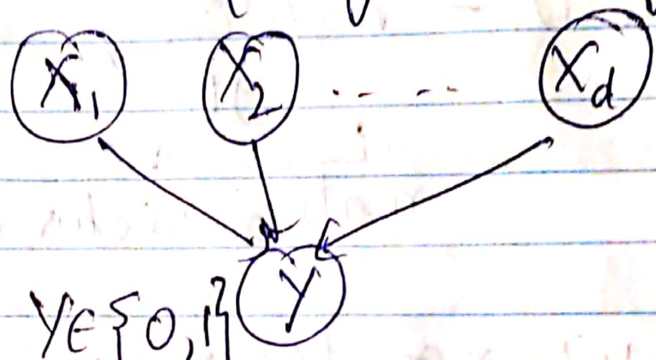
Lecture 10

Case IIB Sigmoid CPTs (logistic regression)

$\vec{X} \in \mathbb{R}^d$



$Y \in \{0,1\}$

- sigmoid CPT
  $$P(Y=1|\vec{X}) = \sigma(\vec{w} \cdot \vec{x})$$
  with $\sigma(z) = \dfrac{1}{1+e^{-z}}$

properties of $\sigma(z)$:

- $\sigma(z) = 1 - \sigma(-z)$
  $$\frac{d}{dz}\sigma(z) = \sigma(z)\,\sigma(-z)$$

* Training Examples: (IID)
$$\{(\vec{X}_t, y_t)\}_{t=1}^{T}$$

- log (conditional) likelihood:
  $$\mathcal{L}(\vec{w}) = \log P(\text{data})$$

$$= \log \prod_{t=1}^{T} P(Y = y_t \mid \vec{X} = \vec{x}_t) \quad \text{(IID)}$$

$$= \sum_t \log P(Y = y_t \mid \vec{X} = \vec{x}_t)$$

$$\cancel{\mathcal{L}(\vec{w}) = \sum_{t=1}^{T} \left[\sigma(\vec{w} \cdot \vec{x}_t)^{y_t} \right.}$$

$$\mathcal{L}(\vec{w}) = \sum_{t=1}^{T} \log \left[\sigma(\vec{w} \cdot \vec{x}_t)^{y_t} \, \sigma(-\vec{w} \cdot \vec{x}_t)^{1-y_t} \right]$$

with $y_t \in \{0,1\}$

$$\mathcal{L}(\vec{w}) = \sum_{t=1}^{T} y_t \log \sigma(\vec{w} \cdot \vec{x}_t) + (1-y_t) \log \sigma(-\vec{w} \cdot \vec{x}_t) \Big]$$

To maximize this expression:

$$0 = \frac{\partial \mathcal{L}}{\partial w_\alpha} = \sum_t \left[y_t \frac{1}{\sigma(\vec{w} \cdot \vec{x}_t)} \sigma(\vec{w} \cdot \vec{x}_t) \sigma(-\vec{w} \cdot \vec{x}_t) x_{\alpha t} + \right.$$

$$\left. (1-y_t) \frac{1}{\sigma(-\vec{w} \cdot \vec{x}_t)} \sigma(\vec{w} \cdot \vec{x}_t) \sigma(-\vec{w} \cdot \vec{x}_t) (-x_{\alpha t}) \right]$$

$$= \sum_t x_{\alpha t} \left[y_t \sigma(-\vec{w} \cdot \vec{x}_t) - (1-y_t) \sigma(\vec{w} \cdot \vec{x}_t) \right]$$

$$= \sum_t x_{\alpha t} \left[y_t (1 - \sigma(\vec{w} \cdot \vec{x}_t)) - (1-y_t) \sigma(\vec{w} \cdot \vec{x}_t) \right]$$

$$= \sum_t x_{\alpha t} \left[y_t - \sigma(\vec{w} \cdot \vec{x}_t) \right]$$

difference between target value $y_t \in \{0,1\}$ and $P(Y=1 \mid \vec{X}_t)$ modeling our prediction.

$$0 = \frac{\partial \mathcal{L}}{\partial w_\alpha} \quad \text{for } \alpha = 1, 2, \ldots, d$$

These are NON-LINEAR equations.

