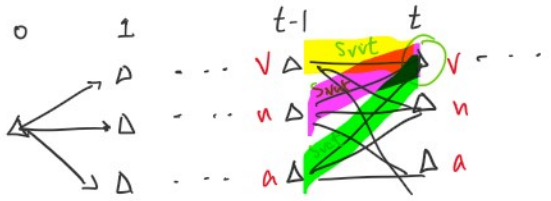
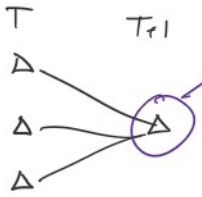
Section A

Wednesday, March 29, 2023 9:31 AM

Viterbi Algo (edge weights)





$W_{T+1}(END)$ = weight of the maximum weight path in the trellis

$$w_t(v) = \max\left(w_{t-1}(v)\, s_{vvt},\ w_{t-1}(n)\, s_{nvt},\ w_{t-1}(a)\, s_{avt}\right)$$

for $t = 1 \ldots T$:

  for $k = 1 \ldots K$:

$$w_t(k) = \max_{j \in \{1,\ldots,K\}} w_{t-1}(j)\, s_{kjt}$$

For HMM:

$$s_{kjt} = p(y_t = k \mid y_{t-1} = j)\, p(x_t \mid y_t = k)$$

$$b_t(k) = \operatorname{argmax}_{j \in \{1,\ldots,K\}} w_{t-1}(j)\, s_{kjt}$$

← backpointer

Ex: CMU going to the moon?

$N = 1$ ⇒ Bloomberg article
$M = 1$ ⇒ CMU is going to the moon
$H = 1$ ⇒ Elaborate hoax
$A = 1$ ⇒ April 1st (or close)
$B = 1$ ⇒ CMU researchers are bored
$G = 1$ ⇒ building control center in GHC
$C = 1$ ⇒ actually just a carnival booth
$S = 1$ ⇒ CMU students are amazing

Poll Q1

Q: How to represent $p(N, M, H, A, B, G, C, S)$?

Idea #1: Condition on the observables

$$P(M, H, B, G, C, S \mid N=1, A=1)$$

Idea #2: Chain Rule

$$P(N, M, \ldots, S) = P(N \mid M \ldots S)\, P(M \mid A \ldots S) \cdots P(S)$$

pro: looks compact

pro: looks compact
con: actually not compact

Idea #3: Full Joint Table

N	M	H	A	B	G	C	S	p(·)
0	0	0	0	0	0	0	0	θ_1
0	0	0	0	0	0	0	1	θ_2
0	0	0	0	0	0	1	0	θ_3
								⋮
1	1	1	1	1	1	1	1	θ_M

pro: can encode any distribution
con: enormous

Idea #4: Full Independence

$$P(N, M, \ldots, S) = P(N)\, P(M) \cdots P(S)$$

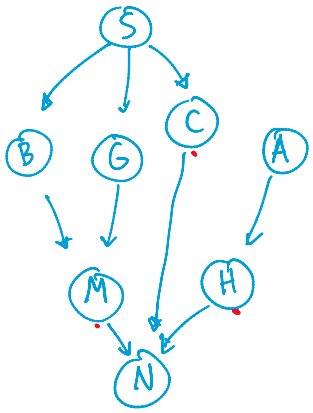
pro: compact
con: completely unrealistic

Idea #5: Naive Bayes (label, feature)

$$P(N, M, \ldots, S) = P(M)\, P(N|M) \cdots P(S|M)$$

pro: still compact
con: missing a lot of dependencies

Idea #6: Bayes Net Network



w/ carefully chosen order

① Write out the chain rule
② Remove some RHS variables

$p(N, M, H, C, B, A, G, S) = P(N|M, H, C, \cancel{B}, \cancel{A}, \cancel{G}, \cancel{S})$
$P(M|\cancel{H}, \cancel{C}, B, \cancel{A}, G, \cancel{S})$
$P(H|\cancel{C}, \cancel{B}, A, \cancel{G}, \cancel{S})$
$P(C|\cancel{B}, \cancel{A}, \cancel{G}, S)$
$P(B|\cancel{A}, \cancel{G}, S)$
$P(A|\cancel{G}, \cancel{S})$
$P(G|S)$

