CS 281 - Lecture 1

- What is machine learning? AI meets data/stats

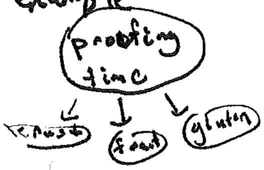
Expert knowledge, sophisticated reasoning and search ↘

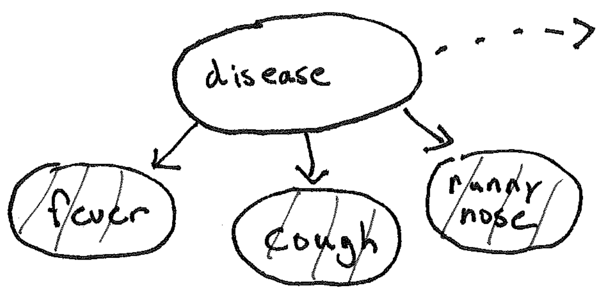
learn, train data, simpler models ↙

ML combines these

- We're going to focus on one particular area (take 181 for breadth): generative models of the world

2nd example





could be cold, flu, allergies

sup: $p(y \mid x, \theta)$
unsup: $p(x \mid \theta)$
param: $p(\theta \mid x, y)$

① disease generates symptoms (but then we infer diseases from symptoms)

② this picture is a "graphical model" shows generative dependencies

③ we could also learn the params from data (e.g. what the diseases are) — supervised vs. unsupervised

⓪ Write Bayes rule:

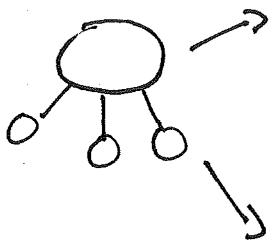
$$p(\text{disease} \mid \text{data}) \propto p(\text{fever} \mid \text{disease})\, p(\text{cough} \mid \text{disease}) \cdot \ldots \cdot p(\text{disease})$$

$$\propto \underbrace{\prod p(\text{symptom} \mid \text{disease})}_{\text{likelihood}} \; \underbrace{p(\text{disease})}_{\text{prior}}$$

likelihood
p(data | model)
"prob of data"

- in some ways like rule-based systems, need models
- note Bayes rule can be used with some symtoms missing (missing data). By modeling everything we can answer any question (drawbacks & benefits) "reasoning under uncertainty

4



Models: Simple discrete models, Gaussian models, Markov random fields, GLM, Factor analysis, HMMs, Latent Dirichlet Allocation, Deep models

Inference & Learning: Belief Propagation, Junction Tree, Variational, MCMC, LP Relaxations

We'll alternate between models and methods.

☐ Format of class / course

- Class: first 10 min. reading check: be on time, have a web-enabled device
- Lecture + turn to neighbor
- Course: 5 psets, first today! midterm
  Math & code reviews → section: stanchef training
- Final Project! (UAI/ICML/KDD) Turn to your partner and talk about research / why you're in this class
- piazza! sign-up.

- Remind Class
  - HW1 release today
  - HW0 due friday
  - try out quiz
  - Help with piazza

- Discrete Models:

Take on countable values $\{0,1\}$, {cold, flu, asthma}

Today: some of the simplest discrete models, how to manipulate, "tactics" Murphy 3.3

Running Example: $p(\text{heads}) = \theta$

- Prior #1: Three manufacturing processes: $\theta = .4, .5, .6$ w.p. $.1\ .8\ .1$ (empirical dist.)

→ Expert knowledge told us this fact. Mixture models coming later.

for now: $p(\theta) = .1\,\delta(\theta = .4) + .8\,\delta(\theta = .5) + .1\,\delta(\theta = .6)$

Likelihood: $\text{Bin}(N_1 | N, \theta) = \underbrace{\binom{N}{N_1}}_{\substack{\#\text{ of ways} \\ \text{to get } N_1 \text{ heads} \\ \text{in } N;\text{ constant} \\ \text{w.r.t. } \theta}} \underbrace{\theta^{N_1}(1-\theta)^{N-N_1}}_{\theta^{x}(1-\theta)^{x} \text{ per coin}}$

Suppose we observe $N_0, N_1, \ldots$ how do we perform inference? $\Rightarrow p(\theta | x)$

(N_0 = tails)

Several Options:

☐ Maximum Likelihood [MLE]

$$\theta_{MLE} = \arg\max_{\theta} p(N_0, N_1 | \theta) = \arg\max_{\theta} \log p(N_0, N_1 | \theta)$$

$$= \underbrace{\log \binom{N_1+N_0}{N}}_{\text{doesn't depend on } \theta \ldots} + \underbrace{N_1 \log \theta + N_0 \log(1-\theta)}_{\text{take derivatives wrt } \theta}$$

doesn't depend on θ... in fact we can ignore, pretend one sequence of heads

take derivatives wrt θ

$$\frac{d()}{d\theta} = \frac{N_1}{\theta} + \frac{N_0}{1-\theta}(-1) = 0$$

$$\theta = \frac{N_1}{N_0+N_1}$$

No surprises!

(?) What if you needed to predict whether a coin was going to be heads? would you guess w.p. θ? (No partial credit) No!

- Proportion right if you always guess H: θ
- " " " guess w.p. θ: $\theta^2 + (1-\theta)^2 \Rightarrow 1 - 2\theta + 2\theta^2$

If we plug-in .6 : .6 vs. .52

⇒ Predicting ≠ Decisions!

☐ MAP / full posterior [discrete_coins.m] $p(\theta|x) \propto p(x|\theta)\,p(\theta)$

$$p(\theta=.4 | N_0, N_1) \propto \underbrace{\binom{N_0+N_1}{N_1}}_{\text{can ignore}} \underbrace{(.4)^{N_1}(1-.4)^{N_0}}_{\text{unnormalized likelihood}} \underbrace{(.1)}_{\text{prior}}$$

$p(\theta=.45 | N_0, N_1) = 0$ [Big effect]

$p(\theta=.5 |\ )$ $p(\theta=.6|()) \rightarrow$ same as above normalize after

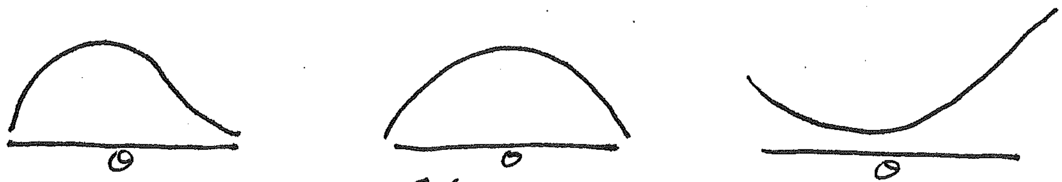
full posterior

MAP $\arg\max_{\theta} p(\theta|N)$ ⇓ not MLE!

(?) when will MAP = MLE

□ Prior #2

Beta distribution:



note pdf non-finite if $\alpha < 1$

$$p(\theta \mid \alpha_0, \alpha_1) = \frac{\Gamma(\alpha_1 + \alpha_0)}{\Gamma(\alpha_0)\Gamma(\alpha_1)} \underbrace{\theta^{\alpha_1 - 1}(1-\theta)^{\alpha_0 - 1}}_{\text{here, } \alpha_0\ \alpha_1 \text{ are pseudo-counts}}$$

plot_beta

$$p(\theta \mid N_0, N_1) = \frac{p(N_0, N_1 \mid \theta)\, p(\theta)}{p(N_0, N_1)} \quad \} \text{ constant}$$

$$= \frac{\Gamma(\alpha_0 + \alpha_1)}{\Gamma(\alpha_0)\Gamma(\alpha_1)} \theta^{\alpha_1 - 1}(1-\theta)^{\alpha_0 - 1} \theta^{N_1}(1-\theta)^{N_0} \quad (\text{constant})$$

$$= \quad '' \quad \underbrace{\theta^{N_1 + \alpha_1 - 1}(1-\theta)^{N_0 + \alpha_0 - 1}}$$

only part with θ. need to integrate to 1.

- oh but it has form of beta, so:

$$= \frac{\Gamma(N_1 + N_0 + \alpha_0 + \alpha_1)}{\Gamma(N_1 + \alpha_1)\Gamma(N_0 + \alpha_0)} \theta^{N_1 + \alpha_1 - 1}(1-\theta)^{N_0 + \alpha_0 - 1}$$

? A ↻↺ distribution corresponds to a "sparse" prior that things are either 0 or 1. How does this change if we had:

→ 10 counts to place

→ 8 heads, 1 tails, where would the 10^{th} go

No more sparsity once we have evidence →

no "noise" model

☐ Predictive Distribution

$$p(x \mid N_0, N_1) = \int p(x \mid N_0, N_1)\, p(\theta \mid N_0, N_1)\, d\theta$$

$$= \int p(x \mid \theta)\, p(\theta \mid N_0, N_1)\, d\theta = \int \theta\, p(\theta \mid N_0, N_1)\, d\theta$$

$$= \mathbb{E}_{\theta \sim p(\theta \mid N_0 N_1)}[\theta] = \frac{\alpha_1 + N_1}{\alpha_0 + \alpha_1 + N_0 + N_1} \Rightarrow$$ mean of Beta dist.

note: doesn't care about $\circlearrowleft$ or $\leftrightsquigarrow$

☐ Marginal Likelihood

$p(\text{data} \mid \alpha)$ = prob of model

$$p(N_0, N_1) = \int_\theta p(x_1 \dots x_N \mid \theta)\, p(\theta)\, d\theta \Rightarrow$$

$$\int_\theta \frac{\Gamma(\alpha_1 + \alpha_0)}{\Gamma(\alpha_0)\Gamma(\alpha_1)}\, \theta^{\alpha_1 - 1 + N_1} (1-\theta)^{\alpha_0 - 1\, N_0}\, d\theta = \text{"} \frac{\Gamma(\alpha_1 + N_1)\Gamma(\alpha_0 + N_0)}{\Gamma(\alpha_1 + N_1 + \alpha_0 + N_0)}$$

☐ Extensions: What can we do with coins?