- **Hessian Matrix:**

$$H_{\alpha\beta} = \frac{\partial^2 \mathcal{L}}{\partial w_\alpha w_\beta} = \frac{\partial}{\partial w_\beta} \left\{ \sum_t \left[y_t - \sigma(\vec{w} \cdot \vec{x}_t) \right] x_{\alpha t} \right\}$$

$$= -\sum_t \sigma(\vec{w} \cdot \vec{x}_t)\, \sigma(-\vec{w} \cdot \vec{x}_t)\, x_{\beta t}\, x_{\alpha t}$$

Vector form:

$$\frac{\partial L}{\partial \vec{w}} = \sum_t \left[y_t - \sigma(\vec{w} \cdot \vec{x}_t) \right] \vec{x}_t$$

$$\frac{\partial^2 L}{\partial \vec{w} \partial \vec{w}^T} = -\sum_t \sigma(\vec{w} \cdot \vec{x}_t)\, \sigma(-\vec{w} \cdot \vec{x}_t)\, \vec{x}_t\, \vec{x}_t^{\,T}$$

ML Estimation:

1) Gradient ~~Descent~~ Ascent : update $\vec{w} \leftarrow \vec{w} + \eta \left(\frac{\partial L}{\partial \vec{w}} \right)$

suggest: $\eta = \frac{0.2}{T}$

2) Newton's Method:

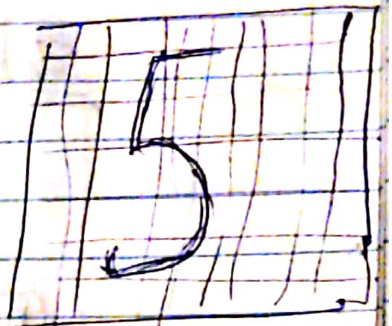
update: $\vec{w} \leftarrow \vec{w} - H^{-1} \left(\frac{\partial \mathcal{L}}{\partial \vec{w}} \right)$

suggest $\vec{w} = [0, \ldots\ldots 0]^T$

HW5: Classify

$y=0$

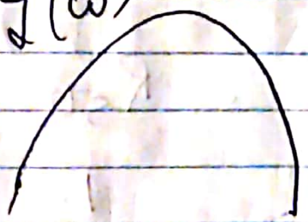
vs



$y=1$

- Global Optimality: ← it can be shown that $\mathcal{L}(\vec{w})$ is concave for logistic regression and has no spurious local maxima.

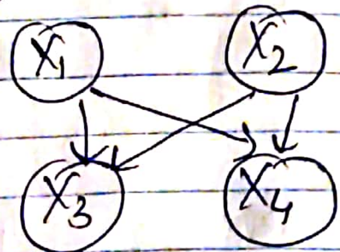
$\mathcal{L}(\vec{w})$



Case III: fixed DAG, discrete nodes, lookup CPTs,

INCOMPLETE DATA.

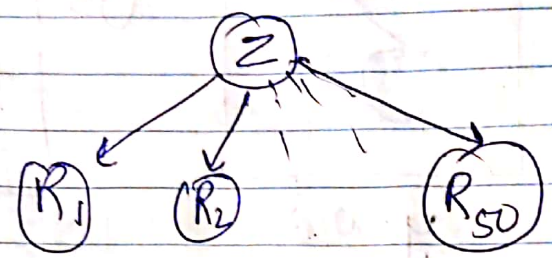
TOY Example:



binary nodes $X_i \in \{0,1\}$

t	X_1	X_2	X_3	X_4
1	?	1	0	1
2	0	?	1	1
3	?	?	0	1
⋮	⋮	⋮	⋮	⋮
T				

Ex: Movie Recommender System:



$Z \in \{1, 2, \dots K\}$ types of movie goer

$R_i \in \{0, 1\}$ movie rating

$i = 1$ Avengers
$i = 2$ Toy Story
$i = 3$ Star Wars

t	Z	R_1	R_2	. . .	R_{50}
1	?	1	?		1
2	?	?	0		0
3	?	?	?		1
⋮	⋮	⋮	⋮		
256	?				

↓
no. of students

* Variables in BN

H = set of Hidden (unobserved) variables
V = set of visible (observed) variables

Can vary with example

$X = HUV$ (all nodes)

* <u>log-likelihood:</u>

$$\mathcal{L} = \log P(\text{data})$$

$$= \log \prod_{t=1}^{T} P(V^{(t)})$$ marginal probs.

visible nodes on t^{th} example

$$= \sum_t \log P(V^{(t)})$$

$$= \sum_t \log \sum_h P(H=h, V^{(t)})$$ marginalization.

$$\mathcal{L} = \sum_t \log \sum_h \left\{ \prod_{i=1}^{n} P(X_i = x / pa_i = \pi) \right\} \Bigg|_{\substack{H=h \\ V=V^{(t)}}}$$ in form of CPTs

For complete data:
CPTs decoupled $\Leftrightarrow$ many independent optimizations.

Now, for incomplete data:
many (or all) CPTs are coupled.

How to optimize?

Options?

1.) Gradient Descent: $\vec{\theta} \leftarrow \vec{\theta} + \eta \frac{\partial L}{\partial \vec{\theta}}$ must tune $\eta > 0$ asymptotic but not monotonic convg.

2.) Newton's Method: $\vec{\theta} \leftarrow \vec{\theta} - H^{-1} \frac{\partial L}{\partial \vec{\theta}}$ expensive, fast but unstable.

