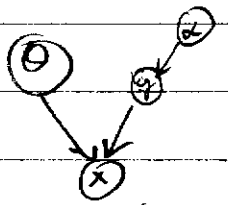
-classifier

discrete (multinomial)
vector (Gaussian)

i.e. for M classes have M-Gaussians

$p(y,x)$



$$p(x,y|\theta) = \left(\prod_j \alpha_j^{y^j}\right)\left(\sum_m \delta(y=m)\, N(x|\mu_m, \Sigma_m)\right)$$

$$\sum_i \log p(x_i,y_i|\theta) = \sum_i \log p(y_i,x_i|\theta)$$

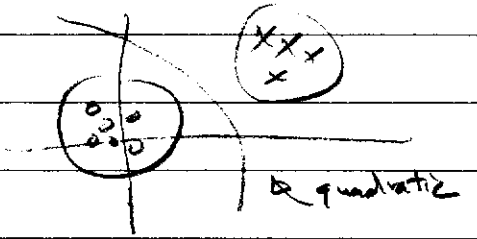
$$= \sum_i \log\left(\prod_j \alpha_j^{y_i^j}\right)\left(\sum_m \delta(m=y_i)\, N(x|\mu_m,\Sigma_m)\right)$$

$$= \sum_i \sum_j y_i^j \log \alpha_j + \sum_i \log \sum_m \delta(m=y_i)\, N(x_i|\mu_m,\Sigma_m)$$

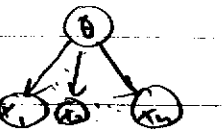
$\frac{\partial}{\partial \mu_m}$ ↓

$$\sum_i \delta(m=y_i) \log N(x_i|\mu_m,\Sigma_m)$$

$$\mu_m = \frac{\sum_i x_i\, \delta(m=y_i)}{\sum_i \delta(m=y_i)}$$

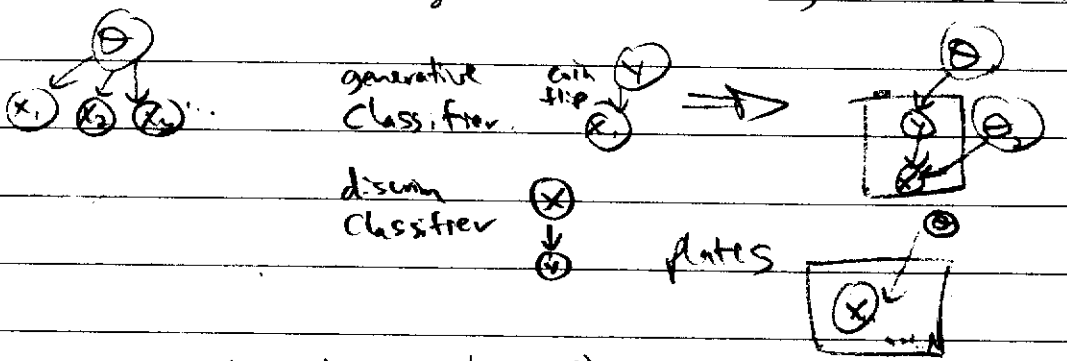


- Readings: Chapter 2 (first half)
  Chapter 4
  Chapter 7 (next week)

- Quick Review of Graphical Models: $p(x_i|pa(x_i))$ 
- Quick Rev of why pdf (all vars)
- Quick Rev. of Bayesian Learning post = $\frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$

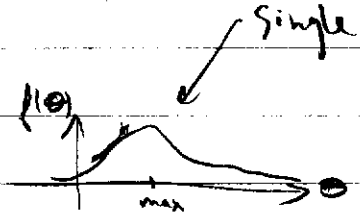
$$p(x|X) \propto \int p(x|\theta) \prod p(x_i|\theta)\, p(\theta)\, d\theta$$

- Quick Rev of ML → intuitive mean & cov → ML
- Questions / Conversations / Other Distribs ⇒ e-family
  - ↳ scalar coins discrete + integer
  - nice properties for ML / Bayes
- Review Optimization
- MAP Gaussian & Bayesian Gaussian (page 8-8b13)
- Bernoulli / Multinomial / simplest hypercube / heuristics then ML derivation (page 7-8)
- Generative Models → regenerate the data, later Discrim.

