Pillow 2008 reading notes 2/14/2022

eq. 3.2 $$p(y \mid x, \theta) = \prod_i p(y_i \mid x_i, \theta) \quad \text{where } \theta = \{k, \phi_f\}$$

↳ params of fn f

Poisson likelihood: $p(x_j) = e^{-\lambda_0} \frac{1}{x_j!} \lambda_0^{x_j}$

let $\lambda_0 \Leftarrow \Delta f(k \cdot x_i)$, then:

eq. 3.3 $$p(y_j \mid x_j, \theta) = e^{-\Delta f(k \cdot x_j)} \frac{1}{y_j!} \left(\Delta f(k \cdot x_j)\right)^{y_j}$$

$$p(y \mid x, \theta) = \prod_i \left[e^{-\Delta f(k \cdot x_i)} \frac{1}{y_i!} \left(\Delta f(k \cdot x_i)\right)^{y_i}\right]$$

$$= \prod_i \left[e^{-\Delta f(k \cdot x_i)} \left(\Delta^y f(k \cdot x_i)^y\right) / y_i!\right]$$

eq. 3.4 $$= \Delta^n \prod_i \left[e^{-\Delta f(k \cdot x_i)} f(k \cdot x_i)^y / y_i!\right] \quad \text{where } n = \sum_i y_i$$

find MLE of $\hat{\theta} = \{\hat{k}, \hat{\phi}_f\}$ by max. log-likelihood:

$$\log\left[p(y \mid x, \theta)\right] = \log\left[\Delta^n \prod_i \left[e^{-\Delta f(k \cdot x_i)} f(k \cdot x_i)^y / y_i!\right]\right]$$

$$= \log \Delta^n + \log\left[\prod_i e^{-\Delta f(k \cdot x_i)}\right] + \log\left[\prod_i f(k \cdot x_i)^y\right] - \log\left[\prod_i y_i!\right]$$

$$= c_1 + \sum_i \log\left(e^{-\Delta f(k \cdot x_i)}\right) + \sum \log\left(f(k \cdot x_i)^y\right) + c_2$$

eq. 3.5 $$= -\Delta \sum_i f(k \cdot x_i) + \sum_i y_i \log\left(f(k \cdot x_i)\right) + c$$

constrain k to be a unit vector: $\|k\| = 1$

⇒ $k \cdot x_i$ will only change direction of x_i, not magnitude

now estimate k using "angular error"

$$\frac{d}{dk}\left(\log\left[p(y|x,\theta)\right]\right) = \frac{d}{dk}\left[-\Delta \sum_i f(k\cdot x_i) + \sum_i y_i \log(f(k\cdot x_i)) + c\right]$$

$$= -\Delta \sum_i x_i \cdot f'(k\cdot x_i) + \sum_i y_i \cdot x_i \cdot f'(k\cdot x_i) \cdot \frac{1}{f(k\cdot x_i)} + 0$$

$$= -\Delta \sum_i x_i f'(k\cdot x_i) + \sum_i y_i x_i \frac{f'(k\cdot x_i)}{f(k\cdot x_i)}$$

use "method of Lagrange multipliers to find the local max / min:

find max/min of $f(x)$ w/ constraint $g(x)=0$, then

$$\mathcal{L}(x,\lambda) = f(x) - \underbrace{\lambda}_{\text{"Lagrange multiplier"}} g(x)$$

set $\frac{d}{dx}\left[\log(p(y|x,\theta))\right] = 0 = g(x)$ (where x here is our k)

$$\mathcal{L}(k,\lambda) = \left(-\Delta \sum_i f(k\cdot x_i) + \sum_i y_i \log(f(k\cdot x_i)) + c\right) - \lambda\left[-\Delta \sum_i x_i f'(k\cdot x_i) + \sum_i y_i x_i \frac{f'(k\cdot x_i)}{f(k\cdot x_i)}\right]$$

↳

- find critical points of $\mathcal{L}$
- find which critical point gives the min/max w.r.t k, λ
- that λ is the "Lagrange multiplier"

