School of Computer Science
Carnegie Mellon

10-708 Probabilistic Graphical Models

Generalized Linear Models
+
Learning Fully Observed Bayes Nets

Readings:
KF Chap. 17
Jordan Chap. 8
Jordan Chap. 9.1 – 9.2

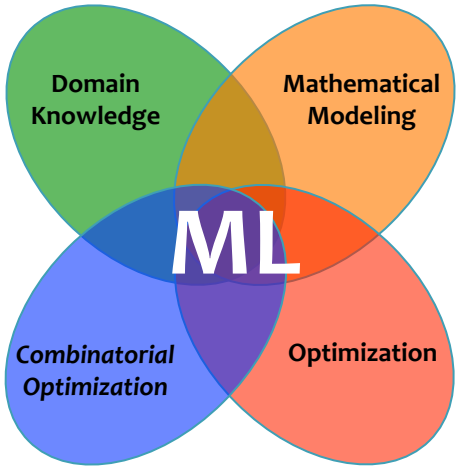
Matt Gormley
Lecture 5
January 27, 2016

Machine Learning

The **data** inspires the structures we want to predict

Our **model** defines a score for each structure

It also tells us what to optimize



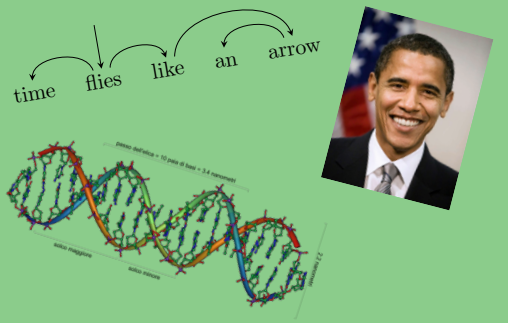
Inference finds {best structure, marginals, partition function} for a new observation

(**Inference** is usually called as a subroutine in learning)

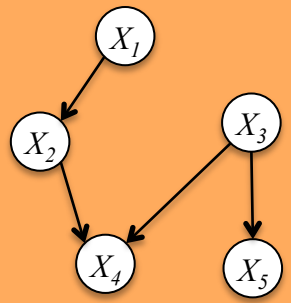
Learning tunes the parameters of the model

Machine Learning

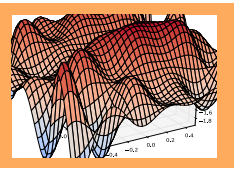
Data



Model



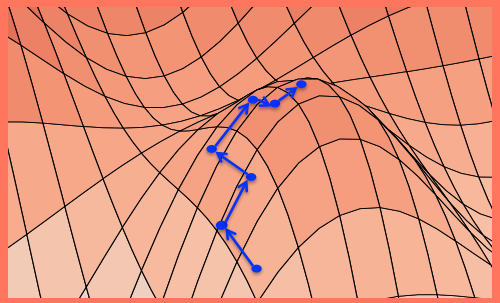
Objective



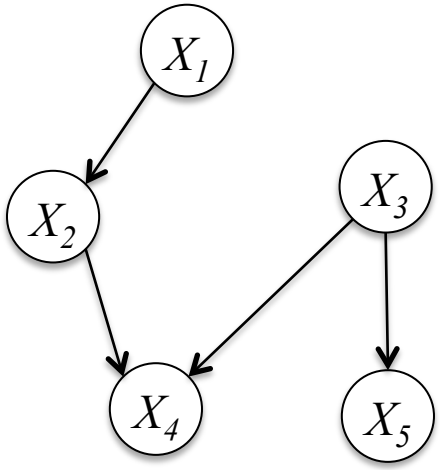
Inference

(**Inference** is usually called as a subroutine in learning)

Learning

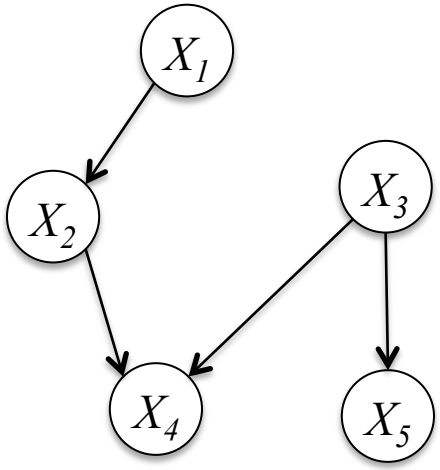


Today's Lecture



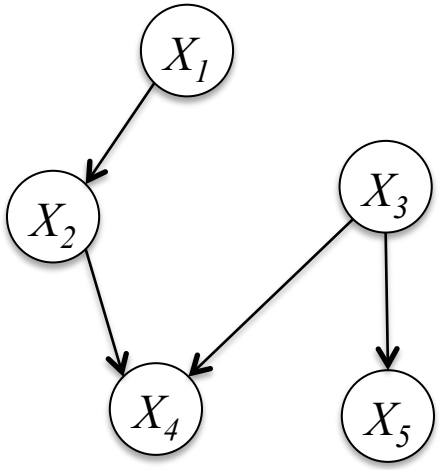
$$p(X_1, X_2, X_3, X_4, X_5) =$$
$$p(X_5|X_3)p(X_4|X_2, X_3)$$
$$p(X_3)p(X_2|X_1)p(X_1)$$

