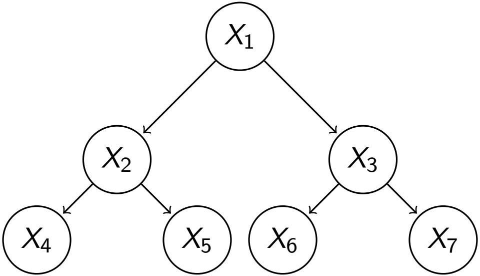
COMP 790-125: Goals for today

- Recap of Bayes Nets
- Exponential family
- Learning in exponential family

Bayes Nets



A graphical representation, a DAG, of a probabilistic model that captures conditional independencies.

$$p(X_1, \ldots, X_N) = \prod_i p(X_i | X_{\mathbf{pa}(i)})$$

This specifies a “skeleton” for the model. We still have to specify these conditional probabilities – we consider some candidates next.

Exponential family

$$p(\mathbf{x}|\eta) = h(\mathbf{x}) \exp\left\{\langle \eta, T(\mathbf{x}) \rangle - A(\eta)\right\}$$

$h(\mathbf{x})$	**base measure**
η	**natural parameter**
$T(\mathbf{x})$	**sufficient statistic**
$A(\eta)$	**log-normalization function**

Exponential family – log-normalization function $A(\eta)$

$$p(x|\eta) = h(\mathbf{x}) \exp\left\{\langle \eta, T(\mathbf{x})\rangle - A(\eta)\right\} = \frac{1}{\exp\left\{A(\eta)\right\}} h(\mathbf{x}) \exp\left\{\langle \eta, T(\mathbf{x})\rangle\right\}$$

We mentioned **normalization function** before, it normalizes a distribution.

The log normalization function $A(\eta)$

$$A(\eta) = \log \int h(\mathbf{x}) \exp\left\{\langle \eta, T(\mathbf{x})\rangle\right\} d\mathbf{x}$$

Exponential family – sufficient statistic $T(\mathbf{x})$

$$p(\mathbf{x}|\eta) = h(\mathbf{x}) \exp\{\langle \eta, T(\mathbf{x}) \rangle - A(\eta)\}$$

The value of $T(\mathbf{x})$ is a *sufficient* summary of **x** for purposes of computing the probability.

Examples of exponential family distributions

- Bernoulli
- Categorical
- Multinomial
- Gaussian
- Laplace (with a known mean)
- and many others

Exponential family – Bernoulli

You would use Bernoulli to model a single toss of a biased coin that gives heads ($x = 1$) with probability π and tails ($x = 0$) with probability $1 - \pi$.

$$
\begin{aligned}
p(x|\pi) &= \pi^x (1-\pi)^{1-x} = \exp\left\{\log\left\{\pi\right\} x + \log\left\{1-\pi\right\}(1-x)\right\} \\
&= \underbrace{1}_{h(x)} \exp\left\{ \underbrace{\log\left\{\frac{\pi}{1-\pi}\right\}}_{\eta} \underbrace{x}_{T(x)} - \underbrace{(-\log 1 - \pi)}_{A(\eta)} \right\}
\end{aligned}
$$

So for Bernoulli

$$
\begin{aligned}
h(x) &= 1 \\
T(x) &= x \\
\eta &= \log\left\{\frac{\pi}{1-\pi}\right\} \\
A(\eta) &= -\log\left\{1-\pi\right\} = \log\left\{1+\exp\eta\right\}
\end{aligned}
$$

Exponential family – Categorical

Generalization of Bernoulli to k outcomes (multi-sided die), parameterized by $\pi_1, \pi_2, \ldots, \pi_k$ such that $\pi_i \geq 0, \sum_i \pi_i = 1$.

$$p(x|\pi) = \prod_{i=1}^{k} \pi_i^{[x=i]} = \exp\left\{ \sum_{i=1}^{k-1} \log\left\{\frac{\pi_i}{\pi_k}\right\} [x=i] - (-\log \pi_k) \right\}$$

and in exponential family form

$$h(x) = 1 \quad T(x) = \begin{bmatrix} [x=1] \\ [x=2] \\ \vdots \\ [x=k-1] \end{bmatrix} \quad \eta = \begin{bmatrix} \log \frac{\pi_1}{\pi_k} \\ \log \frac{\pi_2}{\pi_k} \\ \vdots \\ \log \frac{\pi_{k-1}}{\pi_k} \end{bmatrix}$$

where $\pi_k = 1 - \sum_{i=1}^{k-1} \pi_i$, since it is not a free parameter.
And finally

$$A(\eta) = -\log \pi_k = \log\left\{ 1 + \sum_{i=1}^{k-1} \exp \eta_i \right\}$$

