Weak Supervision Nuggets

- Indep case $\longrightarrow$ Simple estimation trick
- Correlations $\longrightarrow$ Inverse Covariance & Graph Structure.

Given: $x^{(1)} \ldots x^{(n)} \in \mathbb{R}^d$

$\lambda_1 \cdots \lambda_m \ : \ \lambda_i : \mathbb{R}^d \to \{-1, 1\} \cup \{\text{abstain}\}$

find $P(y \mid \bar{\lambda}, x^{(i)})$ $\qquad y \in \{-1, 1\}$

Idea: λ_i is a noisy function (inaccurate, incomplete)

λ_1 : "name in dictionary"

λ_2 : "upper case word"

... Programmatic labels ...

Model 0: No abstains, independent errors (clearly correlated)

Each labeler has a hidden accuracy.

with prob p_i $\quad \lambda_j(x^{(i)}) = y^{(i)}$ $\qquad$ "λ_j is right"

$1 - p_i$ $\quad \lambda_j(x^{(i)}) = -y^{(i)}$ $\qquad$ "λ_j is wrong"

Sadly, we don't see y, but we do see $\lambda_j(x^{(i)})$

λ_i, λ_j's errors independent

1. $\mathbb{E}[\lambda_i Y]$ $\qquad$ observe if $\lambda_i \wedge Y$ agree value is 1

$\lambda_i \wedge Y$, disagree value is -1

$\therefore \ = P_i \, 1 + (1 - P_i)(-1)$

$= 2P_i - 1 \triangleq q_i$ (define q_i this way) $\qquad q_i \in [-1, 1]$

② $\mathbb{E}[\lambda_i \lambda_j] = 1$ if $i=j$

$\mathbb{E}[\lambda_i \lambda_j \mid Y=1] = P_1P_2 \cdot 1 + (1-P_1)(1-P_2) \cdot 1$ "Agree"

$(1-P_1)P_2 \cdot (-1) + P_1(1-P_2)(-1)$ "disagree"

$= a_1 \cdot a_2$

NOTE we didn't use $\mid Y=1$, same true for $\mid Y=-1$

$$\mathbb{E}[\lambda_i \lambda_j] = \sum_{b \in \{-1,1\}} \mathbb{E}[\lambda_i \lambda_j \mid Y=b]\, P(Y=b) = a_1 a_2 \xcancel{\sum_b P(Y=b)}^{\,1}$$

didn't need to know $P(Y=b)$.

We form a matrix $M \in \mathbb{R}^{m \times m}$ $\quad M_{ij} = \mathbb{E}[\lambda_i \lambda_j]$

NB: M can be estimated - unlike Y!

"Agreements and disagreements" $\Rightarrow$ Key idea don't need to see y

Simple Algorithm: for any i,j,k distinct, $M_{ij}, M_{jk}, M_{ik} \neq 0$

$$\frac{M_{ij} M_{jk}}{M_{ik}} = \frac{a_i a_j^2 a_k}{a_j a_k} = a_j^2$$

So we can solve upto sign of a_i.

NOTE: If we knew $\text{sign}(a_1) = s$

then $M_{1k} = a_1 a_k$

$\text{sign}(M_{1k})\,\text{sign}(a_1) = \text{sign}(a_k)$ $\therefore$ can solve for all signs. from one.

So a & $-a$ are solutions...

Assume $\sum_{i=1}^{n} a_i > 0$, breaks symmetry "good on average".

- What if $M_{ij} = 0$? $a_i = 0$ or $a_j = 0$.

this means $a_i = 0 \Rightarrow 2P_i - 1 = 0 \Rightarrow P_i = \frac{1}{2} \Rightarrow$ Random noise!

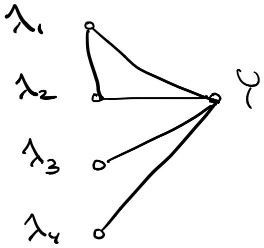
have every a_i bounded away from random noise

(CAN HANDLE w/ fancier tricks).

THEORY SAYS " let $\delta = \min_{j=1..m} |p_j - \frac{1}{2}|$, NEED SAMPLES PROPORTIONAL TO $\frac{1}{\delta^2}$ " (lot of work)

RECAP: Symmetry, simple ALGEBRAIC learns (go called we GD)

<u>WHAT if LABELERS ARE CORRELATED?</u>

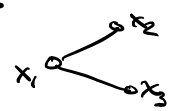


<u>Key Concept</u>: PROBABILITY distribution on GRAPHS.

$$\mathbb{E}[\lambda_i \lambda_j | Y] = \mathbb{E}[\lambda_i | Y]\, \mathbb{E}[\lambda_j | Y]$$

if $(i,j) \notin$ Edge ABOVE.

