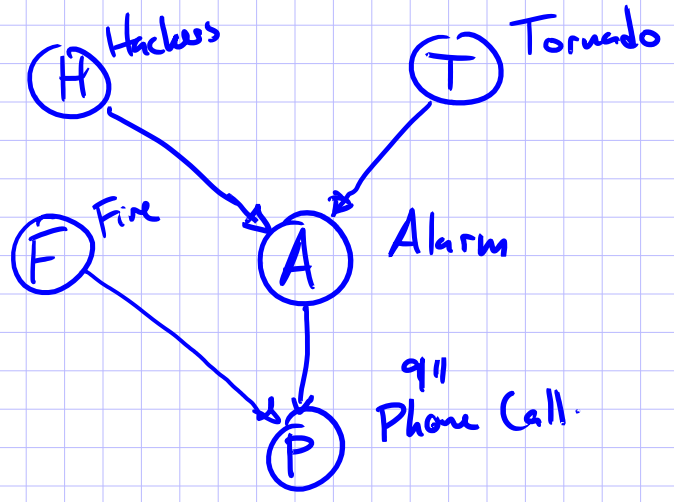
Idea #4: Causality

1. Write out chain
2. Remove RHS variables

$P(T,H,A,P,F) = P(P|A,F,\cancel{T,H})$
$P(F|\cancel{A,T,H})$
$P(A|T,H)$
$P(T\cancel{|H})$
$P(H)$

- Fairly Compact
- Mostly realistic

Ex: Bayesian Network



Bayesian Networks

whose nodes are vars.

Def: a **Bayes Net** consists of a directed acyclic graph G and conditional probabilities P over r.v.s $Y_1, \ldots, Y_J$ s.t.

$$P(Y_1, \ldots, Y_J) = \prod_{j=1}^{J} P(Y_j \mid \text{parents}(Y_j ; G))$$

set of variables which are parents of Y_j in G

Two Parts

1. Qualitative Spec: G ← usually given by an expert
2. Quantitative Spec: P ← usually learned from data.