→ many coins, correlated θ? a model of binary data

→ many-sided coins? a model of categorical data

$$p(\theta \mid \alpha) = \frac{\Gamma(\sum \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \theta_k^{\alpha_k - 1}$$ "Dirichlet"

$$p(\text{data} \mid \theta) = \frac{(\sum x_k)!}{\prod_k x_k!} \prod_k \theta_k^{x_k}$$ "multinomial"

☐ Think of a language with distribution over words

the	.5
some	.35
squid	.05
bow-tie	.05
octopus	.05

discrete
2·5

Change the model:

$$p(x\mid\alpha) = \int \mathrm{Mult}(x\mid\theta)\, p(\theta\mid\alpha)\, d\theta$$

$$p(\theta\mid x,\alpha) = \frac{1}{Z} \prod \theta_k^{x_k+\alpha_k-1}$$

$$\cancel{\frac{(\sum_k x_k+\alpha_k-1)!}{\prod(x_k+\alpha_k}} \quad \frac{\Gamma(\sum x_k+\alpha_k)}{\prod \Gamma(x_k+\alpha_k)}$$

? How to "train"? (inference over parameters)

Coming weeks

▷ inv orthonormal istrans
▷ eigen decomp $\Sigma = U \Lambda U$
▷ pos. def $\Lambda > 0$

Gaussians
3-1

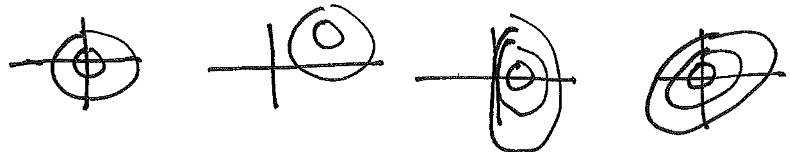
HW0 proves important properties

□ Project Ideas

□ Review Gaussians

$$\mathcal{N}(x \mid \mu, \Sigma) = \underbrace{(2\pi)^{-\frac{D}{2}} |\Sigma|^{-\frac{1}{2}}}_{Z} \exp\{-\tfrac{1}{2} \underbrace{(x-\mu)^T \Sigma^{-1} (x-\mu)}_{\text{important part quadratic form}}\}$$

? What do μ, Σ look like for:



normal

Decompose: $(x-\mu)^T \Sigma^{-1} (x-\mu) = (x-\mu)^T \left(\sum_d \frac{1}{\lambda_d} u_d u_d^T \right) (x-\mu)$

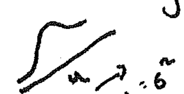
$$= \sum_d \frac{1}{\lambda_d} \Big[\underbrace{(x-\mu)^T u_d}_{y_d \text{ scalar}} \Big] \Big[u_d^T (x-\mu) \Big]$$

$\rightarrow$ gives ellipse for each direction d and elongation λ

? suppose $u_1 = [1 \ 0]$ $\lambda_1 = 3$
$u_2 = [0 \ 1]$ $\lambda_2 = 0.1$

what do we get? u_1 picks out 1st dim

? Tiled gaussians



□ Manipulating: stretch, rotate, shift [Change of variables]

let $y = Ax + b$ $\quad x \sim \mathcal{N}(0, I)$

recall: $y = f(x)$ $\quad p(x)\,dx \rightarrow \overset{p(y)}{p}\left(f^{-1}(y)\right) \left| \frac{d}{dy} f^{-1}(y) \right| dy$

$f^{-1}(y) = x$ $\quad p(y) = p(x) \left| \frac{dx}{dy} \right|$

here: $x = A^{-1}(y - b)$

$$\left| \frac{d}{dy} f^{-1}(y) \right| = A^{-1}$$

So we get

$$p(y) = (2\pi)^{-\frac{D}{2}} |A|^{-1} \exp\{-\tfrac{1}{2} x^T x\}$$ initial

$$= \quad " \quad " \quad " \{-\tfrac{1}{2} (A^{-1}(y-b))^T (A^{-1}(y-b))\}$$

$$\{-\tfrac{1}{2} (y-b)^T A^{-1^T} A^{-1} (y-b)\}$$

$$\mu_y = b \qquad \Sigma_y = A^T A$$

But wait, easier for Gaussians.

$$\mathbb{E}[y] = \mathbb{E}(Ax+b) = A\,\mathbb{E}(x) + b = b$$

$$\mathrm{Cov}[y] = \cancel{\mathbb{E}[(y-b)^T[y-b]]} = A^T A \quad \text{(exercise)}$$

true in general, but completely define gaussians (first 2 moments)

☐ Now write $\Sigma = U^T \Lambda U$ and $\mu = b$

$$y = \underset{\text{rotate}}{U}\ \underset{\text{scale}}{\Lambda^{\frac{1}{2}}} x + \underset{\text{shift}}{b}$$

plot
bigauss 2

☐ Detour: High-dim Gaussians

$x \sim N(0, \frac{1}{D} I)$

? What is expected length $\|x\|^2$?

$$\mathbb{E}[\|x\|^2] = \mathbb{E}[\textstyle\sum_d x_d^2] = D\sigma^2 = D\tfrac{1}{D} = 1$$

? What is the variance? $\mathbb{E}[x^4] = 3\sigma^4$

$$\mathrm{var}[\|x\|^2] \Rightarrow \mathbb{E}[x^4] - \mathbb{E}[x^2]^2 = 3\sigma^4 - \sigma^4 = 2\sigma^4 = D\tfrac{2}{D^2} = \tfrac{2}{D}$$

? implications?

- Key Formulas for MVN

$$\mu = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

$$p(x_1, x_2) = \mathcal{N}(\overset{x_1 x_2}{\mu}, \Sigma)$$

Marginalization: $p(x_1) = \int_{x_2} p(x_1, x_2)$

$$p(x_1) = \int_{x_2} \frac{1}{Z} \exp\{(x_1-\mu_1)^T \Sigma_{11}^{-1} (x_1-\mu_2) + 2(x_1-\mu_1)^T \Sigma_{12}^{-1} (x_2-\mu_2) + (x_2-\mu_2)^T \Sigma_{22}^{-1} (x_2-\mu_2)\}$$

$$= p(x_1) \int_{x_2} p(x_2|x_1) = \mathcal{N}(x_1 | \mu_1, \Sigma_{11})$$

Conditionals : $p(x_1|x_2)$

$$p(x_1|x_2) = \cancel{\mathcal{N}(\mu_{x_1|x_2}, \Sigma_{x_1|x_2})} = \mathcal{N}(\mu_{1|2}, \Sigma_{1|2})$$

$$\mu_{1|2} = \mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (x_2 - \mu_2) \Leftarrow \text{residual [later]}$$

$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \qquad \text{(proof in Murphy)}$$

Information Form

Instead of Σ and μ, use Σ^{-1} and $\Sigma^{-1}\mu$

This format makes conditioning trival (Σ_{11}^{-1})

but marginals more complicated.

□ MLE

$$\underset{\mu, \Sigma}{\text{argmax}} \sum_{n=1}^{N} \left\{ -\frac{1}{2} \ln |\Sigma| - \frac{1}{2} (x_n - \mu)^T \Sigma^{-1} (x_n - \mu) \right\}$$

$$\frac{d}{d\mu} : \Sigma^{-1} \sum_{n=1}^{N} (x_n - \mu) = 0 \Rightarrow \mu = \sum_{n=1}^{N} x_n / N$$ (sample mean)

$$\frac{d}{d\Sigma}$$ [Leave as exercise]

$$\frac{\partial}{\partial A} \ln |A| = A^{-1}$$

$$\frac{\partial}{\partial A} tr\{BA\} = B^T$$

$$tr(ABC) = tr(CAB) = tr(BCA)$$

$$\Sigma = \frac{1}{N} \sum_n x_n x_n^T$$ (sample covariance)

□ Conjugate Priors [just mean]

□ Predictive

☐ Regression

inputs x, outputs y

for now $\mathbb{R}$ in $\mathbb{R}$

☐ Gaussian Noise Model

$$p(y|x,\theta) = \mathcal{N}(y|w^T x, \sigma^2)$$

Write down likelihood:

$$\begin{aligned} \ell(\theta) &= \log p(D|\theta) \\ &= \sum_n \log p(y_n|x_n,\theta) \\ &= \sum_n \log \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(y_n - w^T x_n)^2\right\} \\ &= \underbrace{-\frac{1}{2\sigma^2}}_{\text{scaling}} \underbrace{\sum_n (y_n - w^T x_n)^2}_{\text{residual sum of squares}} - \underbrace{\frac{N}{2}\log(2\pi\sigma^2)}_{\text{constant}} \end{aligned}$$

Note how gaussian $\Leftrightarrow$ squared loss

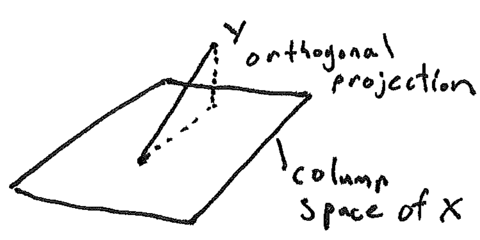
MLE: $(y - Xw)^T(y - Xw) = w^T X^T X w - 2 w^T X^T y$

$$\frac{d}{dw} \quad 2X^T X w - 2X^T y = 0$$

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

☐ Geometry of MLE

$y = \begin{bmatrix} | \\ y \\ | \end{bmatrix} \quad X = \begin{bmatrix} | & & | \\ x & \cdots & x \\ | & & | \end{bmatrix}$



□ Being Bayesian

$$Y = W^T X + \varepsilon$$

observed random fixed random

$p(w)$? → In HW1, consider $\mathcal{N}(w|0,\tau^2)$

$$p(y|X, w, \mu, \sigma^2) = \mathcal{N}(y|\mu + Xw, \sigma^2 I)$$

assume 0 centered   assume known

$$\propto \exp\{-\tfrac{1}{2}\tfrac{1}{\sigma^2}\|y - Xw\|_2^2\}$$

? why can we ignore Z?
↳ only care about w.

Put a prior: $p(w|w_0, V_0) \propto \exp\{-\frac{1}{2}(w-w_0)^T V_0^{-1}(w-w_0)\}$

$$p(w|x, y, \mu, \sigma^2) = \underbrace{}_{\mathcal{N}} \quad \underbrace{}_{\mathcal{N}}$$

$$= \mathcal{N}(w|w_N, V_N)$$

After algebra:

$$w_N = V_N V_0^{-1} w_0 + \frac{1}{\sigma^2} V_N X^T y = V_N (V_0^{-1} w_0 + \frac{X^T}{\sigma^2} y)$$

original mean

$$V_N = (V_0^{-1} + \frac{X^T X}{\sigma^2})^{-1}$$

original information

Note that $V_0 \to \infty$
becomes MLE $(X^T X)^{-1} X^T y$

□ Posterior Predictive

$$\int \mathcal{N}(y|x^T w, \sigma^2)\, \mathcal{N}(w|w_N, V_N)\, dw = \mathcal{N}(w_N^T x, \sigma^2 + x^T V_N x)$$

noise   variance on x

Variance depends on x!

