Generative Classification, Exponential Family Distributions

Piyush Rai

Topics in Probabilistic Modeling and Inference (CS698X)

Jan 30, 2018

Recap: (Bayesian) Logistic Regression

- Models the conditional probability of label given features as

$$p(y=1|\boldsymbol{x},\boldsymbol{w}) = \sigma(\boldsymbol{w}^\top\boldsymbol{x}) = \frac{1}{1+\exp(-\boldsymbol{w}^\top\boldsymbol{x})} = \frac{\exp(\boldsymbol{w}^\top\boldsymbol{x})}{1+\exp(\boldsymbol{w}^\top\boldsymbol{x})}$$

  where $\boldsymbol{w} \in \mathbb{R}^D$ is the weight vector. Can assume a prior $p(\boldsymbol{w}) = \mathcal{N}(0, \lambda^{-1}\mathbf{I})$

- MLE, MAP, fully Bayesian inference can be carried out for this model
  - Due to non-conjugacy, fully Bayesian approach requires approximate inference!

- Can extend it to multinomial logistic (aka "softmax") regression for $K > 2$ classes

$$p(y=k|\boldsymbol{x},\mathbf{W}) = \frac{\exp(\boldsymbol{w}_k^\top\boldsymbol{x})}{\sum_{k=1}^K \exp(\boldsymbol{w}_k^\top\boldsymbol{x})} \qquad k = 1,\ldots,K$$

  where $\mathbf{W} = [\boldsymbol{w}_1,\ldots,\boldsymbol{w}_K]$. Inference similar to logistic regression.

Generative Classification

- Logistic/Softmax Regression models the conditional probability of labels given features directly

$$\text{Logistic:} \quad p(y=1|\boldsymbol{x},\boldsymbol{w}) = \frac{\exp(\boldsymbol{w}^\top \boldsymbol{x})}{1+\exp(\boldsymbol{w}^\top \boldsymbol{x})}$$

$$\text{Softmax:} \quad p(y=k|\boldsymbol{x},\mathbf{W}) = \frac{\exp(\boldsymbol{w}_k^\top \boldsymbol{x})}{\sum_{k=1}^K \exp(\boldsymbol{w}_k^\top \boldsymbol{x})} \qquad k=1,\ldots,K$$

- These are examples of Discriminative Classification (directly define a class discriminator function!)
- Another alternative is to use Generative Classification
  - Doesn't define the distribution of y given $\boldsymbol{x}$ directly but as a <u>two step process</u>
  - Step 1: Using training data, estimate class-marginals $p(y)$ and class-conditional $p(\boldsymbol{x}|y)$
  - Step 2: Use these to compute conditional probability of label given features as

$$p(y=k|\boldsymbol{x}) = \frac{p(\boldsymbol{x},y=k)}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x},y=k)}{\sum_{k=1}^K p(\boldsymbol{x},y=k)} = \frac{p(y=k)p(\boldsymbol{x}|y=k)}{\sum_{k=1}^K p(y=k)p(\boldsymbol{x}|y=k)}$$

- Note: $p(y)$ and $p(\boldsymbol{x}|y)$ can be learned using point estimation or fully Bayesian inference(we have already seen several examples of learning such distributions from data)

Generative Classification: The "Generative Story"

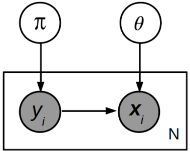
- Assuming binary labels, we can define a "generative story" for each example $(\boldsymbol{x}_i, y_i)$
  - First draw ("generate") a binary label $y_i \in \{0, 1\}$

$$y_i \sim \text{Bernoulli}(\pi)$$

  - Now draw ("generate") the feature vector $\boldsymbol{x}_i$ from the distribution of class $y_i \in \{0, 1\}$

$$\boldsymbol{x}_i | y_i \sim p(\boldsymbol{x} | \theta_{y_i})$$

- Writing $\theta = (\theta_0, \theta_1)$, the above generative model shown in "plate notation" (shaded = observed)



- Note that the above implicity defines a joint distribution of the form $p(\boldsymbol{x}, y) = p(y)p(\boldsymbol{x}|y)$

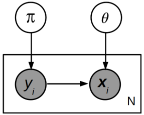
Generative Classification: Learning Procedure

