CPSC 440: Advanced Machine Learning

Exponential Families

Mark Schmidt

University of British Columbia

Winter 2022

Previously: Density Estimation with Categorical/Gaussian Distributions

- We have discussed density estimation with categorical and Gaussian distribution.
  - Binary is special case of categorical.

- These distributions have a lot of nice properties for learning/inference.
  - NLL is convex, and MLE has closed-form (statistics in training data).
  - Exists conjugate prior, so posterior is prior with "updated hyper-parameters".

- But these distributions make restrictive assumptions:
  - Categorical assumes categories are unordered, non-hierarchical, and finite.
  - Gaussian assumes symmetry, full support, no outliers, uni-modal.

- Many alternatives to categorical/Gaussian exist (examples later).
  - Whether or not they maintain nice properties is related to exponential family.

Exponential Family: Definition

- General form of exponential family likelihood for data x with parameters θ is

$$p(x \mid \theta) = \frac{h(x)\exp(\eta(\theta)^T s(x))}{Z(\theta)}.$$

- The value $s(x)$ is called the sufficient statistics.
  - $s(x)$ tells us everything that is relevant to θ about data x.

- The parameter function η controls how parameters θ interact with statistics.
  - We focus a lot on $\eta(\theta) = \theta$, which is called the cannonical form.

- The support function h contains terms that do not depend on w.
  - Also called the base measure.

- The normalizing constant Z ensures it sums/integrates to 1 over x.
  - Also called the partition function.

Bernoulli as Exponential Family

- Is Bernoulli in the exponential family for some parameters w?

$$p(x \mid \theta) = \theta^x (1-\theta)^{1-x} \stackrel{?}{=} \frac{h(x)\exp(\eta(\theta)^T F(x))}{Z(\theta)}.$$

- To get an exponential, take log of exp (cancelling operations),

$$\begin{aligned}
p(x \mid \theta) = \theta^x (1-\theta)^{1-x} &= \exp(\log(\theta^x (1-\theta)^{1-x})) \\
&= \exp(x \log \theta + (1-x)\log(1-\theta)) \\
&= (1-\theta)(\exp\left(x \log\left(\frac{\theta}{1-\theta}\right)\right).
\end{aligned}$$

- The sufficient statistic is $s(x) = x$ and normalizing constant is $Z(\theta) = 1/(1-\theta)$.
- The parameter is $\eta(\theta) = \log(\theta/(1-\theta))$ (the log odds).
  - Not in canonical form. Canonical form would use log odds directly as the parameter.
- For the support function, $h(x) = 1$ if $x = 0$ or $x = 1$ and $h(x) = 0$ otherwise.
  - There are other ways to write Bernoulli as an exponential family.

Gaussian as Exponential Family

- Writing univariate Gaussian as an exponential family:

$$\begin{aligned}
p(x \mid \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-(x-\mu)^2/2\sigma^2\right) \\
&= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-x^2/2\sigma^2 + \mu x/\sigma^2 - \mu^2/2\sigma^2\right) \\
&= \frac{1}{\sqrt{2\pi}} \frac{\exp\left(-\mu^2/2\sigma^2\right)}{\sigma} \exp\left(\begin{bmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{bmatrix}^T \begin{bmatrix} x \\ x^2 \end{bmatrix} \right).
\end{aligned}$$

- The sufficient statistics are x and x^2, and parameters are μ/σ^2 and $-1/2\sigma^2$
- The normalizing constant is $\sigma \exp(\mu^2/2\sigma^2)$, and support is $1/\sqrt{2\pi}$.

- Again, there is more than one way to represent as an exponential family.
  - If σ^2 is not a parameter, then x/σ^2 is the sufficient statistic and μ is cannonical.

Learning with Exponential Families

- With n IID examples and cannonical paramaters, the likelihood can be written

$$
\begin{aligned}
p(X \mid \theta) &= \prod_{i=1}^{n} h(x^i) \frac{\exp(\theta^T s(x^i))}{Z(\theta)} \\
&= \frac{1}{Z(\theta)^n} \exp\left(\theta^T \sum_{i=1}^{n} s(x^i)\right) \prod_{j=1}^{n} h(x^i) \\
&= \frac{\exp(\theta^T s(X))}{Z(\theta)^n} \prod_{j=1}^{n} h(x^i),
\end{aligned}
$$

  where the sufficient statistics of the data are $s(X) = \sum_{i=1}^{n} s(x^i)$.