$y = x^T w + \epsilon$ sum of gaussians

MVN MVN

fixed variance

Versus $p(y|x, w_{MAP}, \sigma^2)$

Show Demo
Samples

Before: Regression $x \to y$ $\mathbb{R}$

Now $y \in \{0,1\}$ (Later multiclass)

□ Naive Bayes

$y \sim \text{Cat}(\pi)$ $x_j \sim \underline{\quad}(y)$. Generative model of x, y

$$p(x \mid y=c, \theta) = \underbrace{\prod_{j=1}^{D}}_{\substack{\text{naive conditional} \\ \text{independence} \\ \text{assumption}}} p(x_j \mid y=c, \theta_{jc})$$

1) multivariate Bernoulli Naive Bayes

$x_j \sim \text{Ber}(\mu_{jc})$ if $y=c$

Features are binary-variables

2) Categorical NB

$x_j \sim \text{Cat}(\underline{\mu}_{jc})$ if $y=c$ [recall $p(\vec{x} \mid \mu) = \prod_d \mu^d$]

3) Gaussian NB

$x \sim \mathcal{N}(\mu_c, \Sigma^c_{diag})$ if $y=c$

MLE

$$p(x_i, y_i \mid \theta) = p(y_i \mid \pi) \prod_j p(x_{ij} \mid \theta_j)$$
$$= \prod_c \pi_c^{(y_i=c)} \prod_{j,c} p(x_{ij} \mid \theta_{jc})^{(y_i=c)}$$

$$\underset{\mu, w}{\text{argmax}} \; \log p(x_i, y_i \mid \theta) = \sum_{c=1}^{C} N_c \log \pi_c + \sum_j \sum_c \underbrace{\sum_{i: y_i=c} \log p(x_{ij} \mid \theta_{jc})}_{\text{MLE for distribution}}$$

$\hat{\pi}_c = \frac{N_c}{N}$ (class counts)

Bayesian NB

Use a factored Prior, assume parameter independence.

$$p(\vec{\theta}) = p(\pi) \prod_j \prod_c p(\theta_{jc})$$

(?) Which priors should we use here?

π parameters for Cat - Dirichlet

θ parameters for class-conditional, so - Beta or Dirichlet or MVN

Multivariate Bernoulli case

$$\pi \sim Dir(\alpha)$$
$$\theta_{jc} \sim Beta(\beta_0, \beta_1)$$

(often set to 1 assume 1 pseudo-count for features and class)

Recall that posterior updates have a simple form

$$p(\pi | D) = Dir(N_1 + \alpha_1, \ldots, N_C + \alpha_C)$$
$$p(\theta_{jc} | D) = Beta((N_c - N_{jc}) + \beta_0, N_{jc} + \beta_1)$$

Exercise: Compute Naive Bayes for gaussian with prior.

Posterior Predictive

$$p(y|x,D) \propto \underbrace{p(y|D)}_{\int \, d\pi} \prod \underbrace{p(x_j|y,D)}_{\int \, d\theta_{jc}}$$

Integrate over params

By earlier class we know this gives the mean of the posterior

$$\bar{\pi}_c = \frac{N_c + \alpha_c}{N + \sum_{c'} \alpha_{c'}} \qquad \bar{\theta}_{jc} = \frac{N_{jc} + \beta_1}{N_c + \beta_1 + \beta_0}$$

$$p(y=c|x,D) \propto \bar{\pi}_c \prod_{j=1} (\theta_{jc})^{(x_j=1)} (1-\theta)^{(x_j=0)}$$

Exponential Form

$$p(y=c|x,D) \propto \pi_c \prod_j \theta_{jc}^{(x_j=1)} \theta_{jc}^{(x_j=0)}$$

Take exp of log

$$\propto \exp\{\log \pi_c + \sum_j x_j \log \theta_{jc} + (1-x_j) \log^{1-} \theta_{jc}\}$$

$$= \exp\{w_c^T x + b_c\} \quad \text{where} \quad b = \log \pi_c + \sum_j \log\{1-\theta_{jc}\}$$

$$w_{jc} = \log \frac{\theta_{jc}}{1-\theta_{jc}}$$

This shows that NB is a transformation of linear functions of x. The formulation is known as the log-odds of the data.

Conversely if we have w we can recover θ

$$\theta = \text{sign}(w) = \sigma(w) = \frac{1}{1+e^{-w}} \quad \text{(exercise)}$$

This is the ~~logi~~ sigmoid function

Finally to normalize the probability distribution

$$p(y=c|x,D) = \frac{\exp(w_c^T x + b_c)}{\sum_{c'} \exp(w_{c'}^T x + b_{c'})} = \text{softmax}\left(\begin{bmatrix} \vdots \\ w_c^T x + b_c \\ \vdots \end{bmatrix}\right)_c$$

where $$\text{softmax}(z)_c = \frac{\exp(z_c)}{\sum_{c'} \exp(z_{c'})}$$

Last two classes linear regression and classification
Both linear at heart, but different outputs.

Today Exponential Families as a unifying concept.

- Central concept behind many core distributions: normal, bernoulli, categorical, gamma, et.
- Provides basis for conjugacy in Bayesian reasoning
- Central tool for graphical models and variational inf.
- We will use to rederive logistic regression etc.

– Warning: notation changes here a bit. Be careful w.r.t previous sections.

Exponential Family

$$p(x|\theta) = \frac{1}{Z(\theta)} h(x) \exp\{\theta^T \phi(x)\}$$
$$= h(x) \exp\{\theta^T \phi(x) - A(\theta)\}$$
$$A(\theta) = \log(Z(\theta)) = \log \int \_\_\_\_ \, dx$$

- θ · the natural parameters (function of "parameters" μ etc.)
- Z, A · the (log) partition function
- $\phi(x)$ · the sufficient statistics (informally features)
- $h(x)$ · scaling (not really important)

A representation is overcomplete if there is

Bernoulli:

$$\mathrm{Ber}(x|\mu) = (\mu)^x (1-\mu)^{(1-x)} = \exp x \log \mu + (1-x)\log(1-\mu)$$
$$= \exp\left\{x \log \frac{\mu}{1-\mu} + \log(1-\mu)\right\}$$

$$A = -\log(1-\mu) = -\log(1-\sigma(\theta)) = \theta + \log(1+e^{-\theta})$$ exercise

$$\phi(x) = x$$

$$\theta = \log\frac{\mu}{1-\mu} \qquad \mu = \underline{\sigma(\theta)} = \frac{1}{1+e^{-\theta}}$$ ψ mean ~~function~~ mapping and inverse

$$h = 1$$

Alternatiely, overcomplete representation:

$$\exp\left\{ \begin{bmatrix} x \\ 1-x \end{bmatrix} \begin{bmatrix} \log \mu \\ \log 1-\mu \end{bmatrix} \right\}$$

Univariate Gaussian

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right]$$

$$= \exp\left[-\frac{1}{2\sigma^2}x^2 + \frac{x\mu}{\sigma^2} - \frac{1}{2\sigma^2}\mu^2\right]$$

$$\phi(x) = [x^2 \quad x]$$

$$\theta = \left[-\frac{1}{2\sigma^2} \quad \frac{\mu}{\sigma^2}\right] \qquad \mu = \theta_1\sigma^2 = \frac{\theta_1}{-2\theta_2} \qquad \sigma^2 = -\frac{1}{2\theta_2}$$ Ψ mean function

$$Z = \sqrt{2\pi}\,\sigma \exp\frac{\mu^2}{2\sigma^2} \qquad A = \log\sqrt{2\pi} + \log\sigma + \frac{\mu^2}{2\sigma^2}$$
$$= \frac{1}{2}\log\sqrt{2\pi} \; \frac{1}{2}\log -2\theta_2 + \frac{\theta_1^2}{-4\theta_2}$$

Key properties

- Derivatives of the log-partition function $\underline{A}$ are the cumulents of the distribution, e.g. $\mathbb{E}[X]$, $\mathrm{var}[X]$, etc.

$$\frac{dA}{d\eta} = \frac{d}{d\eta}\left(\log \int \exp(\eta^T \phi) h(x)\, dx\right)$$

$$= \frac{\int \phi \exp(\eta^T\phi) h(x)\, dx}{\exp(A(\theta))} = \int \phi(x) p(x)\, dx = \mathbb{E}[\phi(x)]$$

Exercise: Show proof for variance and multivariate case

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

$$\frac{dA}{d\theta} = 1 - \frac{e^{-\theta}}{1+e^{-\theta}} = \frac{1}{1+e^{-\theta}} = \sigma(\theta) = \mu$$

Gaussian

$$A = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log(-2\theta_2) - \frac{\theta_1^2}{4\theta_2}$$

$$\frac{dA}{d\theta_1} = -\frac{\theta_1}{2\theta_2} = \frac{\mu/\sigma^2}{1/\sigma^2} = \underline{\mu}$$

$\boxed{\text{Fun}}$

MLE

$$\ell(\theta) = \log p(D|\theta) = \theta^T \underbrace{\phi(D)}_{\text{features of data}} - \underbrace{N}_{\text{data points}} A(\theta)$$

$$\frac{d\,\ell(\theta)}{d\theta} = \phi(D) - N\frac{d}{d\theta}A(\theta) = \phi(D) - N\,\mathbb{E}[\phi(D)]$$

Can show concave, so sufficient that

$$\frac{\phi(D)}{N} = \mathbb{E}[\phi(D)]$$

Moment matching.

Recall for Bernoulli: $\frac{1}{N}\sum_{i=1} I(x_i=1)$

Gaussian: sample mean, sample variance

Generalized Linear Models

Exponential families make it easy to generalize linear regression and classification.

$$p(y|x,w) = h(y) \exp\left(\phi(y)^T \underbrace{\theta(x,w)}_{\text{conditional}} - A(\theta)\right)$$

↳ break from murphy

univariate

$$\theta = \underbrace{\psi}_{\text{transform}} (g^{-1} (\underbrace{w^T x}_{\text{linear}}))$$

f - response function

Linear regression

Select exponential family as gaussian

Set ψ and g^{-1} to identity

Estimate $\underline{w}$ (closed form)

New: Logistic Regression

Select exp. family as Bernoulli:

set ψ to ~~identity~~ logit and $g^{-1}(w^T x) = \text{sigmoid}(w^T x)$

$\mu = \sigma(w^T x)$

$\theta = \log \frac{\mu}{1-\mu}$

Basic idea

use linear model to estimate μ, ~~then use Ψ~~ with squashing g^{-1} to match range. then use Ψ to map to natural parameters.

Exponential Families
6-5

Fitting Models

Graphical Models

- Core tool for rest of semester
- Separate out

$$\begin{bmatrix}\text{Conditional} \\ \text{Independence} \\ \text{Structure}\end{bmatrix} \begin{bmatrix}\text{form "parameterization"} \\ \text{of the model} \\ \text{itself}\end{bmatrix}$$

- Will provide key "modularity" for doing inference

☐ High-Level: When does a joint distribution simplify?