- Recall the rule: $p(y=k|\boldsymbol{x}) = \frac{p(y=k)p(\boldsymbol{x}|y=k)}{\sum_{k=1}^{K} p(y=k)p(\boldsymbol{x}|y=k)}$. How do we learn $p(y)$ and $p(\boldsymbol{x}|y)$?
- Need to choose the distribution $p(y)$. It has to be a discrete distribution, e.g.,
    - For binary y, $p(y) = p(y|\pi) = \text{Bernoulli}(\pi)$ with $\pi \in (0,1)$
    - For $y \in \{1,\ldots,K\}$, $p(y) = p(y|\boldsymbol{\pi}) = \text{multinoulli}(\boldsymbol{\pi})$ where $\boldsymbol{\pi} = [\pi_1,\ldots,\pi_K]$ with $\sum_{k=1}^{K} \pi_k = 1$
- Need to choose the distribution $p(\boldsymbol{x}|y)$. Its form will depend on the type of $\boldsymbol{x}$, e.g.,
    - For $\boldsymbol{x} \in \mathbb{R}^D$, we can choose $p(\boldsymbol{x}|y)$ as a multivariate Gaussian (one for each class)

$$p(\boldsymbol{x}|y=k) = p(\boldsymbol{x}|\theta_k) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- Given training data, we can estimate these distributions. Note: In the fully Bayesian approach
    - Instead of $p(y_*|\hat{\pi})$, we'd infer $p(y_*|\boldsymbol{y}) = \int p(y_*|\pi)p(\pi|\boldsymbol{y})d\pi$
    - Instead of $p(\boldsymbol{x}_*|\hat{\theta}_k)$, we'd infer $p(\boldsymbol{x}_*|\mathbf{X}_k) = \int p(\boldsymbol{x}_*|\theta_k)p(\theta_k|\mathbf{X}_k)d\theta_k$ ($\mathbf{X}_k$: inputs from class k)
- We have already looked at various examples for estimating distributions (via point estimation and Bayesian inference), e.g., for Bernoulli, Gaussian, etc., given data

Generative Classification: Some Benefits



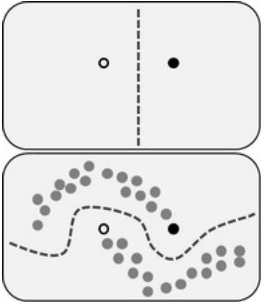
- Can incorporate class prior information (relative frequencies of classes) via $p(y)$
- Can be used in semi-supervised learning (SSL) settings (when only part of the data is labeled)
  - Usually learned in an alterating fashion
    1. "Guess" the missing labels using the current estimates of $p(y)$ and $p(\boldsymbol{x}|y)$

    $$p(y = k|\boldsymbol{x}) = \frac{p(y = k)p(\boldsymbol{x}|y = k)}{\sum_{k=1}^{K} p(y = k)p(\boldsymbol{x}|y = k)}$$

    2. Update $p(y)$ and $p(\boldsymbol{x}|y)$ using the examples with guessed labels + the other labeled examples
- Not limited to learning linear boundaries (unlike logistic/softmax while is a linear models)
  - Form of the class-conditional $p(\boldsymbol{x}|y)$ controls the shape of the learned decision boundary
- Can leverage recent advances in generative models to learn very flexible forms for $p(\boldsymbol{x}|y)$

An Aside: Why Unlabeled Examples Might Help in Classification?

- The unlabeled example give us a (rough) sense of what each class looks like



- SSL algorithms can leverage this information via the class-conditional $p(\boldsymbol{x}|y)$

Exponential Family (Pitman, Darmois, Koopman, Late 1930s)

- Defines a class of distributions. An Exponential Family distribution is of the form

$$\boxed{p(\boldsymbol{x}|\theta) = \frac{1}{Z(\theta)} h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x})] \quad = \quad h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x}) - A(\theta)]}$$

- $\boldsymbol{x} \in \mathcal{X}^m$ is the random variable being modeled (where $\mathcal{X}$ denotes some space, e.g., $\mathbb{R}$ or $\{0,1\}$)
- $\theta \in \mathbb{R}^d$: Natural parameters or canonical parameters defining the distribution
- $\phi(\boldsymbol{x}) \in \mathbb{R}^d$: Sufficient statistics (another random variable)
    - Why "sufficient": $p(\boldsymbol{x}|\theta)$ as a function of θ depends on $\boldsymbol{x}$ only via $\phi(\boldsymbol{x})$
- $Z(\theta) = \int h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x})] d\boldsymbol{x}$: Partition function
- $A(\theta) = \log Z(\theta)$: Log-partition function (also called the cumulant function)
- $h(\boldsymbol{x})$: A constant (doesn't depend on θ)

Expressing a Distribution in Exponential Family Form

A general trick to represent any distribution (assuming it is exp-family dist.) in exp-family form