Exponential family – Gaussian

$$\begin{aligned}
p(x|\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}\\
&= \frac{1}{\sqrt{2\pi}}\exp\left\{\frac{\mu}{\sigma^2}x-\frac{1}{2\sigma^2}x^2-\frac{1}{2\sigma^2}\mu^2-\log\{\sigma\}\right\}\\
&= \frac{1}{\sqrt{2\pi}}\exp\left\{\left\langle\begin{bmatrix}\frac{\mu}{\sigma^2}\\ -\frac{1}{2\sigma^2}\end{bmatrix},\begin{bmatrix}x\\ x^2\end{bmatrix}\right\rangle-\frac{1}{2\sigma^2}\mu^2-\log\{\sigma\}\right\}
\end{aligned}$$

and for exponential family form

$$h(x)=\tfrac{1}{\sqrt{2\pi}}\quad T(x)=\begin{bmatrix}x\\ x^2\end{bmatrix}\quad \eta=\begin{bmatrix}\frac{\mu}{\sigma^2}\\ -\frac{1}{2\sigma^2}\end{bmatrix}\quad \begin{aligned}A(\eta)\quad &= \tfrac{1}{2\sigma^2}\mu^2+\log\{\sigma\}\\ &= \tfrac{\eta_1^2}{2\eta_2}+\tfrac{1}{2}\log\left\{\tfrac{1}{2\eta_2}\right\}\end{aligned}$$

Exponential family

You can take many other distributions and convert them into exponential family form.

Once you put a distribution in the exponential family form it is trivial to read out: sufficient statistics, log normalization function, natural parameters.

Exponential family and sufficiency

So why is $T(x)$ called sufficient statistic?

$$p(\mathbf{x}|\eta) = p(\mathbf{x}|\eta) = h(\mathbf{x}) \exp\left\{\langle \eta, T(\mathbf{x}) \rangle - A(\eta)\right\}$$

We refer to $T(\mathbf{x})$ as sufficient because maximum likelihood estimation of η does not require any additional information about $\mathbf{x}$.

Of course, sufficient statistics for one distribution are not necessarily sufficient for another.

Gradient of $A(\eta)$

We will just work out a partial derivative with respect to a single entry of η

$$\begin{aligned}
\frac{\partial A(\eta)}{\partial \eta_i} &= \frac{\partial}{\partial \eta_i} A(\eta) \\
&= \frac{\partial}{\partial \eta_i} \log \int h(\mathbf{x}) \exp\left\{\langle \eta, T(\mathbf{x}) \rangle\right\} d\mathbf{x} \\
&= \frac{\int h(\mathbf{x}) \exp\left\{\langle \eta, T(\mathbf{x}) \rangle\right\} T_i(\mathbf{x}) d\mathbf{x}}{\int h(\mathbf{z}) \exp\left\{\langle \eta, T(\mathbf{z}) \rangle\right\} d\mathbf{z}} \\
&= \int \frac{h(\mathbf{x}) \exp\left\{\langle \eta, T(\mathbf{x}) \rangle\right\}}{\int h(\mathbf{z}) \exp\left\{\langle \eta, T(\mathbf{z}) \rangle\right\} d\mathbf{z}} T_i(\mathbf{x}) d\mathbf{x} \\
&= \int \frac{h(\mathbf{x}) \exp\left\{\langle \eta, T(\mathbf{x}) \rangle\right\}}{\exp\left\{A(\eta)\right\}} T_i(\mathbf{x}) d\mathbf{x} \\
&= \int h(\mathbf{x}) \exp\left\{\langle \eta, T(\mathbf{x}) \rangle - A(\eta)\right\} T_i(\mathbf{x}) d\mathbf{x} \\
&= \int p(\mathbf{x}|\eta) T_i(\mathbf{x}) d\mathbf{x} = \mathbf{E}\left[T_i(\mathbf{x})\right]
\end{aligned}$$

Gradient of $A(\eta)$

$$\frac{\partial A(\eta)}{\partial \eta_i} = \mathbf{E}\left[T_i(\mathbf{x})|\eta\right]$$

A partial derivative of the log normalization function is equal to mean sufficient statistic.

$$\nabla A(\eta) = \mathbf{E}\left[T(\mathbf{x})|\eta\right]$$

Gradient of the log normalization function is equal to mean sufficient statistics.

Maximum likelihood fitting of exponential family

Suppose we have n data instances $\mathbf{x}_1, \ldots, \mathbf{x}_n$ modeled by an exponential family member

$$p(\mathbf{x}|\eta) = h(\mathbf{x}) \exp\left\{\langle \eta, T(\mathbf{x})\rangle - A(\eta)\right\}$$

Then log-likelihood is

$$\begin{aligned} \mathrm{LL}(\eta) &= \sum_{i=1}^{n} \log h(\mathbf{x}_i) + \langle \eta, T(\mathbf{x}_i)\rangle - A(\eta) \\ &= \sum_{i=1}^{n} \left[\log h(\mathbf{x}_i) + \langle \eta, T(\mathbf{x}_i)\rangle\right] - nA(\eta) \end{aligned}$$

Maximum likelihood fitting of exponential family

To maximize the log-likelihood, and consequently likelihood, we take derivatives with respect to η_j of log-likelihood and set it to 0:

$$\frac{\partial \mathrm{LL}(\eta)}{\partial \eta_j} = \sum_{i=1}^{n} T_j(\mathbf{x}_i) - n\frac{\partial A(\eta)}{\partial \eta_j} = 0$$

and this gives us

$$\frac{1}{n}\sum_{i=1}^{n} T_j(\mathbf{x}_i) = \frac{\partial A(\eta^{\mathrm{ML}})}{\partial \eta_j}$$

$$\frac{1}{n}\sum_{i=1}^{n} T_j(\mathbf{x}_i) = \mathbf{E}\left[T_j(\mathbf{x}) | \eta^{\mathrm{ML}}\right]$$