- Always use chain-rule $p(A,B,C) = p(A|B,C)\,p(B|C)\,p(C)$

But does it factor more? e.g.

$p(A|B)\,p(B|C)\,p(C)$

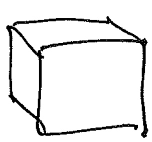
$(A) \leftarrow (B) \leftarrow (C)$ generative "story"

Formally, directed GM or Bayes net,

- Graph $G = (V, E)$ with $(s,t) \in E$ $s \neq t$ — $pa(x)$ parents of x
- Each node corresponds to a random variable.
- Each edge " to a conditioning decision
- Graph is topologically ordered because of chain rule
- Nodes that are shaded indicate observed RVs.

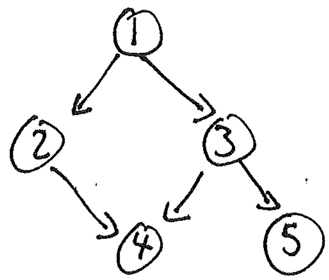
Discrete DGMs

- Each node associated with a sample space set
- Local conditional probabilities defined by a CPT
- CPT size of $p(x_i | x_1 \ldots x_{i-1}) = O(\prod |X|)$



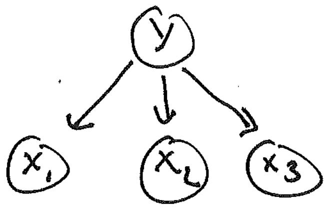
Graphical Models
7-2

Examples

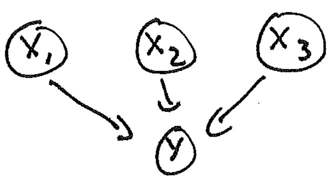


5 CPTs

(?) Write out DGM for naive bayes

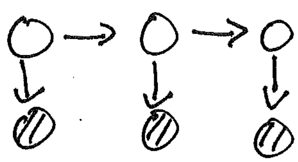


LR

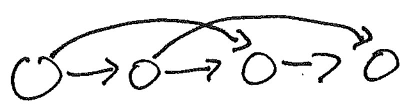


- much worse. simplicity in hidden parameterization

HMM



Auto regressive / Markov chain

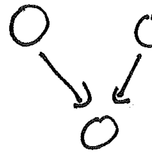


Factorial HMM

☐ Gaussian Directed Models

Special Case: much easier

$$p(x_i \mid pa(x_i)) = \mathcal{N}(x_i \mid \mu_i + w_i^T (pa(x_i)_j - \mu), \sigma_i^2)$$



mean is linear in parent variables

$$x_i = \mu_i + \sum_j w_{ij} (x_j - \underbrace{\mu_j}_{\text{constant for simplicity}}) + \sigma_i z_i \qquad \forall i,\ z_i \sim \mathcal{N}(0,1)$$

Can derive global mean $\underline{\mu}$ as $(\mu_1, \dots, \mu_D)$

Let $S = \mathrm{diag}(\underline{\sigma})$ local standard dev.

$$(\cancel{x_i} - \mu_i) = W(x - \mu) + Sz \qquad \text{[matrix-vector form]}$$

$$Sz = (I - W)^{-1}(x - \mu)$$

$$x - \mu = \overbrace{(I - W)^{-1}}^{U} S z$$

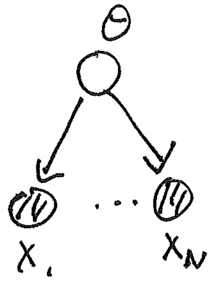
$$\Sigma = \mathrm{cov}[x - \mu] = \mathrm{cov}[USz] = US\,\underbrace{\mathrm{cov}[z]}_{1}\,SU^T = US^2U^T$$

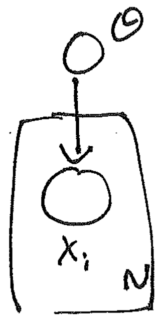
(?) Why is $(I - W)$ invertible?

$$GGM = \mathcal{N}(\mu, US^2U^T)$$

↳ invert weights
S is local variances

Properties of Bayes Nets: Plates

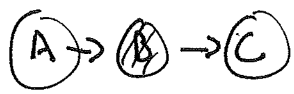




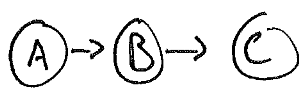
iid

D-separation and conditional independence

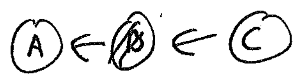
?



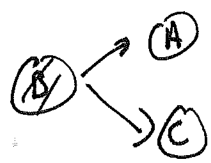
$A \perp C \mid B$ ✓



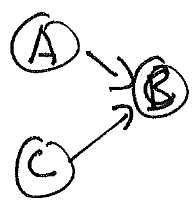
$A \perp C$ ✗



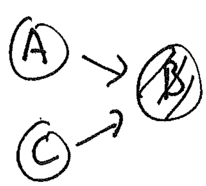
$A \perp C \mid B$ ✓



$A \perp C \mid B$ ✓



$A \perp C$ ✓



$A \perp C \mid B$ ✗ "Explaining Away"

Undirected Graphical Models
8-1

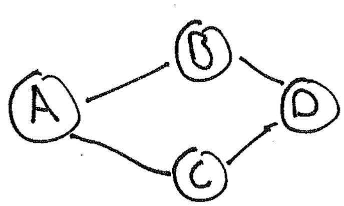
Last class: Directed graphical models. At

- Attempt to describe the conditioning relationships
- can directly use to find local conditional

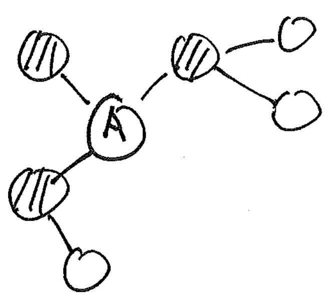
Today: Undirected Graphical models (Markov Random Fields)

- Simpler conditional independence rules
- Describes a different class of distributions
- (personal bias) Often more useful

High-Level: Independence properties



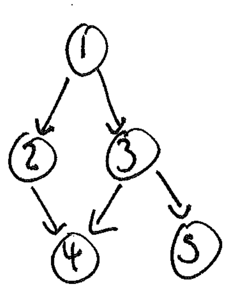
$A \perp D \mid S$ if ~~any~~ no path between A and D that does not cross throus S

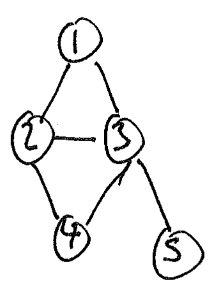


Fundamental property: A is conditionally independent from rest of graph conditioned on its markov blanket (neighbors in G)

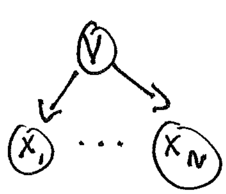
Conversion from directed

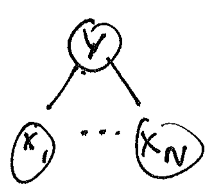
"marry parents"

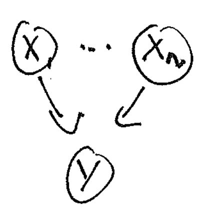


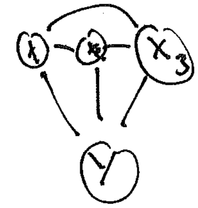


clear why difficult.

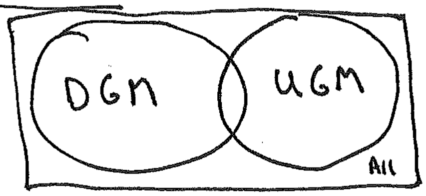


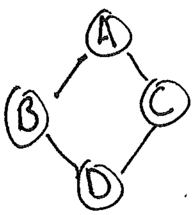




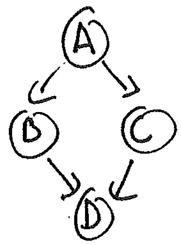


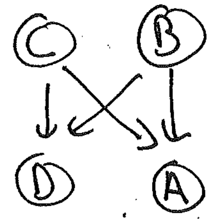
Corner Cases

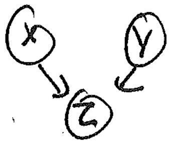




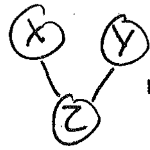
$A \perp D \mid B, C$
$B \perp C \mid A, D$





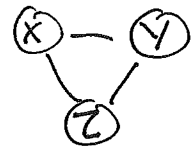


$X \not\perp Y \mid Z$
$X \perp Y$



$X \perp Y \mid Z$

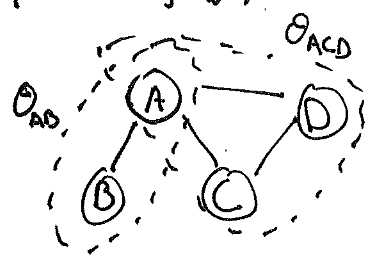
no



$X \perp Y$ X

MRF Parameterization

$$p(x_1, \dots, x_N) = h(x) \exp\{\theta^T \phi(x) - A(\theta)\}$$ (exponential family)



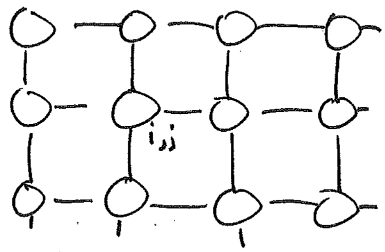
One log-potential function θ associated with each clique in the graph

$$\exp\{\theta_{AD}(x_A, x_D) + \theta_{ADC}(x_A, x_D, x_C) - A(\theta)\}$$

- Note: unlike DGM there are no easy local probabilities
- $\theta_{AD}(x_A, x_D)$ is a local energy, but is unnormalized, compare to CPT
- To compute $p(x_1, \dots x_n)$ need log partition function
- ~~For~~ In general computing A is NP-complete sum or integral over all structure.

Undirected
Graphical models
8-3

Canonical Example



$\exp\{\sum_{ij} \theta^{\uparrow}_{ij}(x_{ij}, x_{i-1,j}) + \theta^{\rightarrow}_{ij}(x_{ij}, x_{ij+1}) \cdots \}$

$-E$

partition $\log \sum_x e^{-E(x)}$: super-intractable!

What next?