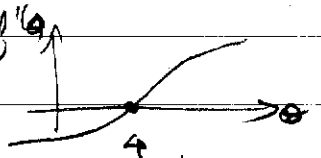


- Classifier $p(y_i, x)$ (page 9)
- When is ML consistent/analytic? & Bayes? E-family.
- E-family: bernoulli, multinomial, Gaussian, Poisson, dirichlet, ... the $\mathbb{R}^n$ >1
  - bernoulli: binary
  - multinomial: discrete
  - Gaussian: $\mathbb{R}^n$
  - Poisson: $\mathbb{Z}^+$ (# of people in a room)

Optimization:



but 
min if function is a bowl

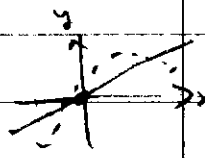


$\frac{\partial L}{\partial \theta}$

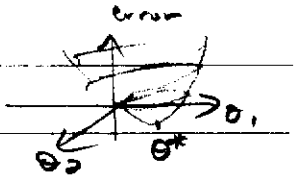
$\nabla_\theta L$ direction of fastest increase

max or min when flat.

$\therefore \frac{\partial L}{\partial \theta} := 0$ & solve.



least squares: $\min_\theta \frac{1}{2} \sum_i (y_i - \theta^T \vec{x_i})^2$



$\frac{\partial}{\partial \theta}$ ↳ $\frac{1}{2}(\vec{y} - X\theta)^T(\vec{y} - X\theta)$

$X^T(\vec{y} - X\theta) := 0$

$X^T X\theta := X^T y$

$\theta = (X^T X)^{-1} X^T y$

weighted least squares: $\min_\theta \frac{1}{2} \sum_i w_i (y_i - \theta^T \vec{x_i})^2$

$\equiv \frac{1}{2}(\vec{y} - X\theta)^T W (\vec{y} - X\theta)$

$\frac{\partial}{\partial \theta}$ $X^T W(\vec{y} - X\theta)$

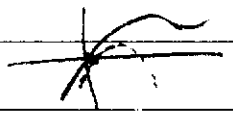
$WX^T X\theta = X^T W y$ ← hessian, ← gradient

$\theta = (WX^T X)^{-1} X^T W y$

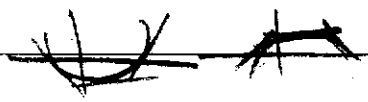
easy to solve

not easy to solve

gradient desc: $\theta^{t+1} = \theta^t + \delta \nabla_\theta L$ ← function to maximize

newton-raphson (taylor-series):  $\theta^{t+1} = \theta^t + H^{-1} \nabla_\theta L \Big|_{\theta^t}$

convex/concave: $\frac{\partial}{\partial \theta}\left(\frac{\partial}{\partial \theta} J(\theta)\right) = H$ if +ve semi-def then convex, else concave

= epigraph 

– Jensen: $\log(\sum_i p_i x_i) \geq \sum_i p_i \log x_i$

– Single max or min.

- General form: $p(x|\theta) = h(x) \exp\left(\theta^T T(x) - A(\theta)\right)$ ← natural params...

$$p(x|\theta) = \exp\left(H(x) + \theta^T x - A(\theta)\right)$$

$H(x) = \log h(x)$; $A(\theta)$ ← convex

$\int p(x|\theta)\, dx = 1$ — Cumulant gen. Func.

$$A(\theta) = \log \int \exp\left(H(x) + \theta^T x\right) dx$$

Laplace transform

$\frac{\partial}{\partial \theta} A(\theta) =$ mean $= E[T(x)] := \mu$

$\frac{\partial^2}{\partial \theta^2} A(\theta) =$ the def., cov. $= E[T(x)\,T(x)^T] - (E[T(x)])^2$

gradient $\frac{\partial}{\partial \theta} A(\theta)$ is unique (convexity)

to a θ

$E[T(x)] = \mu \longleftrightarrow \theta$

- Sufficient Stats. $\sum T(x)$ or $E[T(x)]$ is all you need

e-family has this, get $\sum x$, $\sum x^2$, .. $\sum x^n$, stop.