Today's Lecture



$$p(X_1, X_2, X_3, X_4, X_5) =$$
$$p(X_5|X_3)p(X_4|X_2, X_3)$$
$$p(X_3)p(X_2|X_1)p(X_1)$$

Today's Lecture



$$p(X_1, X_2, X_3, X_4, X_5) =$$
$$p(X_5|X_3)p(X_4|X_2, X_3)$$
$$p(X_3)p(X_2|X_1)p(X_1)$$

How do we define and learn these conditional and marginal distributions for a Bayes Net?

Today's Lecture

1. **Exponential Family Distributions**

   A candidate for marginal distributions, $p(X_i)$

2. **Generalized Linear Models**

   Convenient form for conditional distributions, $p(X_j \mid X_i)$

3. **Learning Fully Observed Bayes Nets**

   Easy thanks to decomposability

A candidate for marginal distributions, $p(X_i)$

1. EXPONENTIAL FAMILY

Why the Exponential Family?

1. **Pitman-Koopman-Darmois theorem:** it is the only family of distributions with **sufficient statistics that do not grow** with the size of the dataset
2. Only family of distributions for which **conjugate priors** exist (see Murphy textbook for a description)
3. It is the distribution that is closest to uniform (i.e. **maximizes entropy**) – subject to moment matching constraints
4. Key to **Generalized Linear Models** (next section)
5. Includes some of your favorite distributions!

Adapted from Murphy (2012) textbook

Whiteboard

- Definition of multivariate exponential family
- Example 1: Categorical distribution
- Example 2: Dirichlet distribution

Exponential family, a basic building block

Extra slides from 2015

- For a numeric random variable X

$$p(x \mid \eta) = h(x)\exp\left\{\eta^T T(x) - A(\eta)\right\}$$
$$= \frac{1}{Z(\eta)} h(x)\exp\left\{\eta^T T(x)\right\}$$

is an **exponential family distribution** with natural (canonical) parameter η

- Function $T(x)$ is a *sufficient statistic*.
- Function $A(\eta) = \log Z(\eta)$ is the log normalizer.
- Examples: Bernoulli, multinomial, Gaussian, Poisson, gamma,...

Example: Multivariate Gaussian Distribution

Extra slides from 2015

- For a continuous vector random variable $X \in R^k$:

$$p(x|\mu,\Sigma) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}$$

Moment parameter

$$= \frac{1}{(2\pi)^{k/2}} \exp\left\{-\tfrac{1}{2}\mathrm{tr}\left(\Sigma^{-1}xx^T\right) + \mu^T\Sigma^{-1}x - \tfrac{1}{2}\mu^T\Sigma^{-1}\mu - \log|\Sigma|\right\}$$

- Exponential family representation

Natural parameter

$$\eta = \left[\Sigma^{-1}\mu; -\tfrac{1}{2}\mathrm{vec}\left(\Sigma^{-1}\right)\right] = \left[\eta_1, \mathrm{vec}(\eta_2)\right], \quad \eta_1 = \Sigma^{-1}\mu \text{ and } \eta_2 = -\tfrac{1}{2}\Sigma^{-1}$$

$$T(x) = \left[x; \mathrm{vec}\left(xx^T\right)\right]$$

$$A(\eta) = \tfrac{1}{2}\mu^T\Sigma^{-1}\mu + \log|\Sigma| = -\tfrac{1}{2}\mathrm{tr}(\eta_2\eta_1\eta_1^T) - \tfrac{1}{2}\log(-2\eta_2)$$

$$h(x) = (2\pi)^{-k/2}$$

- Note: a k-dimensional Gaussian is a (d+d^2)-parameter distribution with a (d+d^2)-element vector of sufficient statistics (but because of symmetry and positivity, parameters are constrained and have lower degree of freedom)

Example: Multinomial distribution

- For a binary vector random variable $x \sim \text{multi}(x \mid \pi)$,

$$p(x|\pi) = \pi_1^{x_1}\pi_2^{x_1}\cdots\pi_K^{x_K} = \exp\left\{\sum_k x_k \ln \pi_k\right\}$$

$$= \exp\left\{\sum_{k=1}^{K-1} x_k \ln \pi_k + \left(1 - \sum_{k=1}^{K-1} x_K\right)\ln\left(1 - \sum_{k=1}^{K-1}\pi_k\right)\right\}$$

$$= \exp\left\{\sum_{k=1}^{K-1} x_k \ln\left(\frac{\pi_k}{1-\sum_{k=1}^{K-1}\pi_k}\right) + \ln\left(1 - \sum_{k=1}^{K-1}\pi_k\right)\right\}$$