$p(x_1 \ldots x_N)$ - likelihood of data

$p(x_i)$ - marginals

$\underset{x_{1N}}{\operatorname{argmax}}\, p(x_1 .. x_N)$ - arg max/MAP

Gaussian

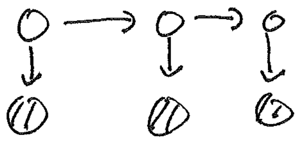
Given "information" form Σ^{-1} can read MRF off of

$\Sigma^{-1}_{ij} \neq 0$ implies $x_i \Leftrightarrow x_j$ edge

$\Sigma^{-1} = (I-W)(S^2)^{-1}(I-W)^T$

$\Sigma = US^2U^T$

Recall



HMM

Gaussian

$$x_t = x_{t-1} + \epsilon \qquad \epsilon \sim N(0, \sigma_t^2)$$

$$z_t = x_t + \epsilon' \qquad \epsilon \sim N(0, \sigma_n^2)$$

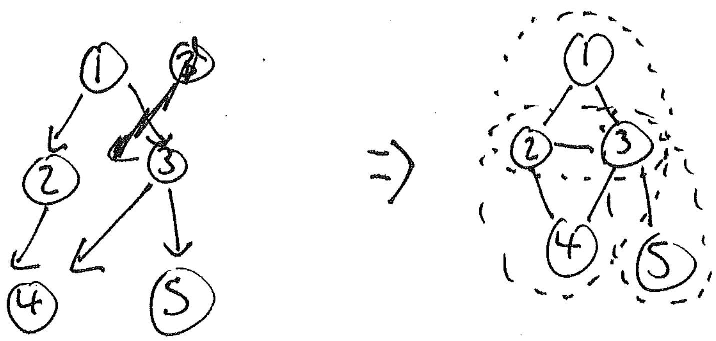
Joint: $\prod_t p(x_t | x_{t-1})\, p(z_t | x_t)$

Time series 6 M

$8\frac{1}{2} - 2$

This Class:

Exact marginal inference in undirected discrete GMs.



$$p(x_1)\,p(x_2|x_1)\,p(x_3|x_1)\,p(x_4|x_2,x_3)\,p(x_5|x_3) =$$
$$\exp\{\theta_{123}(x_1,x_2,x_3) + \theta_{234}(x_2,x_3,x_4) + \theta_{35} - A(\theta)\}$$

where $\theta_{123} = \log p(x_1) + \log p(x_2|x_1) + \log p(x_3|x_1)$
$\theta_{234} = \log p(x_4|x_2,x_3)$
$\theta_{35} = \log p(x_5|x_3)$ [may lose some CI information]

Hammersley-Clifford: A positive distribution satisfies CI properties of a graph iff it can be represented as

$$p(y|\theta) = \exp\{\sum_{C} \theta_C(y_C) - A(\theta)\} \quad \text{or} \quad \underbrace{\frac{1}{Z(\theta)}\prod_C \psi_C(y_C|\theta_C)}$$

Where C is the set of cliques in the graph. Murphy

Therfore for simplicity we will consider UGM to start with.

Ex: Conditional random field.

A CRF is a conditional ~~as~~ UGM

$$p(y_1 .. y_n \mid x) = \exp\{\sum_c \theta_c(y_c; x) - A(\theta)\}$$

They are heavily used for labeling style problems.

Linear chain CRF

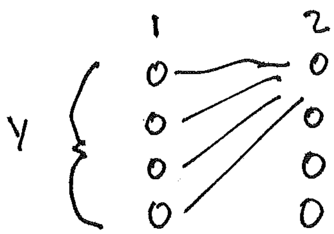
use ψ instead

$$p(y_1 \cdots y_N \mid x) = \exp\{\sum_{i=1}^{N-1} \theta_{i,i+1}(y_{i,i+1}; x) - A(\theta)\}$$

$$A(\theta) = \log \underbrace{\sum_{y'}}_{\text{exponential size set}} e\{\sum_{i=1}^{N-1} \theta_{i,i+1}(y'_{i,i+1})\} = \log \sum_{y'_{12}} \exp\{\theta_{12}(y_{12})\} \sum_{y_3} \exp\{\theta_{23}(y_{23})\} .$$

Dynamic Programming. $\cancel{\alpha_{IJ} = \sum_{y': y_I = J} \exp \sum_{i=1}^{I-1} \theta_{i,i+1}(y_i, y_{i+1})}$

$\cancel{\alpha_{IJ} = \alpha_{I-1}}$



$p(y_t \mid x)$ $\cancel{bel^-_t(y_t) \triangleq \exp\{\sum m^-_{t-1 \to t}(y_t) - A_t\}}$

$\cancel{m^-_{t-1 \to t}(y_t) = \sum_{y_{t-1}} \theta_{t-1 \to t}(x_{t+1}) * bel^-_{t-1}(y_{t-1})}$

$\cancel{m^+_{t \to t+1}(y_{t+1}) = \sum_{y_t} e \theta}$

$\cancel{m_{t+1 \to t}(y_t) = \sum_{y_{t+1}} \exp \theta_{t,t+1}(y_{t,t+1})\, m_{t+1,t-1}(y_t)}$

$$bel^-_t(y_t) \triangleq \frac{1}{Z_t} m^-_{t-1,t}(y_t)$$

$$m^-_{t-1,t}(y_t) = \sum_{y_{t-1}} \psi_{t-1}(y_{t-1}, y_t)\, bel^-_{t-1}(y_{t-1})$$

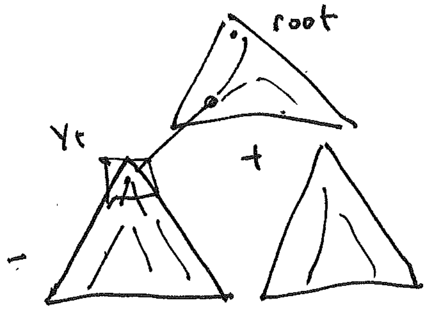
] forward

$$bel_t(y_t) \propto b^-_t(y_t)\, m^+_{t+1 \to t}(y_t) \;\; p(y_t \mid x)$$

$$m^+_{t+1,t}(y_t) = \sum_{y_{t+1}} \psi_t(y_t, y_{t+1})\, m^+_{t+2 \to t+1}(y_{t+1})$$

] backward

General Case: Tree



- multiple nodes below

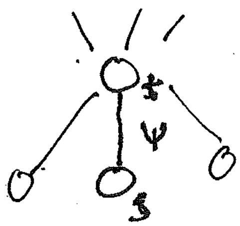
$$bel_t^-(y_t) = \frac{1}{Z} \underbrace{\prod_{c \in ch(t)}} m^-_{c\to t}(y_t)$$

$$m^-_{c\to t}(y_t) = \sum_{y_c} \psi_{ct}(y_c, y_t)\, bel_t^-(y_t)$$

$$bel_s(y_s) \propto bel_s^-(y_s) \prod_{t \in pa(s)} m^+_{t\to s}(y_s)$$ around the sides

$$m^+_{t\to s}(y_s) = \sum_{y_t} \psi_{s,t}(y_s, y_t) \frac{bel_t(y_t)}{m^-_{s\to t}(x_t)} \Big] \text{ saved}$$

$$= \sum_{y_t} \psi_{st}(y_s, y_t) \prod_{c \in ch(t) \setminus s} m^-_{c\to t}(y_t) \prod_{p \in pa(t)} m^+_{p\to t}(y_t)$$



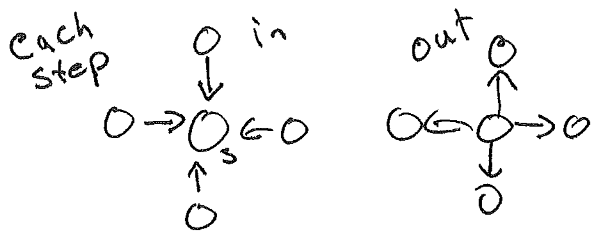
Algorithm

BP implementation

We assumed bottom-up and top down.

However can implement in parallel.

Assume $bel_s^0(x_s) = \text{unif}.$

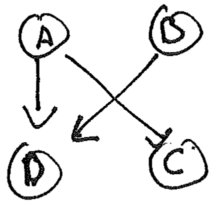


step 1: $bel_s(x_s) \propto \prod_{t \in nbr(s)} m_{t \to s}(x_s)$ (Compute beliefs)

step 2: $m_{s \to t}(x_t) = \sum_{x_s} \left(\psi_{st}(x_s, x_t) \prod_{u \in nbr \setminus t} m_{u \to s}(x_s) \right)$ (marginalize self-out)

Variable Elimination

Consider non-trees



$$p(D) = \sum_{ADC} p(A,B,C,D) = \sum_{ABC} p(D|A,B)\, p(C|A)\, p(A)\, p(B)$$

$$= \sum \psi(D,A,B)\, \psi(A,C)\, \psi(A)\, \psi(B)$$

$$= \sum_{AB} \psi(D,A,B)\, \psi(B) \underbrace{\sum_{C} \psi(A,C)\, \psi(A)}_{p(A)}$$

$$\sum_{A} \psi(A) \sum_{B,D} \psi(D,A,B)\, \psi(B)$$

$$\psi(A)\; p(A)\; p_1(A)$$

Computational complexity, exponential in largest factor

Parallel version of algorithm

So far all methods have been exact.

However this is only for a special case of models.

Most models of interest will be approximations, the focus of 2^{nd} part of class.

→ Before we do that, we need some more fundamentals

Weirdly, these will come from information theory.

Whole textbooks written on this connection (Cover and Thomas 2006, Mackay 2003)

Information Theory

$$\text{source} \rightarrow \text{encoder} \xrightarrow[\text{channel}]{\text{noisy}} \text{decoder} \rightarrow \text{target}$$

$p(x)$ $\quad$ $p(y|x)$ $\quad$ y $\quad$ message?

$$\underbrace{p(x|y)} \propto \underbrace{p(x)}_{\text{Source model}} \underbrace{p(y|x)}_{\text{Channel model}}$$

Fundamental connection: How many bits of source can be recovered.

Entropy

$$H(p) = \mathbb{H}(X) \triangleq - \sum_{k=1}^{K} p(X=k) \log_2 p(x=k) = \mathbb{E}(\log_2 p(x=k))$$

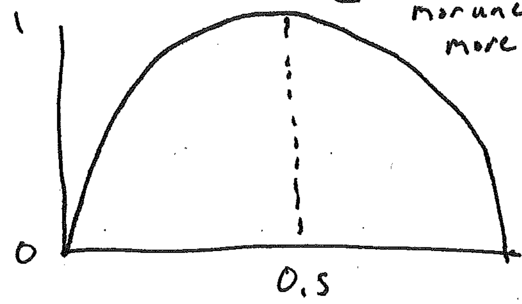
- measure of the uncertainty of the distribution
- unit of measure is "bits" (or nats if ln)
- Avg number of fewer bits required to encode the distribution

$\exp(H(X))$ = perplexity effective uncertainty of distribution

Shannon ~~Game~~: example guess next word

For binary case

$$H(X) = -\Big[\underbrace{\theta \log \frac{\theta}{1-\theta}}_{\text{log odds}} + \log_2 (1-\theta)\Big]$$



Cross Entropy

$$H(p,q) = \sum_k p(X=k) \log q(X=k) = \mathbb{E}_p\left[\log q(X=k)\right]$$

p - true distribution

q - our distribution

- number of avg. bits required when true dist is p, but we use q.