- ML $\prod_{i=1}^{N} \exp(H(x_i) + \theta^T x_i - A(\theta))$: $\frac{\partial}{\partial \theta} A(\theta) = \frac{1}{N} \sum x_i$

Examples Bernoulli: $p(x|\pi) = \pi^x (1-\pi)^{1-x}$ coin

$$\exp\left(\log\left(\frac{\pi}{1-\pi}\right) x + \log(1-\pi)\right)$$

$\theta = \frac{\pi}{1-\pi}$ ⟶ or $\pi = \frac{1}{1+e^{-\theta}}$

$T(x) = x$

$A(\theta) = -\log(1-\pi) = \log(1+e^{\theta})$

$h(x) = 1 \rightarrow H(x) = 0$

example multinomial $p(x|\phi) = \prod_{i=1}^{k} \phi_i^{x_i}$ $\quad x_i \in [0,1]$, $\sum_i x_i = 1$

$p(x|\theta) = \exp\left(\sum_{i=1}^{k} x_i \ln \phi_i\right)$ $\quad \sum_i \phi_i = 1$

$= \exp\left(\sum_{i=1}^{m-1} x_i \ln \phi_i + \left(1 - \sum_{i=1}^{m-1} x_i\right) \ln\left(1 - \sum_{i=1}^{m-1} \pi_i\right)\right)$

$= \exp\left(\sum_{i=1}^{m-1} \ln\left(\frac{\phi_i}{1-\sum_{i=1}^{m-1}\phi_i}\right) x_i + \ln\left(1 - \sum_{i=1}^{m-1} \pi_i\right)\right)$

$\theta_i = \ln\left(\frac{\phi_i}{\phi_m}\right)$ $\quad \phi_i = \frac{e^{\theta_i}}{\sum_{j=1}^{m} e^{\theta_j}}$ $\quad \phi_m = 1 - \sum_{i=1}^{m-1} \phi_i$

$A(\theta) = \ln\left(1 + \sum_{i=1}^{m-1} e^{\theta_i}\right)$

$H(\theta) = 0$

example dirichlet $p(x|\phi) = \frac{\Gamma(\alpha_1 + \dots + \alpha_k)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_k)} \prod_{i=1}^{k} \phi_i^{\alpha_i - 1}$

$\alpha_k > 1$

example Gaussian $\frac{1}{\sqrt{2\pi}\sigma} e^{\frac{1}{2\sigma^2}(x-\mu)^2}$

$= \frac{1}{\sqrt{2\pi}} \exp\left(\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{1}{2\sigma^2}\mu^2 - \ln\sigma\right)$

$\theta = \begin{bmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{bmatrix}$ ← this bottom param is neg.

$T(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix}$

multi-dim. ← $\begin{bmatrix} \Sigma^{-1}\mu \\ -\frac{1}{2}\Sigma^{-1} \end{bmatrix}$

$A(\theta) = \frac{\mu^2}{2\sigma^2} + \ln\sigma = \frac{-\theta_1^2}{4\theta_2} - \frac{1}{2}\ln(-2\theta_2)$

$H(x) = -\frac{1}{2}\log(2\pi)$

example exponential $p(x) = \lambda e^{-\lambda x}$ $\quad x \geq 0$ ← +ve scalars

$p(x) = \exp(-\lambda x + \log\lambda)$

$H(\theta) = 0$ $\quad \theta := -\lambda$

$\forall\ \theta < 0$ $\quad A(\theta) = -\log(-\theta)$

example Poisson $p(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$

$= \frac{1}{x!}\exp(x\log\lambda - \lambda)$

$\theta = \log\lambda$

$T(x) = x$

$A(\theta) = \lambda = e^{\theta}$

$H(x) = -\log x!$

Table

Gauss Mult. Bern Gaussian	Gauss Dir. Beta Normal Inv. Wish.

Common mean ↔ Common mean

more general form: $\exp\left(\nu H(x) + \phi(\theta)^T T(x) - A(\theta)\right)$ — unique pairs

conjugate: $\exp\left(J(\eta,\nu) + \phi(\theta)^T T(x) - \eta A(\theta)\right)$

(legendre transform) — virtual data points of times.

MAP

Bayes $\int P(x|\theta) \prod p(x_i|\theta)\, p(\theta)\, d\theta$ is solvable

$P(x|\theta)$: e-family; $p(\theta)$: conjugate

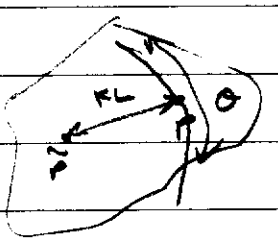
KL-divergence: $\equiv$ ML

$$\hat{p}(x) = \frac{1}{N} \sum_{n=1}^{N} \delta(x, x_n)$$

$$\min D(\hat{p} \| p) = \sum_x \hat{p}(x) \log\left(\frac{\hat{p}(x)}{p(x|\theta)}\right)$$

$$= \sum_x \hat{p}(x) \log \hat{p}(x) - \frac{1}{N} \ell(\theta|D)$$

$\sum_x \hat{p}(x) \log \hat{p}(x)$: negative entropy



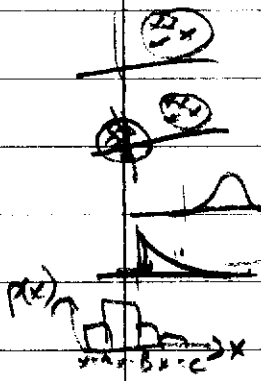
Assignment on Web Page Tomorrow

Read: Chapters 9 & 10

- Assignment #1 on the web page, due Tuesday 19th at 5pm by email or my office. Questions? office hours / email