- Exponential family representation

$$\eta = \left[\ln\left(\pi_k / \pi_K\right); 0\right]$$

$$T(x) = [x]$$

$$A(\eta) = -\ln\left(1 - \sum_{k=1}^{K-1}\pi_k\right) = \ln\left(\sum_{k=1}^{K} e^{\eta_k}\right)$$

$$h(x) = 1$$

Cumulant Generating Property

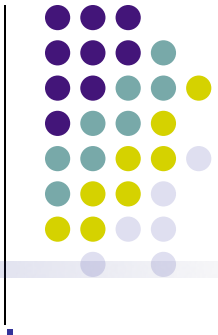
- First cumulant (aka. Mean)

$$\frac{dA}{d\eta} = \frac{d}{d\eta}\log Z(\eta) = \frac{1}{Z(\eta)}\frac{d}{d\eta}Z(\eta)$$
$$= \frac{1}{Z(\eta)}\frac{d}{d\eta}\int h(x)\exp\left\{\eta^T T(x)\right\}dx$$
$$= \int T(x)\frac{h(x)\exp\left\{\eta^T T(x)\right\}}{Z(\eta)}dx$$
$$= E\left[T(x)\right]$$

- Second cumulant (aka. Variance or First central moment)

$$\frac{d^2 A}{d\eta^2} = \int T^2(x)\frac{h(x)\exp\left\{\eta^T T(x)\right\}}{Z(\eta)}dx - \int T(x)\frac{h(x)\exp\left\{\eta^T T(x)\right\}}{Z(\eta)}dx\frac{1}{Z(\eta)}\frac{d}{d\eta}Z(\eta)$$
$$= E\left[T^2(x)\right] - E^2\left[T(x)\right]$$
$$= Var\left[T(x)\right]$$

Moment estimation



- We can easily compute moments of any exponential family distribution by taking the derivatives of the log normalizer $A(\eta)$.
- The q^{th} derivative gives the q^{th} centered moment.

$$\frac{dA(\eta)}{d\eta} = \text{mean}$$

$$\frac{d^2A(\eta)}{d\eta^2} = \text{variance}$$

- When the sufficient statistic is a stacked vector, partial derivatives need to be considered.

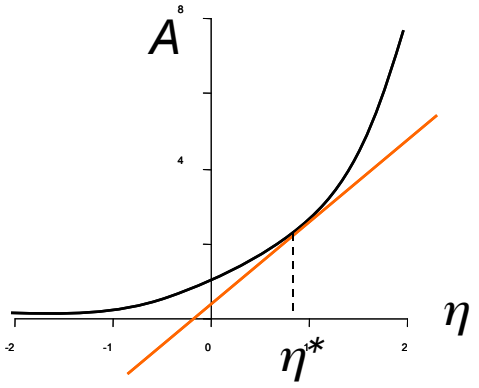
Moment vs canonical parameters

- The moment parameter μ can be derived from the natural (canonical) parameter

$$\frac{dA(\eta)}{d\eta} = E\left[T(x)\right] \overset{\text{def}}{=} \mu$$

- $A(\eta)$ is convex since

$$\frac{d^2 A(\eta)}{d\eta^2} = Var\left[T(x)\right] > 0$$

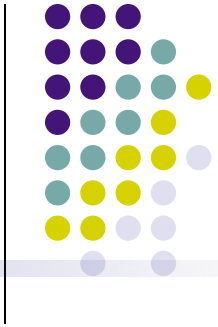


- Hence we can invert the relationship and infer the canonical parameter from the moment parameter (1-to-1):

$$\eta \overset{\text{def}}{=} \psi(\mu)$$

  - A distribution in the exponential family can be parameterized not only by η – the canonical parameterization, but also by μ – the moment parameterization.

MLE for Exponential Family



- For *iid* data, the log-likelihood is

$$\ell(\eta; D) = \log \prod_n h(x_n) \exp\left\{\eta^T T(x_n) - A(\eta)\right\}$$

$$= \sum_n \log h(x_n) + \left(\eta^T \sum_n T(x_n)\right) - NA(\eta)$$

- Take derivatives and set to zero:

$$\frac{\partial \ell}{\partial \eta} = \sum_n T(x_n) - N \frac{\partial A(\eta)}{\partial \eta} = 0$$

$$\Rightarrow \quad \frac{\partial A(\eta)}{\partial \eta} = \frac{1}{N} \sum_n T(x_n)$$

$$\hat{\mu}_{MLE} = \frac{1}{N} \sum_n T(x_n)$$

- This amounts to **moment matching**.
- We can infer the canonical parameters using $\hat{\eta}_{MLE} = \psi(\hat{\mu}_{MLE})$