Example: language model

- Use q to estimate next word, but true dist is p.

$$\mathbb{E}_p\left[\log q(X=k)\right] \approx \sum_{v \in J} \log q(X=k) = H(p,q)$$

- in fact MLE of categorical classifier

$$\log p(y|x) = \sum_n \log p(y_n|x_n) = \sum_n \sum_{y'} \underbrace{\mathbb{I}(y'=y_n)}_{\text{Delta dist. over } y\text{'s}} \underbrace{\log p(y_n|x_n)}_{q \text{ dist as above}}$$

In deep learning, the most common discrete loss is "Cross-entropy" loss.

Information Theory
$9\frac{1}{2}-3$

KL divergence **relative entropy**

- The most common way to compare distances.

$$KL(p\|q) \triangleq \sum_k p_k \log \frac{p_k}{q_k} = \mathbb{E}_p\left[\log \frac{p_k}{q_k}\right] = \underbrace{-H(p)}_{\text{bits to encode } p} + \underbrace{H(p,q)}_{\text{bits to encode } p \text{ with } q.}$$

- extra number of bits needed with q.

$$\underset{q}{\operatorname{argmin}}\ KL(p\|q) = H(p,q) \Rightarrow \text{MLE estimate of } q \text{ with obs } p$$

$$\underset{p}{\operatorname{argmin}}\ KL(p\|q) = \underbrace{-H(p)} + H(p,q) \Rightarrow$$ max entropy p ~~that can be encoded w/~~

Theorem: $KL(p\|q) \geq 0$ (full support)

$$-KL(p\|q) = \mathbb{E}_p\left[\log \frac{q_k}{p_k}\right] \underset{\downarrow}{\leq} \log \mathbb{E}_p\left[\frac{q_k}{p_k}\right] = \log \sum q_k(x)$$

minus

$$= \log 1 = 0$$

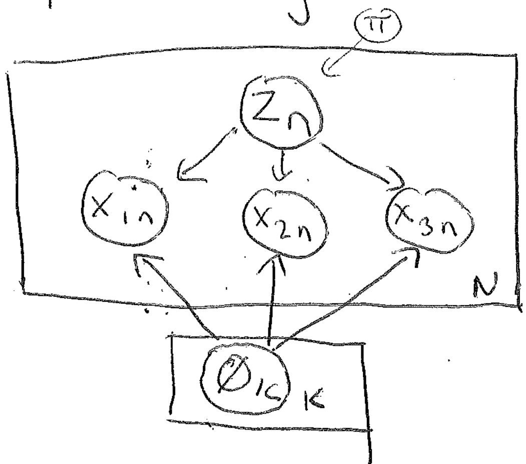
Jensen's Inequality: $f(\mathbb{E}[x]) \leq \mathbb{E}(f(x))$ if f is convex

we will use for $\underline{\log}$

Equality when $p = q$.

- ☐ Untile now: Supervised models
  input: $x \in \mathbb{R}^d$, output $y \in \mathbb{R}$, $\{0,1\}$, GLM

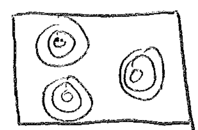
- ☐ Now unsupervised setting: Latent variables z that are unseen.



Z_n controls which ϕ_k generates the data!

Specific instantiations:

1) ϕ are means of gaussians (GMM)



2) ϕ are seperate multinomials (mixtures of multinomials)

- ☐ Why is this challenging? Graphical models are highly connected.

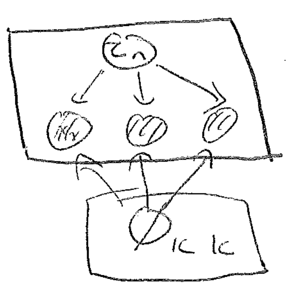
$$\{x_n\}, \pi, \{z_n\}, \{\phi_k\}$$

$$p(\{x_n\},\{z_n\}) = \prod_n \prod_k \left(\pi_k \, p(x_n \mid \phi_k)\right)^{z_{nk}}$$ Complete Data likelihood

$$p(\{x_n\}) = \sum_{\{z_n\}} \prod_n \prod_k \quad ''$$

swap

But $p(\{x_n\},\{z_n\})$ looks okay...



$Z_n \not\perp \phi_k \mid x_n$

$Z_n \not\perp z_i \mid x_n$!

- Let's write with logs.

$$\log p(x_n, z_n \mid \pi, \phi) =$$
$$\sum_n \sum_k z_{nk} \ln \pi_k + z_{nk} \ln p(\cancel{x}_n \mid \phi_k)$$

- Expectation - Maximization → Local coordinate ascent

Goal: Maximize <u>expected</u> complete log-likelihood

1) Suppose we have distribution over z_n.

$$q_{nk} = pr(z_n = k)$$

$$\mathbb{E}_q [\ln p(x_n, z_n \mid \pi, \phi)] = \sum_n \sum_k q_{nk} \ln \pi_k + q_{nk} \ln (x_n \mid \phi_k)$$

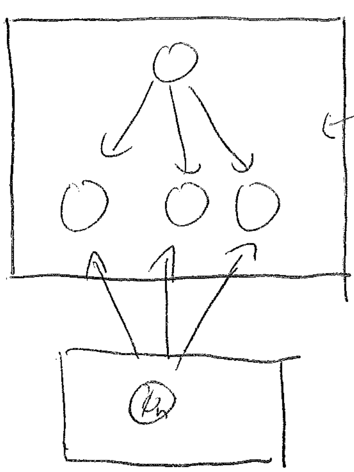
2) E-step: improve approx of q_{nk} given π, ϕ_k estimates

$$q_{nk} \leftarrow \frac{\pi_k \, p(x_n \mid \phi_k)}{\sum_{k'} \pi_{k'} \, p(x_n \mid \phi_{k'})} \qquad \text{Formula for } p(z_n)$$

3) M-step: improve $\{\phi\}, \pi$ assuming q_{nk} is data.

$$\pi_k \propto \sum_n q_{nk} \qquad \text{(categorical)}$$

π ϕ's are model specific.



E-step
$p(z) = q_{nk}$
Can be computed efficiently

Mixture Models
10-3

☐ Justification of EM

$$p(\{x_n\}) = \prod_n \sum_{z_n} p(x_n, z_n \mid \pi, \phi_k)$$

log

$$= \sum_n \log \sum_k p(x_n, z_n = k \mid \pi, \phi_k)$$

$$= \sum_n \log \sum_k q(z_n = k) \frac{p(x_n, z_n = k \mid \ )}{q(z_n = k)}$$

$$= \sum_n \log \mathbb{E}_{k \sim q}\left[\frac{p(x_n, z_n = k)}{q(z_n = k)} \right]$$

jensen
flip log exp

$$\geq \sum_n \mathbb{E}_{k \sim q} \log \text{ ''}$$

$$= \sum_n \underbrace{\mathbb{E}_{k \sim q} \log p(x_n, z_n \mid \ldots)}_{\text{Expected complete data likelihood}} + H[q(z_n = k)]$$

Always a lower bound
we maximize

Now put entropy back

Can show

$$= \sum_n \underbrace{KL[q(z) \,\|\, p(z \mid x, \pi, \phi)]}_{\text{alternating opt, here with marginalization}} + \text{const}$$

$$q = p(z \mid x \pi, \phi$$

- [ ] EM solves the problem by splitting into two parts
  - → local variables z_n
  - → global variables ϕ_k π

Common strategy

Last week: Exact inference rarely possible

Today: Beginning of unit on approximate inference

☑ Variational Inference: idea

if finding $p(z,\theta|D)$ is too hard find $q(z,\theta)$

that is close: $q^* = \underset{q \in Easy}{\mathrm{argmin}}\ d(q,p)$

- if q^* is easy, can compute marginals on q^*
- p can be any distribution.

We will use $d(q,p) = KL(q||p) = \oint q(z,\theta) \log \frac{q(z,\theta)}{p(z,\theta)}$

<u>Rel to EM</u> Recall with EM:

$$\log p(x) = \log \int_\theta p(x,\theta)\, d\theta = \log \int_\theta \underline{q(\theta)} \frac{p(x,\theta)}{\underline{q(\theta)}} d\theta = \log \mathbb{E}_{\theta \sim q} \frac{p(x,\theta)}{q(\theta)}$$

$$\geq \mathbb{E}_{\theta \sim q} \log \frac{p(x,\theta)}{q(\theta)} \quad [\text{Jensen's inequality}]$$

Now consider gap in likelihood

$$\log p(x) - \mathbb{E}_q \log \frac{p(x,\theta)}{q(\theta)} = \cancel{\int_\theta q(\theta) p(x) - \int_\theta}$$

$$\mathbb{E}_q \log(p(x)) - \log \frac{p(x,\theta)}{q(\theta)}$$

$$= \mathbb{E}_q\left[\log \frac{q(\theta)}{p(\theta|x)}\right] = KL(q,p)$$

Directly minimizing KL is equivalent to minimizing lowerbound

Coordinate ascent versus opt.