- Review: optimization: grad descent, or $\frac{\partial}{\partial \theta} = 0$

e-family: ML maximization step $\frac{\partial}{\partial \theta} = 0$ guaranteed unique nice form $\prod_{n=1}^{N} p(x_n|\theta)$ by hits exp



e-family: each distrib has a conjugate

$\text{conj}(\text{conj}(e)) = e$ → $p(X|\theta)$ → $p(\theta|X)$

Gaussian M - Gaussian M

Gaussian τ - Inverse Wishart

Multinomial - Dirichlet

easy Bayes when prior is conjugate

Show examples - multinomial & Gaussian

Then Break e-family → EM → 1) K-means

today Ch. 9(⅓) & ch. 10

next ch. 2 (2nd ½) & ch. 3

2) Conquer & Divide

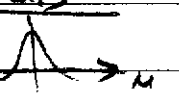
3) Bound Max

4) Filling in nodes

5) Coordinate Ascent

6) Information Geometry

- Bayesian Gaussian



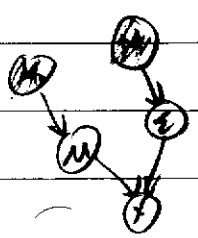
$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n$ τ given

see page 8-8bis

changes τ, makes it bigger

changes μ, pulls it towards prior

- Full integral over covariance params too



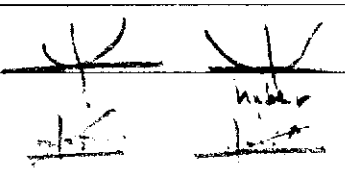
$p(\Sigma) = IW(\Sigma)$

get student "t"

heavy tail

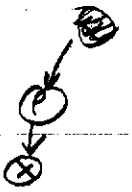
slower than e^{-x^2}

robust stats



as $N \to \infty$ Bayes → ML

Bayesian Multinomial



$p(x|\rho) = \prod_{k=1}^{K} \rho_k^{\delta(x=k)}$ ← indicator function

$p(\rho) = \text{Dirichlet} = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \rho_k^{\alpha_k - 1}$ ← normalized ← hyper params

$\sum_k \rho_k = 1 \quad \rho_k \geq 0$

$\Gamma(n+1) = n!$

$p(x|X) = \frac{1}{p(X)} \int p(x|\rho) \prod_{n=1}^{N} p(x_n|\rho)\, p(\rho)\, d\rho \qquad // \; p(x, X)$

$= \frac{1}{p(X)} \int \prod_{k=1}^{K} \rho_k^{\delta(x=k)} \prod_{n=1}^{N} \prod_{k=1}^{K} \rho_k^{\delta(x_n=k)} \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{K} \rho_k^{\alpha_k - 1} d\rho$

$= \frac{1}{p(X)} \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \int \prod_k \rho_k^{\alpha_k - 1 + \sum_n \delta(x_n=k) + \delta(x=k)} d\rho$

$= \frac{1}{p(X)} \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \frac{\prod_k \Gamma(\alpha_k + \sum_i \delta(x_i=k) + \delta(x=k))}{\Gamma(\sum_k \alpha_k + \sum_i \delta(x_i=k) + \delta(x=k))}$

easier: $\sum_i \delta(x_i=k) = N_k$, $\sum_k N_k = N$

$p(X) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \frac{\prod_k \Gamma(\alpha_k + N_k)}{\Gamma(N + \sum_k \alpha_k)}$

$p(x \mid X) = \frac{\prod_k \Gamma(\alpha_k + N_k + \delta(x=k))}{\Gamma(\sum_k \alpha_k + N + 1)} \frac{\Gamma(N + \sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k + N_k)}$

$p(x=k|X) = \frac{N_k + \alpha_k}{N + \sum_k \alpha_k}$ ← get ... ↗ plug in $x=k$ instead of $\frac{N_k}{N}$