3.) New method: Auxilliary functions $Q(\vec{\theta}, \vec{\theta}')$.
How to minimize $f(\vec{\theta})$?

- Suppose $Q(\vec{\theta}, \vec{\theta}')$ satisfies two properties

EQUALITY (I.) $Q(\vec{\theta}, \vec{\theta}) = f(\vec{\theta})$

BOUND (II.) $Q(\vec{\theta}', \vec{\theta}) \geq f(\vec{\theta}')$ $\forall \vec{\theta}, \vec{\theta}'$

Consider update rule:

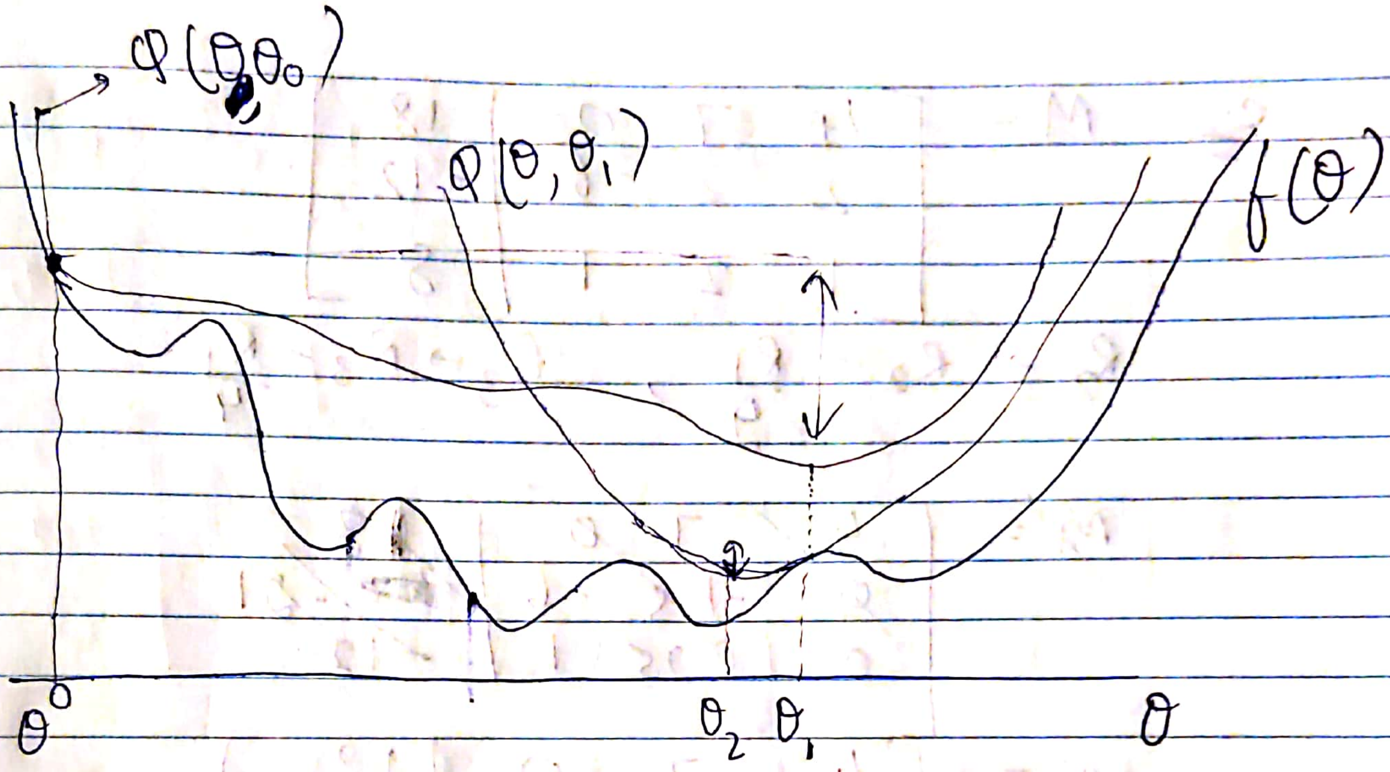
$$\vec{\theta}_{new} = \underset{\vec{\theta}}{\operatorname{argmin}}\, Q(\vec{\theta}, \vec{\theta}_{old})$$

Now, $f(\vec{\theta}_{new}) \leq Q(\vec{\theta}_{new}, \vec{\theta}_{old})$ by property (II.)

$\leq Q(\vec{\theta}_{old}, \vec{\theta}_{old})$ by update rule

$= f(\vec{\theta}_{old})$ by property (I.)

By iterating: $f(\vec{\theta}_0) \geq f(\vec{\theta}_1) \ldots \geq f(\vec{\theta}_n)$



- Properties:
  - no learning rate.
  - monotonic improvement.
  - convergence to stationary point where gradient vanishes: $\frac{\partial f}{\partial \vec{\theta}}$.