$P(A \mid G, S)$

$P(G \mid S)$

$P(S)$

Proof Cond. Indep

Case #1: Cascade

$$p(x,z|y) = \frac{p(x,y,z)}{p(y)}$$

$$= \frac{p(x|y)\,p(y|z)\,p(z)}{p(y)} \quad \text{(BN prob; } p(y,z) \text{ in numerator)}$$

$$= \frac{p(x|y)\,p(z|y)\,\cancel{p(y)}}{\cancel{p(y)}}$$

$$= p(x|y)\,p(z|y)$$

$$\Rightarrow \not\perp X \perp Z \mid Y$$

Case #2: Case #3: > left as exercises

OH

$error(h_{\theta_{MAP}}) > error(h_{\theta_{MLE}})$

$\Rightarrow$ overfit w/ MAP

$\lambda \sum_n \theta_n^2$

L2 : $r(\theta) = \lambda(\|\theta\|_2)^2$

$J'(\theta) = J(\theta) + r(\theta)$

↑ fit data ↑ keep simple

$$p(\theta) = \prod_{m=1}^{M} p(\theta_m)$$

$p(\theta_m) \sim \text{Gaussian}(\mu = 0, \sigma^2)$

$J(\theta) = -\log p(D|\theta)$

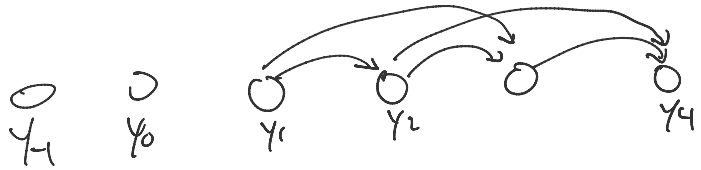
$r(\theta) = -\log p(\theta)$

$$= -\log \prod_{m=1}^{M} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\theta_m - 0)^2}{2\sigma^2}\right)$$

$$= \sum_{m=1}^{M} +\log\left(\sqrt{2\pi\sigma^2}\right) + \frac{\theta_m^2}{2\sigma^2}$$

$$\text{argmin } J'(\theta) = \text{argmin } J(\theta) + \sum_{n=1}^{M} \frac{\theta_n^2}{2\sigma^2} = \text{argmin } J(\theta)$$

$$\lambda = \frac{1}{2\sigma^2}$$

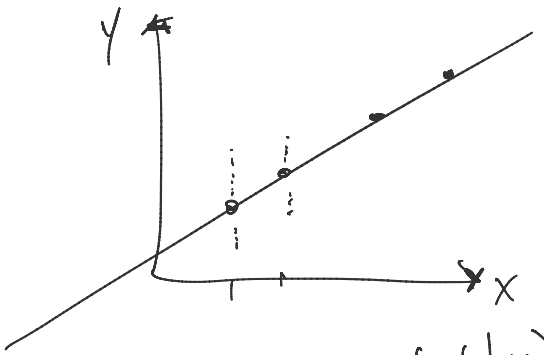


$$p(y_4|y_3,y_2)\, p(y_3|y_2,y_1)\, p(y_2|y_1)\, p(y_1)$$

x_1 weather	x_2 in. of rain	y season
rainy	0.6	S
sunny	0.01	S
cloudy	0.05	F
rainy	0.01	F

regularized cond. ~

x_1'	x_1''	x_1'''	x_2	y
1	0	0	0.6	S
0	1	0	0.01	S
0	0	1	...	~
1	0	0	...	~



$y = wx+b$

$y' \sim wx+b+\epsilon$

$\epsilon \sim N(0,\sigma^2)$

$p(y'|x)$ cond. log. likelihood

$x_0 = \phi_0(x) = x^0$

$x_1 = \phi_1(x) = x^1$

$x_2 = \phi_2(x) = x^2$

1 - 1

$x_2 \quad \phi_2(x) = x$

$p(y=1|\vec{x}) = \sigma(w_1 x_1 + w_0 x_0) \quad d=1$

$p(y=1|\vec{x}) = \sigma(w_1 x_1 + w_2 x_2 + w_0 x_0) \quad d=2$

$\vec{w}' = \underset{\vec{w}}{\operatorname{argmin}}\, J_{d=1}(\vec{w})$

$\vec{w}'' = \underset{}{\operatorname{argmin}}\, J_{d=2}(\vec{w})$

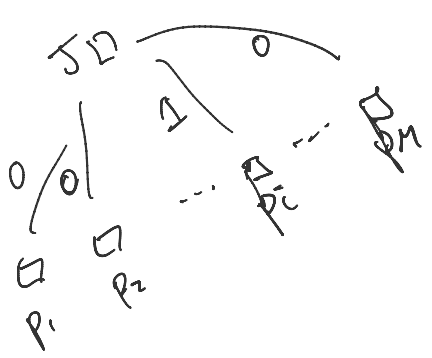
$\vec{w}'' = [w_0', w_1', 0]$

$w_0 \quad w_1 \quad w_2$

$\vec{p} = f(\vec{h})$

$\frac{dJ}{d\vec{p}}$ (vector)

$\frac{dJ}{d\vec{h}} = \left(\frac{d\vec{p}}{d\vec{h}}\right)\left(\frac{dJ}{d\vec{p}}\right)$ (matrix)



$\frac{dJ}{dp_j} = 0 \quad \forall j \neq i$

$\frac{dJ}{dp_j} = 1 \quad \text{for } j = i$