<u>Nugget</u> Structure of INVERSE COVARIANCE for GAUSSIANS



$x_1 \sim N(0,1)$

$x_2 \sim N(x_1, 1)$ $\quad \epsilon_2 \sim N(0,1) \quad x_2 = x_1 + \epsilon_2$

$x_3 \sim N(x_2, 1)$ $\quad \epsilon_3 \sim N(0,1) \quad x_3 = x_1 + \epsilon_3$

1. $\mathbb{E}[x_1] = 0 \qquad \mathbb{E}[x_2] = \cancelto{0}{\mathbb{E}[x_1]} + \cancelto{0}{\mathbb{E}[\epsilon_2]} = 0$

$\mathbb{E}[x_3] = 0.$

$\mathbb{E}[x_1^2] = 1 \qquad \mathbb{E}[x_2^2] = \mathbb{E}[(x_1 + \epsilon_2)^2] = \mathbb{E}[x_1^2] + 2\cancelto{0}{\mathbb{E}[x_1 \epsilon_2]}$

$+ \mathbb{E}[\epsilon_2^2]$

$= 2$

$\mathbb{E}[x_1 x_2] = \mathbb{E}[x_1^2 + x_1 \epsilon_1] = 1$

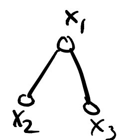
$\mathbb{E}[x_2 x_3] = \mathbb{E}[(x_1 + \epsilon_1)(x_1 + \epsilon_3)] = 1$

$$\Sigma = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}$$

NO REAL STRUCTURE?

$$\Sigma^{-1} = \begin{bmatrix} 3 & -1 & -1 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}$$

No Edge



WE SAY A <u>Probability distribution</u> $p : \mathbb{R}^d \to [0,1]$ agrees or factorizes wrt A graph $G = (V, E)$ if

$$p(x) = C \prod_{e=(v_i,v_j)\in E} p_e(x_i, x_j) \cdot \prod_{v \in V} p_v(x_i)$$

for some functions!

Normalization Constant

Now, let's look @ Gaussians over a graph.

$$\log \exp\{x^T \Sigma^{-1} x\} = \log \prod p_e(x_i, x_j) \prod_{v\in V} p_v(x_i) \quad \text{(ignore } c\text{)}$$

$$x^T \Sigma^{-1} x = \sum_{e\in E} \log p_e(x_i, x_j) + \sum_{v \in V} \log p_v(x_i)$$

let $A = \Sigma^{-1}$ for easy notation

$$\sum_{i,j} A_{ij} x_i x_j =$$

for $i,j \in E$ $\quad \nabla_{x_i x_j} = (A_{ij} + A_{ji}) = 0$ if $(i,j) \notin E$.

But $\Sigma^{-1} = A$ is symmetric, so $A_{ij} = 0$.

Thus, if A Gaussian factors on A graph, Entries of inverse Are 0!

More complex theory for discrete r.v.s   Loh & Wainwright 2014
Ratner et al 2018

Back to our problem

Form "covariance matrix"

$$\begin{array}{c|ccc|c} & \lambda_1 & \lambda_2 & \lambda_3 & y \\ \hline \lambda_1 & & & & \\ \lambda_2 & & \Sigma_o & & \\ \lambda_3 & & & & \\ \hline y & & & & 1 \end{array}$$

We "see" Σ_o but not all of sigma $\mathbb{E}[\lambda_i Y]$ is unobserved.

let $o = \{1,2,3\}$ ⌐ visible terms

$$(\Sigma^{-1})_o = (\Sigma_o - uu^T)^{-1}$$

u → some rank 1 vector.

$(\Sigma^{-1})_o$ → inverse on those entries

let $B = \Sigma_o$

$$(B - uu^T)^{-1} = B^{-1} + \frac{B^{-1}uu^TB^{-T}}{1 - u^TB^{-1}u} \qquad z = \frac{B^{-1}u}{\sqrt{1 - u^TB^{-1}u}}$$

So $(\Sigma_o)^{-1} = \Sigma_o^{-1} + zz^T$

Now, if $(i,j) \notin E$ then we $(\Sigma_o)_{ij} = B_{ij} = 0$.

Hence $(\Sigma_o^{-1})_{ij} = -z_i z_j$

$$(B_{ij})^2 = z_i^2 z_j^2 \mapsto \log B_{ij}^2 = \log z_i^2 + \log z_j^2$$

This is a linear system in z_i^2 & z_j^2, and we can solve (if enough pairwise indep)

In the notes:

- Higher rank versions (handle more correlations)
- How to learn graph structure
- How to handle sampling error (modern ML, lower bounds)

Recap:

- Weak supervision formal theory
- Nuggets about Graphs & Prob. distributions (take graphical models!)
- "Method of moments" style Algorithms