Examples

- Gaussian:

$$\eta = \left[\Sigma^{-1}\mu; -\tfrac{1}{2}\,\mathrm{vec}\left(\Sigma^{-1}\right)\right]$$
$$T(x) = \left[x; \mathrm{vec}\left(xx^T\right)\right]$$
$$A(\eta) = \tfrac{1}{2}\,\mu^T\Sigma^{-1}\mu + \tfrac{1}{2}\log|\Sigma|$$
$$h(x) = (2\pi)^{-k/2}$$

$$\Rightarrow \mu_{MLE} = \frac{1}{N}\sum_n T_1(x_n) = \frac{1}{N}\sum_n x_n$$

- Multinomial:

$$\eta = \left[\ln\left(\pi_k / \pi_K\right); 0\right]$$
$$T(x) = [x]$$
$$A(\eta) = -\ln\left(1 - \sum_{k=1}^{K-1}\pi_k\right) = \ln\left(\sum_{k=1}^{K} e^{\eta_k}\right)$$
$$h(x) = 1$$

$$\Rightarrow \mu_{MLE} = \frac{1}{N}\sum_n x_n$$

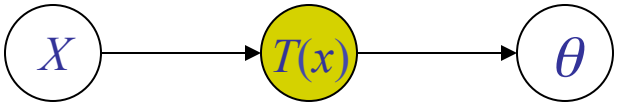
- Poisson:

$$\eta = \log\lambda$$
$$T(x) = x$$
$$A(\eta) = \lambda = e^{\eta}$$
$$h(x) = \frac{1}{x!}$$

$$\Rightarrow \mu_{MLE} = \frac{1}{N}\sum_n x_n$$

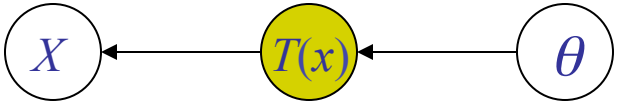
Sufficiency

- For $p(x|q)$, $T(x)$ is *sufficient* for θ if there is no information in X regarding θ beyond that in $T(x)$.
  - We can throw away X for the purpose of inference w.r.t. θ .
  - Bayesian view



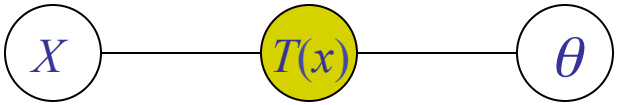
$$p(\theta \mid T(x), x) = p(\theta \mid T(x))$$

  - Frequentist view



$$p(x \mid T(x), \theta) = p(x \mid T(x))$$

  - The Neyman factorization theorem
    - $T(x)$ is *sufficient* for θ if



$$p(x, T(x), \theta) = \psi_1(T(x), \theta)\psi_2(x, T(x))$$

$$\Rightarrow p(x \mid \theta) = g(T(x), \theta)h(x, T(x))$$

Whiteboard

- Bayesian estimation of exponential family

Convenient form for conditional distributions, $p(X_j \mid X_i)$

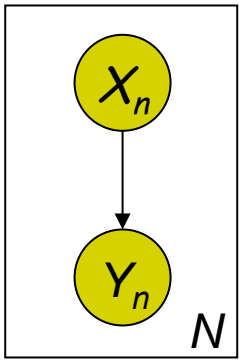
2. GENERALIZED LINEAR MODELS

Why Generalized Linear Models? (GLIMs)

1. Generalization of **linear regression, logistic regression, probit regression**, etc.
2. Provides a **framework for creating new conditional distributions** that come with some convenient properties
3. Special case: GLIMs with canonical response functions are **easy to train** with MLE.
4. *No Free Lunch*: What about **Bayesian estimation of GLIMs?** Unfortunately, we have to turn to approximation techniques since, in general, there isn't a closed form of the posterior.

Generalized Linear Models (GLIMs)

- GLIM
  - The observed input x is assumed to enter into the model via a linear combination of its elements $\xi = \theta^T x$
  - The conditional mean μ is represented as a function $f(\xi)$ of ξ, where f is known as the response function
  - The observed output y is assumed to be characterized by an exponential family distribution with conditional mean μ.



Whiteboard

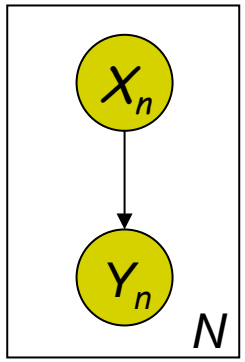
- Constructive definition of GLIMs
- Definition of GLIMs with canonical response functions

Generalized Linear Models (GLIMs)