Solving

- This is an optimization problem any method is okay
- Today Mean-field

1) Select $q(z) = \prod_i q_i(z_i)$, utilize to approximate $p(z)$

2) Recall goal is to maximize lower-bound $KL(q, p)$, we do this by fitting each q_i individually.

$$q_i^*(z_i) \leftarrow \underset{q_i}{\operatorname{argmin}} \; KL(q \| p)$$

$$= \underset{q_i}{\operatorname{argmin}} \int_z \left(\prod_i q_i(z_i)\right) \log\left(\frac{\prod q_i(z_i)}{p(z)}\right)$$

$$= \underset{q_i}{\operatorname{argmin}} \; -H(q_i) - \int_z \left(\prod_i q_i(z_i)\right) \log(p(z)) + \cdots$$

(other entropy terms)

$$= \underset{q_i}{\operatorname{argmin}} \; -H(q_i) - \int_{z_i} q_i(z_i) \underbrace{\int_{z_j \neq z_i} \prod_{j \neq i} q_j(z_j) \log p(z)}_{\mathbb{E}_{j \neq i}[\log p(z)]}$$

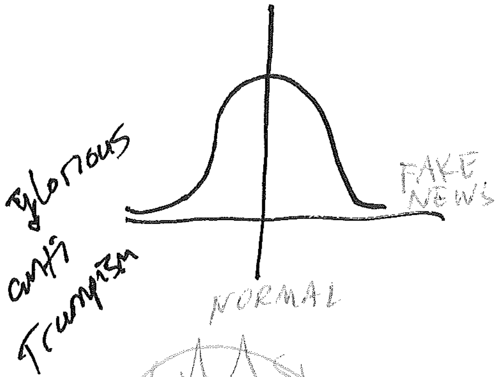
Call this $\log \tilde{p}(z_i)$

$$= \underset{q_i}{\operatorname{argmin}} \; KL(q_i \| \tilde{p})$$

$q_i \leftarrow \tilde{p}$; $\underline{q_i(z_i) \propto \exp\{\mathbb{E}_{j \neq i}[\log p(z)]\}}$

$KL(p\|q)$ version <u>EP</u> does same thing

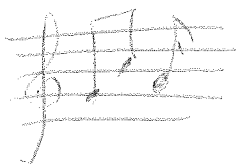
MOST IMPORTANT CLASS
12-2.5



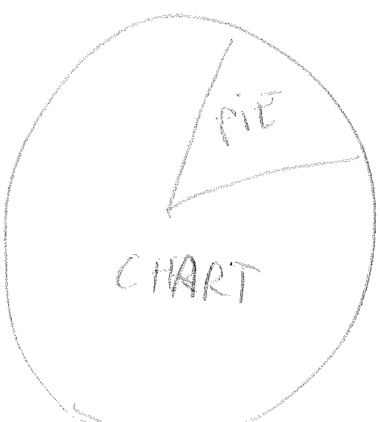
$0.32 \pm 0.46 \cdot SE$

$$\sqrt{\sin\left(\cos\left(x^3\right)^5\right)} = \int_1^{\rho} y^2$$

put the entropy back
now take it out again
put it in!
mix it around
do the hokey pokey
DESPACITO
hey baby ♡



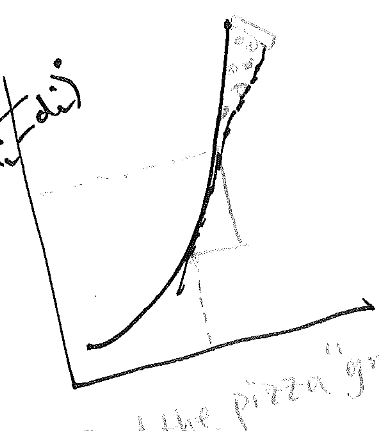
$Pr[\overset{\circ}{1}/x]$



α

π

$$\sqrt{Var(\hat{S}(t))} = \hat{S}(t)^2 \sum_{i:t_i \le t} \frac{d_i}{n_i(n_i - d_i)}$$



"Find the pizza" graph

$$x = \frac{-b \pm \sqrt{4ac^2}}{2a}$$

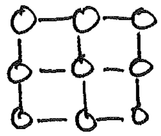
Variational
12-3

□ Example

Ising model

local

$$p(X) \propto \exp\{\theta_v^T x + x^T \theta_E x\} \qquad \mu_i = \mathbb{E}[x_i] = p(x_i=1)1 + p(x_i=0)(0)$$



updates $\log q_i(x_i) \hat{\propto} \mathbb{E}_{j\neq i}\left[\theta_i x_i + \sum_{j\in N(i)} \theta_{ij} x_i x_j\right] + \text{con}$

$$\ln(\mu_i) = \log q_i(x_i) \propto \theta x_i + \mathbb{E}\left[\sum \theta_{ij} x_i x_j\right] = \theta x_i + \sum \theta_{ij} \mu_j$$

$$\mu_i = \frac{1}{1 + \exp\{-(\theta_i + \sum \theta_{ij}\mu_j)\}} = \sigma(-(\theta_i + \sum \theta_{ij}\mu_j))$$

VLB $\log Z \geq H(q,p) + H$

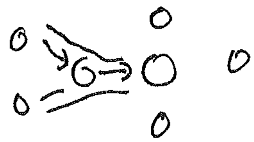
Bayes GMM

Loopy belief propagation

1) $m_{s\to t}(x_t) = \sum_{x_s}\left(\psi_{st}(x_s,x_t)\prod_{u\in nbr(s)\setminus t} m_{u\to s}(x_s)\right) = \sum \exp\{\theta_{st}(x_s,x_t) + \sum m_{s\to t}\}$

2) $bel_s(x_s) \propto \prod_{t\in nbr(s)} m_{t\to s}(x_s)$

$q_s =$



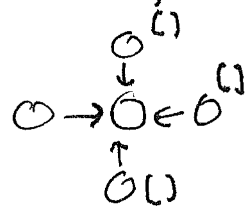
mean field

$\cancel{\log q_s(x_s) \propto \prod_{t\in nbr(s)} \sum_{x_t} \psi_{st}(x_s,x_t)\, q_t(x_t)}$

$\log q_s(x_s) \propto \mathbb{E}_{j\neq s}\left[\sum_{j\in nbr_s} \theta_{sj}(x_s,x_j)\right] + const.$

$\sum_{j\in nbr_s}\sum_{x_j} q(x_j)\,\theta(x_s,x_j)$

$q_s(x_s) \propto \prod_{j\in nbr_s} \exp\left(\sum_{x_j} q(x_j)\,\theta(x_s,x_j)\right)$



$q(x_s) \propto \exp\{ \sum_{j\in N_s} \mathbb{E}_{x_j}\theta(x_s,x_j)$

Adv. Variational
13-2

MAP - Relaxations
14-1

Max product BP

Linear programming relaxation

MAP / Relaxations
14-2

MAP/Relaxations

14-3

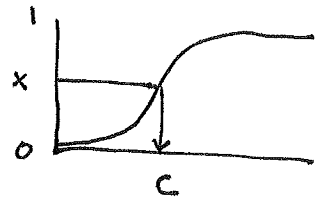
Monte Carlo
15-1

- How to approx a distribution?
  - exact: good
  - variational: fuzzy "marginals"
  - Monte Carlo: lots of samples, "Hard assignments"

- Start at the beginning. Drawing samples.

Assume we have $x \sim \text{Unif}(0,1)$

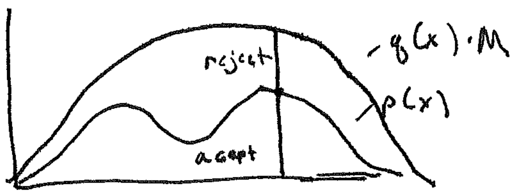
If we know cdf $f(c) = p(y \le c)$ then $f^{-1}(x) = c$ is a sample

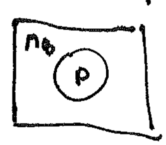


Only really works with univariate, need CDF

- Gaussian Samples (Box mueller trick)

- Rejection Sampling
  - Assume can evaluate $p(x)$ or $\tilde{p}(x)$ but not sample (no partition)





"throw a dart at Mq, see if it lands in p"

Algorithm:

$x_n \sim q(x)$

$u \sim U(0,1)$

if $u < \dfrac{p(x_n)}{Mq(x_n)}$

where $Mq(x) > p(x) \ \forall x$

② or $\underbrace{\tilde{M}}\, q(x) > \tilde{p}(x)$

$\tilde{M} = MZ$

□ Proof: $p(x<x_0 \mid x \text{ accepted}) = \dfrac{p(x<x_0,\, x \text{ accepted})}{p(x \text{ accepted})}$

$$= \frac{\overset{\text{constraint}}{\int_{-\infty}^{x_0}}_x \int_0^1{}_u\, q(x)\cdot \frac{\tilde p(x)}{M q(x)}\, dx\, du}{\int_{-\infty}^{\infty}\int_0^1 \underbrace{q(x)}_{p(x)}\ \underbrace{\frac{\tilde p(x)}{M q(x)}}_{p(\text{accept})}} = \frac{\frac{1}{M}\int_{-\infty}^{x_0} \tilde p(x)\, dx}{\underbrace{\frac{1}{M}\int_{-\infty}^{\infty} \tilde p(x)\, dx}}$$

$$= \int_{-\infty}^{x_0} p(x)\, dx$$ and note that $p(x \text{ accepted})$ small $\underline{M}$

$$\frac{1}{M}\int_{-\infty}^{\infty} \tilde p(x)\, dx = \frac{1}{M'Z}\int_{-\infty}^{\infty} p(x)\, dx = \frac{1}{M'}$$

□ Examples

1) Bayes: let $q(\theta)$ be prior $p(\theta)$ to get samples from posterior $p(\theta \mid x)$

$\tilde p(\theta \mid x) = p(D\mid\theta)\, p(\theta)$

$q(\theta) = p(\theta)$ $\Rightarrow$ $\dfrac{\tilde p(\theta)}{M q(\theta)} = \dfrac{p(D\mid\theta)}{p(D\mid\hat\theta)}$ ← posterior, ← MLE

$M = p(D\mid\hat\theta) = \text{MLE}$

Retain on high-likelihood. Prior controls sampling.

? Why is this okay? MLE ensures $\dfrac{\tilde p(\theta)}{M q(\theta)} \le 1$

2) Gaussians:

$p(x) = \mathcal{N}(0, \sigma_p^2 I)$ $\quad \sigma_q^2 > \sigma_p^2$

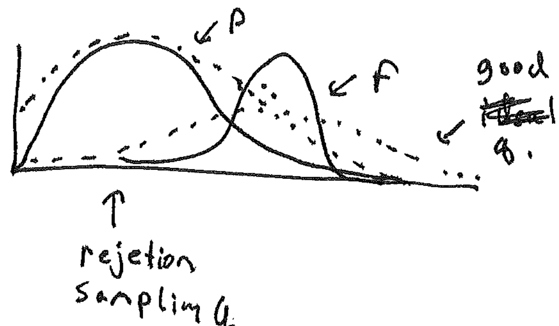
$q(x) = \mathcal{N}(0, \sigma_q^2 I)$

heights $\dfrac{\left(\frac{1}{\sqrt{2\pi}}\right)^D \left(\frac{1}{\sigma_p}\right)^D \exp\{0\}}{\cdot\ 1} = \left(\dfrac{\sigma_q}{\sigma_p}\right)^D = M$ ← dimensionality!