Maximum likelihood fitting of exponential family

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n} T_j(\mathbf{x}_i)}_{\text{sample mean of sufficient statistic}} \qquad = \qquad \underbrace{\mathbf{E}\left[T_j(\mathbf{x})|\eta^{\mathrm{ML}}\right]}_{\text{model mean of sufficient statistic}}$$

The maximum likelihood fit equates the model mean sufficient statistics to the sample mean of sufficient statistics.

The derivative of log-normalization function and its inverse mapping

The other equality is

$$\frac{1}{n}\sum_{i=1}^{n} T_j(\mathbf{x}_i) = \frac{\partial A(\eta)}{\partial \eta_j}$$

and, in principle, we can obtain the optimal η_j

$$\eta_j = \left(\frac{\partial A}{\partial \eta_j}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} T_j(\mathbf{x}_i)\right)$$

Inverse of $\frac{\partial A}{\partial \eta_j}$ for a categorical distribution

$$A(\eta) = \log \left\{ 1 + \sum_{i=1}^{k-1} \exp \eta_i \right\}$$

The gradient and its inverse mapping are

$$\nabla A(\eta) = \begin{bmatrix} \frac{\exp \eta_1}{1+\sum_{i=1}^{k-1} \exp \eta_i} \\ \frac{\exp \eta_2}{1+\sum_{i=1}^{k-1} \exp \eta_i} \\ \vdots \\ \frac{\eta_{k-1}}{1+\sum_{i=1}^{k-1} \exp \eta_i} \end{bmatrix} = \begin{bmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_{k-1} \end{bmatrix} \quad \nabla A^{-1}(\pi) = \begin{bmatrix} \log \frac{\pi_1}{\pi_k} \\ \log \frac{\pi_2}{\pi_k} \\ \vdots \\ \log \frac{\pi_{k-1}}{\pi_k} \end{bmatrix} = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_{k-1} \end{bmatrix}$$

Maximum likelihood fitting of categorical distribution

Compute sample sufficient statistics

$$\frac{1}{n}\sum_{i=1}^{n} T_j(x_i) = \frac{1}{n}\sum_{i=1}^{n}[x_i = j] = \hat{\pi}_j$$

and apply inverse mapping $(\nabla A)^{-1}(\hat{\pi}) = \hat{\eta}$. This yields MLE estimate for natural parameters.

Contrast this to the constrained optimization we derived for PWM problem – we did not need to write out a Lagrangian.

Exponential family

1. Put your distribution into exponential family form
2. Determine sufficient statistic $T(x)$
3. Compute sample mean sufficient statistics $f = \frac{1}{n} \sum_i T(x_i)$
4. Compute $(\nabla A)^{-1}(f)$ to obtain MLE natural parameters.

In some cases this might not be feasible. You still have an alternative of writing out the log-likelihood and maximizing it using an optimization technique.

Fitting completely observed Bayes Net

We will now turn our attention to fitting a Bayes Net that has been fully observed.

This means is that each data instance has values for all random variables in your Bayes Net.

Note that the linear and logistic regression fell into this category, both $\mathbf{X}$ and $\mathbf{y}$ were completely observed.

Fitting a completely observed Bayes Net

The log-likelihood

$$\mathrm{LL}(\theta; \mathbf{x}) = \sum_t \sum_i \log p(x_i^t | \mathbf{x}_{\mathbf{pa}}^t(i), \theta_i)$$

where $\mathbf{x}^t$ is t^{th} instance of completely observed state of a Bayes Net.

We also note that we assumed that each conditional probability $p(x_i^t | \mathbf{x}_{\mathbf{pa}}^t(i))$ has separate parameters θ_i.

As a result we can separate the maximization of LL into separate optimization problems of type

$$\operatorname*{argmax}_{\theta_i} \sum_t \log p(x_i^t | \mathbf{x}_{\mathbf{pa}}^t(i), \theta_i)$$

Fitting a completely observed Bayes Net

$$\operatorname*{argmax}_{\theta_i} \sum_t \log p(x_i^t | \mathbf{x}^t_{\mathbf{pa}}(i), \theta_i)$$

If $p(x_i^t | \mathbf{x}^t_{\mathbf{pa}}(i), \theta_i)$ is in exponential family and θ_i are natural parameters then

$$\hat{\theta}_i = \left(\frac{\partial A}{\partial \eta_j}\right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} T_j \left(\begin{bmatrix} \mathbf{x}_i \\ \mathbf{x}_{\mathbf{pa}(i)} \end{bmatrix}\right)\right)$$

The inverse mappings can be found in any material that talks about exponential families (even wikipedia).

We did ...

- Exponential Family
- MLE fitting of Exponential Family
- Fitting completely observed Bayesian Networks