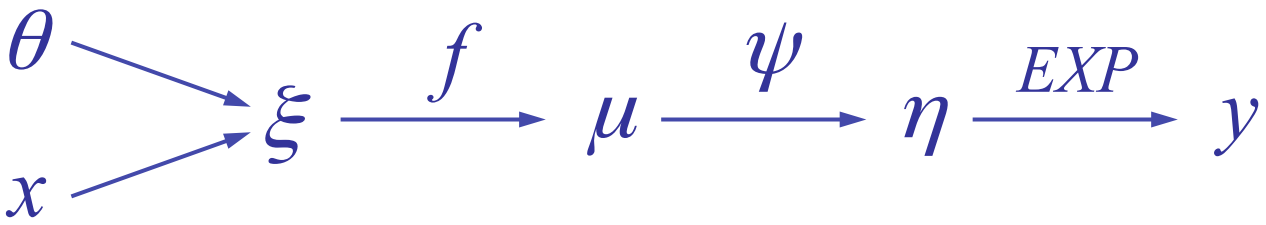
Extra slides from 2015

- The graphical model
  - Linear regression
  - Discriminative linear classification
  - Commonality:

    model $\mathrm{E}_p(Y)=\mu=f(\theta^\top X)$
    - What is $p()$? the cond. dist. of Y.
    - What is $f()$? the response function.



GLIM, cont.

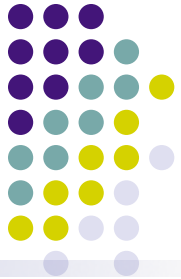


$$p(y \mid \eta) = h(y)\exp\left\{\eta^T(x)y - A(\eta)\right\}$$

$$\Rightarrow p(y \mid \eta, \phi) = h(y, \phi)\exp\left\{\tfrac{1}{\phi}\left(\eta^T(x)y - A(\eta)\right)\right\}$$

- The choice of exp family is constrained by the nature of the data y
  - Example: y is a continuous vector → multivariate Gaussian
    y is a class label → Bernoulli or multinomial
- The choice of the response function
  - Following some mild constrains, e.g., [0,1]. Positivity …
  - Canonical response function: $f = \psi^{-1}(\cdot)$
    - In this case $\theta^T x$ directly corresponds to canonical parameter η.

Example canonical response functions



Model	Canonical response function
Gaussian	$\mu = \eta$
Bernoulli	$\mu = 1/(1+e^{-\eta})$
multinomial	$\mu_i = \eta_i / \sum_j e^{\eta_j}$
Poisson	$\mu = e^{\eta}$
gamma	$\mu = -\eta^{-1}$

MLE for GLIMs with natural response

- Log-likelihood

$$\ell = \sum_n \log h(y_n) + \sum_n \left(\theta^T x_n y_n - A(\eta_n)\right)$$

- Derivative of Log-likelihood

$$\begin{aligned}\frac{d\ell}{d\theta} &= \sum_n \left(x_n y_n - \frac{dA(\eta_n)}{d\eta_n}\frac{d\eta_n}{d\theta}\right) \\ &= \sum_n \left(y_n - \mu_n\right)x_n \\ &= X^T(y-\mu)\end{aligned}$$

This is a fixed point function because μ is a function of θ

- Online learning for canonical GLIMs
  - Stochastic gradient ascent = least mean squares (LMS) algorithm:

$$\theta^{t+1} = \theta^t + \rho\left(y_n - \mu_n^t\right)x_n$$

where $\mu_n^t = \left(\theta^t\right)^T x_n$ and ρ is a step size

Batch learning for canonical GLIMs

- The Hessian matrix

$$H = \frac{d^2\ell}{d\theta d\theta^T} = \frac{d}{d\theta^T}\sum_n (y_n - \mu_n)x_n = \sum_n x_n \frac{d\mu_n}{d\theta^T}$$

$$= -\sum_n x_n \frac{d\mu_n}{d\eta_n}\frac{d\eta_n}{d\theta^T}$$

$$= -\sum_n x_n \frac{d\mu_n}{d\eta_n} x_n^T \quad \text{since } \eta_n = \theta^T x_n$$

$$= -X^T W X$$

$$\mathbf{X} = \begin{bmatrix} -- & \mathbf{x}_1 & -- \\ -- & \mathbf{x}_2 & -- \\ \vdots & \vdots & \vdots \\ -- & \mathbf{x}_n & -- \end{bmatrix}$$

$$\vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

where $X = \left[x_n^T\right]$ is the design matrix and

$$W = \text{diag}\left(\frac{d\mu_1}{d\eta_1}, \ldots, \frac{d\mu_N}{d\eta_N}\right)$$

which can be computed by calculating the 2nd derivative of $A(\eta_n)$

Recall LMS

- Cost function in matrix form:

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i^T\theta - y_i)^2$$
$$= \frac{1}{2}(\mathbf{X}\theta - \vec{y})^T(\mathbf{X}\theta - \vec{y})$$

$$\mathbf{X} = \begin{bmatrix} -- & \mathbf{x}_1 & -- \\ -- & \mathbf{x}_2 & -- \\ \vdots & \vdots & \vdots \\ -- & \mathbf{x}_n & -- \end{bmatrix} \qquad \vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