☐ Importance Sampling

- In practice, we are usually sampling to compute expectations $\mathbb{E}_p[f(x)] = \int p(x) f(x)\, dx$
- If we have access to f, it wastes time when $p(x) \uparrow f(x) \downarrow$

Example:



Recall variational ~~bound~~. approach.

$$KL(q \| p) = \int q(x) \log \frac{p(x)}{q(x)} dx$$

Here we introduce q into expectation

$$\int q(x) \frac{p(x)}{q(x)} f(x)\, dx = \mathbb{E}_q\left[f(x) \frac{p(x)}{q(x)}\right] \approx \frac{1}{N} \underbrace{\sum_{x \sim q}}_{\text{Samples from } q} f(x) \underbrace{\frac{p(x)}{q(x)}}_{\text{Weight } w(x)}$$

We can show optimal $q^* = \dfrac{|f(x)|\, p(x)}{\int |f(x')|\, p(x')\, dx}$

Not sure if we need this...

- [ ] What about high-dimensional $x_1 \ldots x_N$

(Next Class)

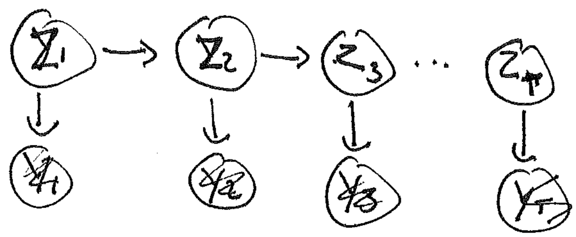
Last class: Importance Sampling $\int f(x)p(x)dx$
utilize function $q(x)$ to draw samples and then reweight.

Today: $x_1, \dots, x_D$ many <u>variables</u> compute sequentially

First note: $p(x') = \int f_{x'}(x)p(x)\,dx$ when $f_{x'}(x) = \delta_{x'}(x)$

~~Can~~ use

Return to time series models



Assume we have seen $y_1 \dots y_t$, want to estimate $p(z_1 \dots z_t | y_1 \dots y_t)$ by sampling

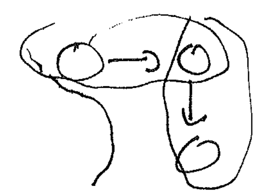
$$p(z_1, \dots z_t | y_1, \dots, y_t) = \sum_{z'} p(z'_1 \dots z'_t | y_1 \dots y_t)\, \delta_{z'}(z) \approx \sum_s \underbrace{\hat{w}_t^s}_{\text{weight}} \underbrace{\delta_{z^s}(z_{1:t})}_{\text{sample same as } z}$$

Again need a proposal $q(z_{1:t}|y_{1:t})$ as before

$$\tilde{w}_t^s = \frac{p(z_{1:t}|y_{1:t})}{q(z_{1:t}|y_{1:t})}$$

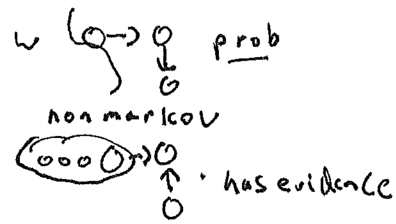
Now let's exploit the conditional independence structure

$$p(z_{1:t}|y_{1:t}) \propto p(z_{1:t}|y_{1:t-1})\, p(y_t|z_{1:t}, y_{1:t-1})$$
$$\propto p(y_t|z_t)\, p(z_t|z_{t-1})\, p(z_{1:t-1}|y_{t-1})$$



$$q(z_{1:t}|y_{1:t}) = q(z_t|z_{1:t-1}, y_{1:t})\, q(z_{1:t-1}|y_{1:t-1})$$] no indep. needed

$$\tilde{w}_t^s = \tilde{w}_{t-1}^s \frac{p(y_t|z_t^s)\, p(z_t^s|z_{t-1}^s)}{q(z_t^s | z_{1:t-1}, y_{1:t})}$$



↓ or markov

$$q(z_t^s | z_{t-1}^s, y_{1:t})$$

□ Algorithm

- sample "particles" $z_t^s \sim q(z_t \mid z_{t-1}^s, y_t)$
- weight particles as $\tilde{\omega}_t^s = \dfrac{p(y_t \mid z_t^s)\, p(z_t^s \mid z_{t-1}^s)}{q(z_t \mid z_{t-1}^s, y_t)}\, \omega_{t-1}^s$
- Normalize particles to compute "filter" $p(z_t \mid y_1 \ldots y_t) = \sum_s \hat{\omega}_t^s\, \delta_{z_t^s}(z_t)$

□ Issue: Sampling in a very high dimensional space

Use appoximation of coverage of distribution.

$$\hat{S}_{eff} = \frac{1}{\sum_s (\tilde{\omega}_t^s)^2}$$

i.e how ~~much of the posterior is q capturing?~~ many of the samples are being used.

two solutions

1) Resample:

Each time step compute $p(z_t \mid y_1 \ldots y_t)$
If $\hat{S}_{eff} \leq$ cutoff, then resample from $p(z_t \mid y_1, \ldots, y_t)$ and start with $\omega_t^s = \frac{1}{S}$ (uniform weighting)

2) $q(z_t \mid z_{t-1}^s, y_t) = p(z_t \mid z_{t-1}^s, y_t)$

not

$q(z_t \mid$ " $= p(z_t \mid z_{t-1}^s)$

$$\begin{matrix} O & \to & O \\ & & \uparrow \\ & & y_t \end{matrix}$$ most kept

$O \to O$ most lost

Particle Filtering
16->3

Application: Linear Gaussian

Show updates for

MCMC
18-1

☐ Monte-Carlo Principle

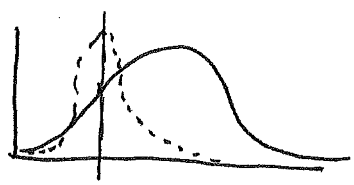
$$\int f(x)p(x)\,dx \approx \frac{1}{N}\sum_n f(x_n) \quad x_n \sim p(x)$$

unbiased, variance $1/N$

How to get $x_n \sim p(x)$?

- exactly (invert, univariate)
- rejection (perfect samples)
- importance (factor in $f(x)$), sample from q.

Philosophical.



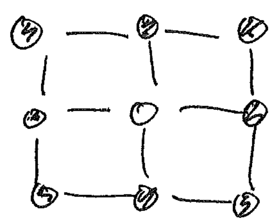
Sample here. This sample is very good.
Why forget about it? No memory in rej./importan

☐ Markov Chain Monte-Carlo

Tradeoff - correlated samples, lose independence, but
more exploration in high prob. regions.

Example: Gibbs Sampling

Idea: assume we have $x_1, \ldots, x_D$, sample x_d



$p(x_d \mid x_1, \ldots x_{d-1} x_{d+1}, x_D) \propto \exp\{x_d \theta \ldots$

Sample each variable in turn

Comparison to mean field.

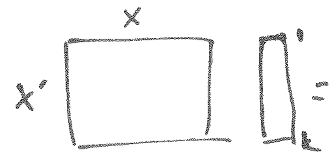
- Similar update process, fix markov blanket
- Mean field: Compute expected value, using expectations of neighbors
- Gibbs : Compute hard assignment by sampling

Gibbs can often by easier to compute.

Markov Chain

$\bigcirc \to \bigcirc \to \bigcirc \to \cdots \to \bigcirc$

- Transition distribution $T(x'|x)$
- Finite example: Transition matrix $\mathbb{R}''$



  - start with ~~π_0~~ π_0 initial dist.
  - Apply transition distribution $\underline{t}$ times

$$\pi_t = T^t \pi_0$$

  - fundamental theorem. Will converge

$$\underline{\pi} = T\underline{\pi}$$

  - equilibrium point $\underline{\pi}$ is "stationary" distribution
  - By definition this is eigenvector with $\lambda = 1$
  - 2nd largest eigenvalue gives rate.
    dominance of rank 1 matrix

- Few more requirements of MCMC
  - irreducible / ergodic: if $\pi(x) > 0$ must be able to reach x from x' in finite steps
  - aperiodic: $\nexists x$ only acessible at times
  - invariance: generalize from matrix

$$\pi(x') = \int T(x'|x)\, \pi(x)\, dx$$

Question: Can we pick T to make $\underline{\pi}(x) = p(x)$

Detailed Balance : Sufficient condition

$$\begin{tikzcd} x \arrow[r, bend left, "T(x'|x)"] & x' \arrow[l, bend left, "T(x|x')"] \end{tikzcd}$$

Suppose we don't know direction / reversible chain / orthogonal matrix

$$\pi(x)\, T(x'|x) = \pi(x')\, T(x|x')$$

↳ implies that

$$\int \pi(x)\, T(x'|x)\, dx = \pi(x')$$

D Properties: $\mathbb{E}\left[\frac{1}{N} \sum_n f(x)\right] = \frac{1}{N} \sum_n \mathbb{E}\left[f(x_n)\right] = \frac{1}{N} \sum_n \int \pi(x) f(x)\, dx = \mathbb{E}[f(x)]$

↳ $x \sim$ MCMC chain

(at stationary dist. asymptotically)

variance: $\mathrm{var}\left(\frac{1}{N} \sum_n f(x_n)\right) = \frac{1}{N^2} \mathrm{var}\left(\sum_n f(x_n)\right)$

$$= \frac{1}{N^2}\left[\underbrace{\sum_n \mathrm{var}(f(x_n))}_{\text{good / as expected } \frac{1}{N}} + 2 \sum_n \sum_{n' < n} \underbrace{\mathrm{cov}(f(x_n), f(x_{n'}))}_{\text{big hit if chain is highly correlated}}\right]$$

(trade off, easy samples versus hard uncorrelated)

☐ Metropolis-Hastings

- Define random walk $q(x'|x)$ "proposal dist"
- Reject if $q(x'|x)$ takes us outside distribution

Formaly

$\frac{1}{Z}$ cancels

1) draw x' from $q(x'|x)$
2) accept w.r.t. $\min\left(1, \frac{p(x')}{p(x)} \frac{q(x|x')}{q(x'|x)}\right)$

$$\Pi \cancel{\leftarrow \mathbb{E}[f(x)] = \int f(x) \frac{p(x')}{p(x)} dx = \int q(x|x') \frac{p(x')}{p(x)\, q(x'|x)}} \quad \text{(Book)}$$

Proof: reverse

$x' \neq x$

$$p(x)\, q(x'|x) \min\left(1, \frac{p(x')}{p(x)} \frac{q(x|x')}{q(x'|x)}\right)$$

$$= \underbrace{\frac{p(x)}{p(x)} \frac{q(x'|x)}{q(x'|x)}}_{\text{cancels}} \underbrace{\min\left(p(x)\, q(x'|x),\, p(x')\, q(x|x')\right)}_{\text{symmetric.}}$$

Stationary if $p(x)$

Mixing Considerations

(?) What if $x \to x_{map}$ w.p. 1 but $x_{map} \to$ unif?

Show pictures