$STA = \frac{1}{n}\sum_i y_i x_i$ $\therefore$ $\hat{k}$ is a weighted STA w/ weights $\frac{f'}{f}$

$\hat{k}$ is same as ordinary STA if $f(z) = e^{az+b}$ i.e. is exponential

$f'(z) = (a)e^{az+b}$

$\Rightarrow f'/f = ae^{az+b} / e^{az+b} = a$

$$p(y_i \mid x_i, \theta) = \frac{1}{y_i!} \left[\Delta f(k \cdot x_i) \right]^{y_i} e^{-\Delta f(k \cdot x_i)}$$

if : $f(z) = e^{az+b}$

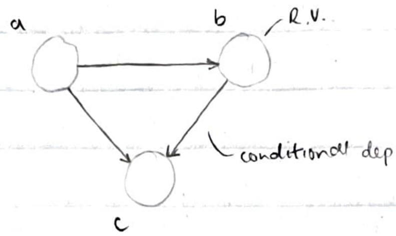
$$p(y_i \mid x_i, \theta) = \frac{1}{y_i!} \left[\Delta e^{(a(k \cdot x_i)+b)} \right]^{y_i} e^{-\Delta e^{(a(k \cdot x_i)+b)}}$$

$=$

Notes on PRML ch. 8 2/15/2022

by the product rule (of prob):

$$p(a,b,c) = p(c|a,b)\,p(a,b) = p(c|a,b)\,p(b|a)\,p(a)$$



For a graph of K nodes, the joint dist.: $p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k | pa_k)$

where pa_k denotes the parents of x_k

example: polynomial regression

$\mathbf{w}$: vector of polynomial coef.

$\mathbf{t}$: vector of observed data $= (t_1, \ldots, t_N)^T$

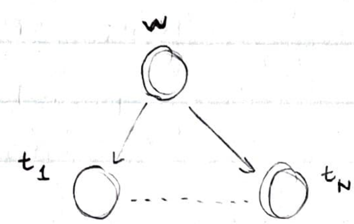
$\mathbf{x}$: input data vector $= (x_1, \ldots, x_N)^T$

σ^2 : variance

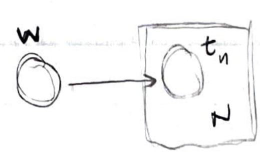
α : hyperparam. representing precision of the Gaussian prior over $\mathbf{w}$

joint dist.:

$$p(\mathbf{t}, \mathbf{w}) = p(\mathbf{w}) \prod_{n=1}^{N} p(t_n | \mathbf{w})$$



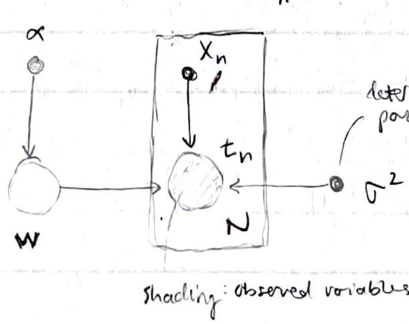
equivalent to



re-write joint dist. more explicitly:

$$p(\mathbf{t}, \mathbf{w} | \mathbf{x}, \alpha, \sigma^2) = p(\mathbf{w}|\alpha) \prod_{n=1}^{N} p(t_n | \mathbf{w}, x_n, \sigma^2)$$

more explicitly



shading: observed variables

posterior distribution of $\mathbf{w}$:

$$p(\mathbf{w} \mid \mathbf{T}) \propto \underbrace{p(\mathbf{w})}_{\text{prior}} \underbrace{\prod_{n=1}^{N} p(t_n \mid \mathbf{w})}_{\text{likelihood}}$$

8.2 Conditional Independence

given: $p(a \mid b, c) = p(a \mid c)$ $\therefore$ a is independent of b given c

alternative expression: $p(a, b \mid c) = p(a \mid b, c)\, p(b \mid c)$

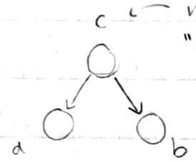
$$= p(a \mid c)\, p(b \mid c)$$

"a & b are statistically independent given c"