- To minimize $J(\theta)$, take derivative and set to zero:

$$\nabla_\theta J = \frac{1}{2}\nabla_\theta \mathrm{tr}(\theta^T X^T X\theta - \theta^T X^T\vec{y} - \vec{y}^T X\theta + \vec{y}^T\vec{y})$$
$$= \frac{1}{2}(\nabla_\theta \mathrm{tr}\,\theta^T X^T X\theta - 2\nabla_\theta \mathrm{tr}\vec{y}^T X\theta + \nabla_\theta \mathrm{tr}\vec{y}^T\vec{y})$$
$$= \frac{1}{2}(X^T X\theta + X^T X\theta - 2X^T\vec{y})$$
$$= X^T X\theta - X^T\vec{y} = 0$$

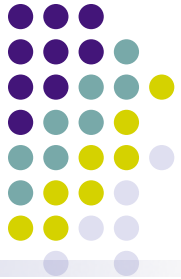
$$\Rightarrow \boxed{X^T X\theta = X^T\vec{y}}$$

The normal equations

$$\Downarrow$$

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

Iteratively Reweighted Least Squares (IRLS)



- Recall Newton-Raphson methods with cost function J

$$\theta^{t+1} = \theta^t - H^{-1}\nabla_\theta J$$

- We now have

$$\nabla_\theta J = X^T(y-\mu)$$

$$H = -X^T W X$$

$$\theta^* = \left(X^T X\right)^{-1} X^T \vec{y}$$

- Now:

$$\begin{aligned}\theta^{t+1} &= \theta^t + H^{-1}\nabla_\theta \ell \\ &= \left(X^T W^t X\right)^{-1}\left[X^T W^t X\theta^t + X^T(y-\mu^t)\right] \\ &= \left(X^T W^t X\right)^{-1} X^T W^t z^t\end{aligned}$$

- where the adjusted response is $z^t = X\theta^t + \left(W^t\right)^{-1}(y-\mu^t)$

- This can be understood as solving the following " Iteratively reweighted least squares " problem

$$\theta^{t+1} = \arg\min_\theta (z - X\theta)^T W (z - X\theta)$$

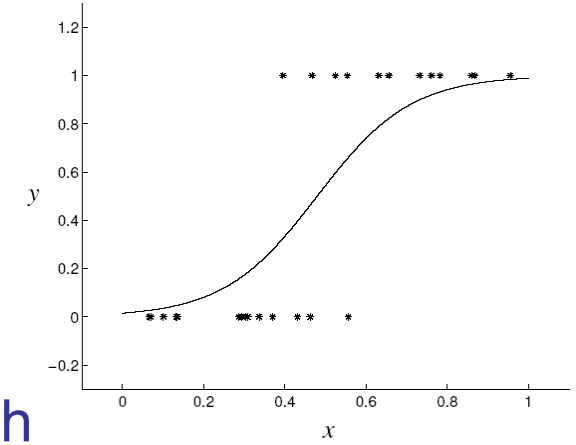
Example 1: logistic regression (sigmoid classifier)

- The condition distribution: a Bernoulli

$$p(y \mid x) = \mu(x)^y (1-\mu(x))^{1-y}$$

where μ is a logistic function

$$\mu(x) = \frac{1}{1+e^{-\eta(x)}}$$



- $p(y|x)$ is an exponential family function, with
  - mean:

$$E[y \mid x] = \mu = \frac{1}{1+e^{-\eta(x)}}$$

  - and canonical response function

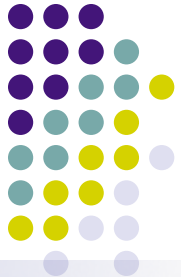
$$\eta = \xi = \theta^T x$$

- IRLS

$$\frac{d\mu}{d\eta} = \mu(1-\mu)$$

$$W = \begin{pmatrix} \mu_1(1-\mu_1) & & \\ & \ddots & \\ & & \mu_N(1-\mu_N) \end{pmatrix}$$

Logistic regression: practical issues



- It is very common to use *regularized* maximum likelihood.

$$p(y = \pm 1 | x, \theta) = \frac{1}{1 + e^{-y\theta^T x}} = \sigma(y\theta^T x)$$

$$p(\theta) \sim \text{Normal}(0, \lambda^{-1} I)$$

$$l(\theta) = \sum_n \log\left(\sigma(y_n \theta^T x_n)\right) - \frac{\lambda}{2}\theta^T\theta$$

  - IRLS takes $O(Nd^3)$ per iteration, where N = number of training cases and d = dimension of input x.
  - Quasi-Newton methods, that approximate the Hessian, work faster.
  - Conjugate gradient takes $O(Nd)$ per iteration, and usually works best in practice.
  - Stochastic gradient descent can also be used if N is large c.f. perceptron rule:

$$\nabla_\theta \ell = \left(1 - \sigma(y_n \theta^T x_n)\right) y_n x_n - \lambda\theta$$