- The sufficient statistics of the data $s(X)$ contain everything relevant for learning.
  - For Gaussians, only knowledge of data we need is $\sum_{i=1}^{n} x^i$ and $\sum_{i=1}^{n} (x^i)^2$.

Learning with Exponential Families

- With n IID examples and cannonical paramaters, the NLL can be written

$$f(\theta) = -\theta^T s(X) + n \log Z(\theta) + \text{const},$$

where we see that once we know $s(X)$, we can throw away data.
  - No point in using SGD, you just compute s on each example once.
- The gradient divided by n (average NLL) for a feature j has the form

$$\begin{aligned}
\frac{1}{n}\nabla_{\theta_j} f(\theta) &= -\frac{1}{n}s_j(X) + \sum_x h(x)\frac{\exp(\theta^T s(X))}{Z(\theta)} s_j(X) \quad (\text{use} \int \text{for continuous } x) \\
&= -\frac{1}{n}s_j(X) + \sum_x p(x \mid \theta) s_j(X) \\
&= -\mathbb{E}_{\text{data}}[s_j(X)] + \mathbb{E}_{\text{model}}[s_j(X)].
\end{aligned}$$

- The stationary points where $\nabla f(\theta) = 0$ correspond to moment matching:
  - Set parameters θ so that expected sufficient statistics equal to statistics in data.
  - This is the source of the simple/intuitive closed-form MLEs we have seen.

Convexity and Entropy in Exponential Families

- If you take the second derivative of the NLL you get

$$\nabla^2 f(\theta) = \mathbb{V}[s(X)],$$

  the covariance of the sufficient statistics.
  - Covariances are positive semi-definite, $\mathbb{V}[s(X)] \succeq 0$, so NLL is convex.
  - This is why "setting the gradient to zero and solve for θ" gives MLE.
- Higher-order derivatives give higher-order moments.
  - We call $\log(Z)$ the cumulant function.
- Can show MLE maximizes entropy over all distributions that match moments.
  - Entropy is a measure of "how random" a distribution is.
  - So Gaussian is "most random" distribution that fits means and covariance of data.
    - Or you can think of this as Gaussian makes "least assumptions".
  - Details for special case of $h(x) = 1$ in bonus slides.

Conjugate Priors in Exponential Family

- Exponential families in canonical form are guaranteed to have conjugate priors.
  - For example, we could choose

$$p(\theta \mid \alpha) \propto \frac{\exp(\theta^T \alpha)}{Z(\theta)^k},$$

    where α represent "pseudo-counts" for the sufficient statistics.
    - And k modifies stength of prior (Z above is normalizer for the likelihood).
  - Posterior would have the same form,

$$p(\theta \mid X, \alpha) \propto \frac{\exp(\theta^T (s(X) + \alpha))}{Z(\theta)^{n+k}}.$$

- Can use prior's normalizing constant for Bayesian inference.
  - Ratio of normalizing constants gives posterior predicttive and marginal likelihood.

Discriminative Models and the Exponential Family

- Going from an exponential family to a discriminative supervised learning:
  - Set cannonical parameter to $w^T x^i$.
  - Gives a convex NLL, where MLE tries to match dasta/model's conditional statistics.

- For example, consider Gaussian with fixed variance for y^i.
  - Cannonical parameter is μ, and we know setting $\mu = w^T x^i$ gives least squares.

- If we start with Bernoulli for y^i, we obtain logistic regression.
  - Canonical parmaeter is log-odds.
  - Set $w^T x^i = \log(y^i/(1-y^i))$ and solve for y^i to get sigmoid function.
    - This is my very-delayed answer to "why use the sigmoid function?".

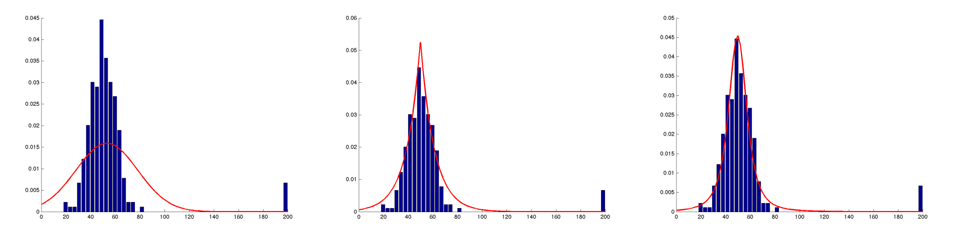
- You can obtain regression models for other settings using this appraoch.
  - Set canonical parameters to $v^T h(W^2 h(W^1 x^i))$ for neural networks.
  - Use a different exponential family to handle a different type of data.