↳ shorthand: $a \perp\!\!\!\perp b \mid c$

<u>example 1</u>:

$p(a,b,c) = p(a \mid c)\, p(b \mid c)\, p(c)$ →



← node c is "tail-to-tail"

test if a & b are independent by marginalizing out c:

$$p(a,b) \stackrel{?}{=} p(a)\,p(b) = \sum_c p(a \mid c)\, p(b \mid c)\, p(c)$$

→ generally does not factorize ⇒ $a \not\!\perp\!\!\!\perp b \mid \emptyset$

└ "empty set"

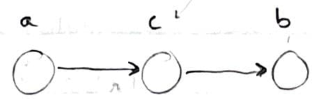
instead show a & b are independent w.r.t c:

$$p(a,b \mid c) = p(a,b,c) / p(c) = p(a \mid c)\, p(b \mid c)$$

⇒ $a \perp\!\!\!\perp b \mid c$

$\therefore$ a & b are conditionally independent

"conditioning on node c "blocks" the path b/w. a & b" node c is "head-to-tail"



example 2:

given:

$$p(a,b,c) = p(a)\,p(c|a)\,p(b|c)$$

test if a & b are independent by marginalizing over c:

$$p(a,b) \stackrel{?}{=} p(a)p(b) = p(a) \sum_c p(c|a)\,p(b|c)$$
$$= p(a)\,p(b|a)$$

→ generally does not factorize to $p(a)p(b)$ $\therefore$ $a \not\perp\!\!\!\perp b \mid \emptyset$

if condition on c:

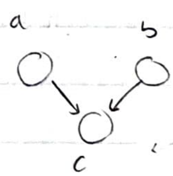
$$p(a,b|c) = p(a,b,c)/p(c)$$
$$= p(a)p(c|a)p(b|c)/p(c)$$
$$= p(a|c)\,p(b|c)$$

$\Rightarrow$ $a \perp\!\!\!\perp b \mid c$ "a & b are conditionally independent w.r.t. c"

example 3:

given:

$$p(a,b,c) = p(a)p(b)p(c|a,b) \longrightarrow$$



node c is "head-to-head"

test if a & b are independent by marginalizing over c: $p(a,b) = p(a)p(b)$

$\Rightarrow$ $a \perp\!\!\!\perp b \mid \emptyset$

test if a & b are conditionally independent on c:

$$p(a,b|c) = p(a,b,c)/p(c)$$
$$= p(a)p(b)p(c|a,b)/p(c)$$

→ does not generally factorize into $p(a),p(b)$ $\therefore$ $a \not\perp\!\!\!\perp b \mid c$

- if c is not observed, then a & b are independent
- if c is observed, then a & b are dependent

Notes on PRML ch. 9. Mixture Models and EM 2/16/2022

9.2 Mixtures of Gaussians

Gaussian mixture : linear superposition of Gaussians :

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k N(\mathbf{x} | \mu_k, \Sigma_k)$$

$\mathbf{z}$: K-dim binary r.v. where; $z_k \in \{0,1\}$ and $\sum_k^K z_k = 1$

$p(z_k = 1) = \pi_k$ where $0 \leq \pi_k \leq 1$

$\sum_k^K \pi_k = 1$

let $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{z}) p(\mathbf{x}|\mathbf{z})$

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k} \quad \leftarrow \{0 \text{ or } 1\} \qquad p(\mathbf{x} | z_k = 1) = N(\mathbf{x} | \mu_k, \Sigma_k)$$

$$p(\mathbf{x} | \mathbf{z}) = \prod_{k=1}^{K} N(\mathbf{x} | \mu_k, \Sigma_k)^{z_k} \quad \leftarrow \{0,1\}$$

get $p(\mathbf{x})$ by marginalizing out $\mathbf{z}$ from joint dist. $p(\mathbf{x}, \mathbf{z})$:

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

sum over all values of $\mathbf{z}$:

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z}) p(\mathbf{x} | \mathbf{z})$$

$$= \sum_{\mathbf{z}} \left[\prod_{k=1}^{K} \pi_k^{z_k} \cdot \prod_{k=1}^{K} N(\mathbf{x} | \mu_k, \Sigma_k)^{z_k} \right]$$