Example 2: linear regression

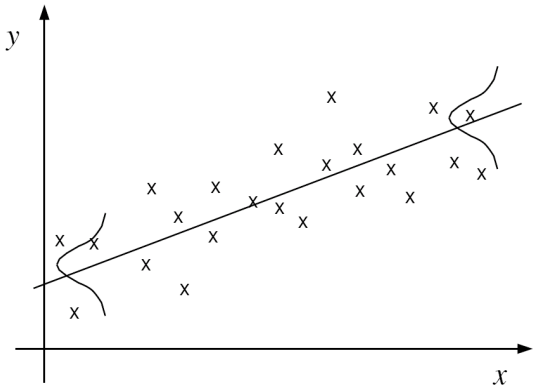
- The condition distribution: a Gaussian

$$p(y|x,\theta,\Sigma) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(y-\mu(x))^T\Sigma^{-1}(y-\mu(x))\right\}$$

Rescale $\Rightarrow h(x)\exp\left\{-\frac{1}{2}\Sigma^{-1}\left(\eta^T(x)y - A(\eta)\right)\right\}$

where μ is a linear function

$$\mu(x) = \theta^T x = \eta(x)$$



- $p(y|x)$ is an exponential family function, with
  - mean: $E[y \mid x] = \mu = \theta^T x$
  - and canonical response function $\eta_1 = \xi = \theta^T x$

- IRLS $\frac{d\mu}{d\eta} = 1$, $W = I$

$$\Rightarrow \quad \theta^{t+1} = \left(X^T W^t X\right)^{-1} X^T W^t z^t = \left(X^T X\right)^{-1} X^T\left(X\theta^t + (y-\mu^t)\right) = \theta^t + \left(X^T X\right)^{-1} X^T (y-\mu^t)$$

Steepest descent

$$\xrightarrow{t\to\infty} \quad \theta = (X^T X)^{-1} X^T Y$$

Normal equation

Today's Lecture

1. **Exponential Family Distributions**

   A candidate for marginal distributions, $p(X_i)$

2. **Generalized Linear Models**

   Convenient form for conditional distributions, $p(X_j \mid X_i)$

3. **Learning Fully Observed Bayes Nets**

   Easy thanks to decomposability

Easy thanks to decomposability

3. LEARNING FULLY OBSERVED BNS

Simple GMs are the building blocks of complex BNs

Density estimation

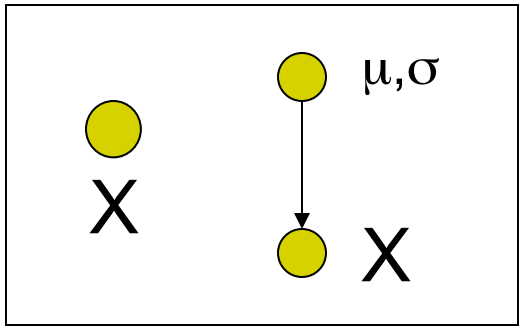
Parametric and nonparametric methods

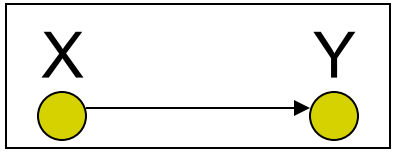
Regression

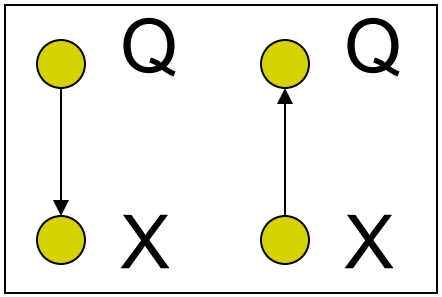
Linear, conditional mixture, nonparametric