Examples of Exponential Families

- Bernoulli: distribution on $\{0, 1\}$.
- Categorical: distribution on $\{1, 2, \ldots, k\}$.
- Gaussian: distribution on $\mathbb{R}^d$.
- Beta: distribution on $[0, 1]$ (including uniform).
- Dirichlet: distribution on discrete probabilities.
- Wishart: distribution on positive-definite matrices.
- Poisson: distribution on non-negative integers.
- Gamma: distribution on positive real numbers.
- Many others, see here:
  - `en.wikipedia.org/wiki/Exponential_family#Table_of_distributions`

Non-Examples of Exponential Families

- Laplace and student t distribution are not exponential families.



  - "Heavy-tailed": have larger probability that data is far from mean.
  - More robust to outliers than Gaussian.
- Ordinal logistic regression is not in exponential family.
  - Can be used for categorical variables where ordering matters.
- In these cases, we may not have nice properties:
  - MLE may not be intuitive or closed-form, NLL may not be convex.
  - May not have conjugate prior, so need Monte Carlo or variational methods.

Convex Conjugate and Entropy

- The convex conjugate of a function A is given by

$$A^*(\mu) = \sup_{w \in \mathcal{W}} \{\mu^T w - A(w)\}.$$

- E.g., if we consider for logistic regression

$$A(w) = \log(1 + \exp(w)),$$

  we have that $A^*(\mu)$ satisfies $w = \log(\mu)/\log(1-\mu)$.
  - When $0 < \mu < 1$ we have

$$\begin{aligned} A^*(\mu) &= \mu \log(\mu) + (1-\mu)\log(1-\mu) \\ &= -H(p_\mu), \end{aligned}$$

    negative entropy of binary distribution with mean μ.
  - If μ does not satisfy boundary constraint, $\sup$ is ∞.

Convex Conjugate and Entropy

- More generally, if $A(w) = \log(Z(w))$ for an exponential family then

$$A^*(\mu) = -H(p_\mu),$$

subject to boundary constraints on μ and constraint:

$$\mu = \nabla A(w) = \mathbb{E}[s(X)].$$

- Convex set satisfying these is called marginal polytope $\mathcal{M}$.
- If A is convex (and LSC), $A^{**} = A$. So we have

$$A(w) = \sup_{\mu \in \mathcal{U}} \{w^T \mu - A^*(\mu)\}.$$

and when $A(w) = \log(Z(w))$ we have

$$\log(Z(w)) = \sup_{\mu \in \mathcal{M}} \{w^T \mu + H(p_\mu)\}.$$

- This can be used to derive variational methods, since we have written computing $\log(Z)$ as a convex optimization problem.

Maximum Likelihood and Maximum Entropy

- The maximum likelihood parameters w in exponential family satisfy:

$$
\begin{aligned}
&\min_{w\in\mathbb{R}^d} -w^T s(D) + \log(Z(w)) \\
&= \min_{w\in\mathbb{R}^d} -w^T s(D) + \sup_{\mu\in\mathcal{M}} \{w^T \mu + H(p_\mu)\} && \text{(convex conjugate)} \\
&= \min_{w\in\mathbb{R}^d} \sup_{\mu\in\mathcal{M}} \{-w^T s(D) + w^T \mu + H(p_\mu)\} \\
&= \sup_{\mu\in\mathcal{M}} \{ \min_{w\in\mathbb{R}^d} -w^T s(D) + w^T \mu + H(p_\mu)\} && \text{(convex/concave)}
\end{aligned}
$$

  which is $-\infty$ unless $s(D) = \mu$ (e.g., maximum likelihood w), so we have

$$
\begin{aligned}
&\min_{w\in\mathbb{R}^d} -w^T s(D) + \log(Z(w)) \\
&= \max_{\mu\in\mathcal{M}} H(p_\mu),
\end{aligned}
$$

  subject to $s(D) = \mu$.
- Maximum likelihood $\Rightarrow$ maximum entropy $+$ moment constraints.