?

$$= \sum_{k=1}^{K} \pi_k N(\mathbf{x} | \mu_k \Sigma_k)$$

* for each observation $\mathbf{x}_n$, there is a corresponding $\mathbf{z}_n$

9.2.2. EM for Gaussian mixtures

conditions that must be satisfied at a max. of the likelihood:

log likelihood: $\ln p(\mathbf{X}|\pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln \left[\sum_{k=1}^{K} \pi_k N(x_n | \mu_k, \Sigma_k) \right]$

$\frac{d}{d\mu} \ln p(\mathbf{X}|\pi, \mu, \Sigma)$; set to 0:

$$0 = - \sum_{n=1}^{N} \left[\frac{\pi_k N(x_n | \mu_k, \Sigma_k)}{\sum_j \pi_j N(x_n | \mu_j, \Sigma_j)} \right] \Sigma_k (x_n - \mu_k)$$

then $\mu_k = \frac{1}{N_k} \sum_{k=1}^{N} \gamma(z_{nk}) x_n$

where: $N_k = \sum_{n=1}^{N} \gamma(z_{nk})$ "effective # of data pts. assigned to class k"

$\gamma(z_{nk}) \equiv p(z_{nk} = 1 | x_n)$ see eq. 9.13 for derivation

∴ mean of k^{th} Gaussian μ_k is the mean of all the data points weighted by their probability of being in class k

$\frac{d}{d\Sigma} \ln p(\mathbf{X}|\pi, \mu, \Sigma)$; set to 0:

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(x_n - \mu_k)(x_n - \mu_k)^T$$

∴ covariance matrix for Gaussian k (Σ_k) is a weighted average of covariance of all data points

$\frac{d}{d\pi} \ln p(\mathbf{X}|\pi, \mu, \Sigma)$; set to 0: $\pi_k = \frac{N_k}{N}$

posterior prob. given x for class k

∴ the mixing coef. for class k is the average "responsibility" which the component takes for explaining the data points

9.3. An alternative view of EM

* goal of EM : find the ML (max. likelihood) solutions for models having latent variables

$\mathbf{X}$: all observed data ; n^{th} row represents x_n^T

$\mathbf{Z}$: all latent vars. ; n^{th} row represents z_n^T

θ : all model parameters

log likelihood :

$$\ln p(\mathbf{X}|\theta) = \ln\left[\sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\theta)\right]$$

complete data set : $\{\mathbf{X},\mathbf{Z}\}$ as if we knew the latent variable values

incomplete data set : $\mathbf{X}$ alone

Expectation step : E

1. use the current values θ^{old} to find the posterior dist. of the latent variables by $p(\mathbf{Z}|\mathbf{X},\theta^{old})$
2. use this posterior dist. to find the expectation of the complete-data log likelihood :

$$Q(\theta,\theta^{old}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}\theta^{old}) \ln p(\mathbf{X},\mathbf{Z}|\theta)$$

Maximization step : M

3. determine the revised paramter estimate θ^{new} by maximizing $Q(\theta,\theta^{old})$

$$\theta^{new} = \operatorname{argmax}_{\theta} Q(\theta,\theta^{old})$$

Session 4 Lecture Notes — From single neurons to pop. dynamics 2/16/2022

1. Generalized linear models recap

2. Conditional independence & graphical models
   - when talking about d-sep., "observing" a value for a node = conditioning on the node

3. Mixture models and the EM algorithm
   - latent var. formalization of Gaussian mix. model is done so can use E.M.

Exercise #2 notes

2/20/2022

[black → left   white → right]

DATA

"activity" : [timepoint x neuron] spikecounts

"cue" : experimental cue ; 0 = black , 1 = white

"cho" : choice ; 0 = left , 1 = right

"corr" : correct choice ; 0 = incorrect , 1 : correct

"prev_corr" : previous outcome ; " " " "

section 2.3 : Fitting Poisson GLM

$\log(y) = w_0 + w_1 x_1 + \ldots$ w: weights   x: data/features