Classification

Generative and discriminative approach







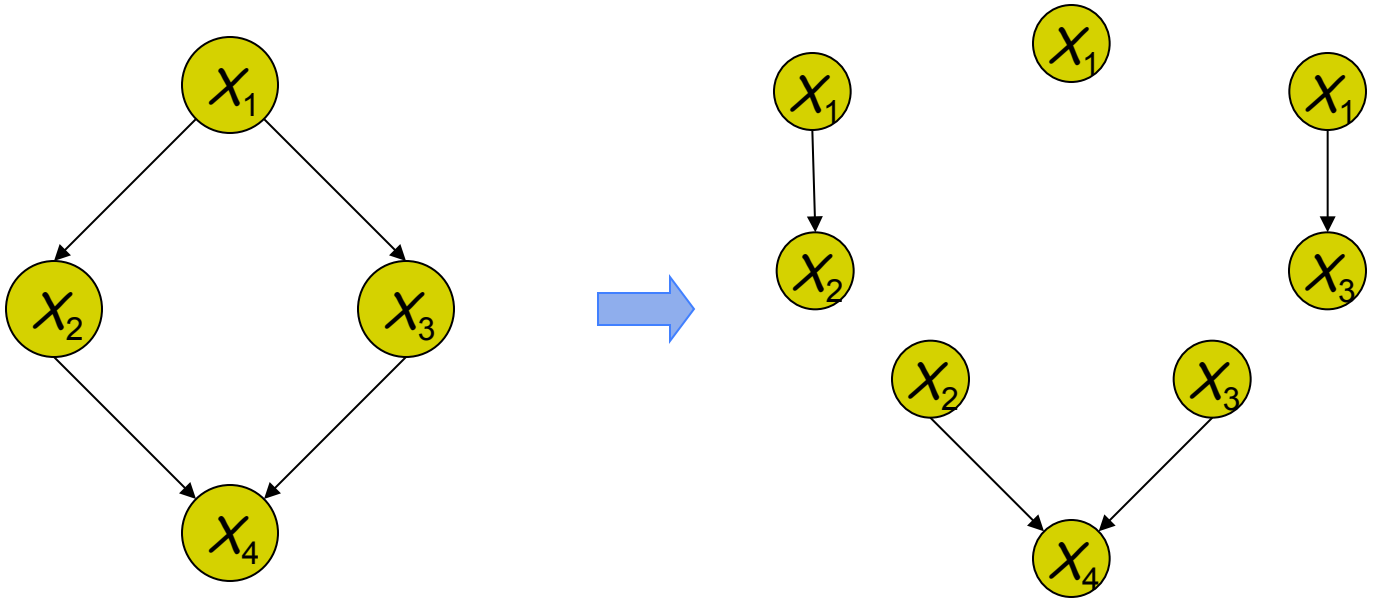
Recall from Lecture 4…

Decomposable likelihood of a BN

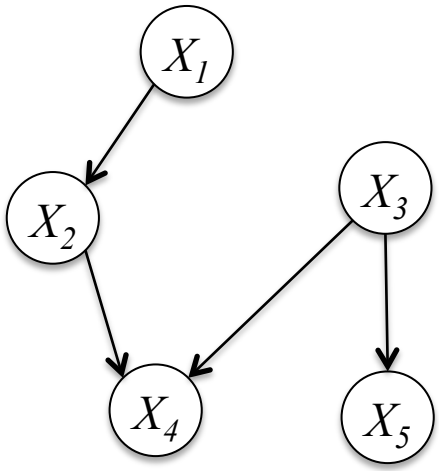
- Consider the distribution defined by the directed acyclic GM:

$$p(x \mid \theta) = p(x_1 \mid \theta_1)\, p(x_2 \mid x_1, \theta_2)\, p(x_3 \mid x_1, \theta_3)\, p(x_4 \mid x_2, x_3, \theta_4)$$

- This is exactly like learning four separate small BNs, each of which consists of a node and its parents.



Learning Fully Observed BNs



$$\begin{aligned}\boldsymbol{\theta}^* &= \operatorname*{argmax}_{\boldsymbol{\theta}} \log p(X_1, X_2, X_3, X_4, X_5) \\ &= \operatorname*{argmax}_{\boldsymbol{\theta}} \log p(X_5|X_3, \theta_5) + \log p(X_4|X_2, X_3, \theta_4) \\ &\qquad + \log p(X_3|\theta_3) + \log p(X_2|X_1, \theta_2) \\ &\qquad + \log p(X_1|\theta_1)\end{aligned}$$

$$\theta_1^* = \operatorname*{argmax}_{\theta_1} \log p(X_1|\theta_1)$$

$$\theta_2^* = \operatorname*{argmax}_{\theta_2} \log p(X_2|X_1, \theta_2)$$

$$\theta_3^* = \operatorname*{argmax}_{\theta_3} \log p(X_3|\theta_3)$$

$$\theta_4^* = \operatorname*{argmax}_{\theta_4} \log p(X_4|X_2, X_3, \theta_4)$$

$$\theta_5^* = \operatorname*{argmax}_{\theta_5} \log p(X_5|X_3, \theta_5)$$

Summary

1. **Exponential Family Distributions**
   - A candidate for marginal distributions, $p(X_i)$
   - Examples: Multinomial, Dirichlet, Gaussian, Gamma, Poisson
   - MLE has closed form solution
   - Bayesian estimation easy with conjugate priors
   - Sufficient statistics by inspection
2. **Generalized Linear Models**
   - Convenient form for conditional distributions, $p(X_j \mid X_i)$
   - Special case: GLIMs with canonical response
     - Output $\boldsymbol{y}$ follows an exponential family
     - Input $\boldsymbol{x}$ introduced via a linear combination
   - MLE for GLIMs with canonical response by SGD
   - In general, Bayesian estimation relies on approximations
3. **Learning Fully Observed Bayes Nets**
   - Easy thanks to decomposability