- Write the given distribution as $\exp(\log p())$ and simplify, e.g., for the Binomial

$$
\begin{aligned}
\exp\left(\log \text{Binomial}(x|N,p)\right) &= \exp\left(\log \binom{N}{x} p^x (1-p)^{N-x}\right) \\
&= \exp\left(\log \binom{N}{x} + x\log p + (N-x)\log(1-p)\right) \\
&= \binom{N}{x} \exp\left(x \log \frac{p}{1-p} - N\log(1-p)\right)
\end{aligned}
$$

- Now compare the resulting expression with the exponential family form

$$p(\boldsymbol{x}|\theta) = h(\boldsymbol{x})\exp(\theta^\top \phi(\boldsymbol{x}) - A(\theta))$$

.. to identify the natural parameters, sufficient statistics, log-partition function, etc.

(Univariate) Gaussian as Exponential Family

- Let's try to write a univariate Gaussian in the exponential family form

$$p(\boldsymbol{x}|\theta) = h(\boldsymbol{x})\exp[\theta^\top \phi(\boldsymbol{x}) - A(\theta)]$$

- Recall the standard definition of a univariate Gaussian

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

$$= \frac{1}{\sqrt{2\pi}}\exp\left[\begin{bmatrix}\frac{\mu}{\sigma^2}\\ -\frac{1}{2\sigma^2}\end{bmatrix}^\top \begin{bmatrix}x\\ x^2\end{bmatrix} - \left(\frac{\mu^2}{2\sigma^2} - \log\sigma\right)\right]$$

- $h(x) = \frac{1}{\sqrt{2\pi}}$
- $\theta = \begin{bmatrix}\frac{\mu}{\sigma^2}\\ -\frac{1}{2\sigma^2}\end{bmatrix} = \begin{bmatrix}\theta_1\\ \theta_2\end{bmatrix}$, and $\begin{bmatrix}\mu\\ \sigma^2\end{bmatrix} = \begin{bmatrix}-\frac{\theta_1}{2\theta_2}\\ -\frac{1}{2\theta_2}\end{bmatrix}$
- $\phi(x) = \begin{bmatrix}x\\ x^2\end{bmatrix}$
- $A(\theta) = \frac{\mu^2}{2\sigma^2} + \log\sigma = \frac{-\theta_1^2}{4\theta_2} - \frac{1}{2}\log(-2\theta_2) - \frac{1}{2}\log(2\pi)$

Other Examples

- Many other distribution belong to the exponential family
  - Bernoulli
  - Beta
  - Gamma
  - Multinoulli/Multinomial
  - Dirichlet
  - Multivariate Gaussian
  - .. and many more (`https://en.wikipedia.org/wiki/Exponential_family`)
- Note: Not all distributions belong to the exponential family, e.g.,
  - Uniform distribution ($x \sim \text{Unif}(a, b)$)
  - Student-t distribution
  - Mixture distributions (e.g., mixture of Gaussians)
- If the support of the distribution depends on its parameters, then it is not an exp. family dist.

Log-Partition Function

- $A(\theta) = \log Z(\theta) = \log \int h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x})] d\boldsymbol{x}$ is the log-partition function
- $A(\theta)$ is also called the cumulant function
- Derivatives of $A(\theta)$ can be used to generate the cumulants of the sufficient statistics $\phi(\boldsymbol{x})$
- Exercise: Assume θ to be a scalar (thus $\phi(\boldsymbol{x})$ is also scalar). Show that the first and the second derivatives of $A(\theta)$ are

$$\begin{aligned} \frac{dA}{d\theta} &= \mathbb{E}_{p(\boldsymbol{x}|\theta)}[\phi(\boldsymbol{x})] = \text{mean}[\phi(\boldsymbol{x})] \\ \frac{d^2A}{d\theta^2} &= \mathbb{E}_{p(\boldsymbol{x}|\theta)}[\phi^2(\boldsymbol{x})] - \left[\mathbb{E}_{p(\boldsymbol{x}|\theta)}[\phi(\boldsymbol{x})]\right]^2 = \text{var}[\phi(\boldsymbol{x})] \end{aligned}$$

- Note: The above result also holds when θ and $\phi(\boldsymbol{x})$ are vector-valued (the "var" will be "covar")
- Important: $A(\theta)$ is a **convex function** of θ. Why?
- Exercise: For Binomial, using its expression of $A(\theta)$, derive the first and second cumulants of $\phi(x)$

MLE for Exponential Family Distributions

- Suppose we have data $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ drawn i.i.d. from an exponential family distribution

$$p(\boldsymbol{x}|\theta) = h(\boldsymbol{x}) \exp\left[\theta^\top \phi(\boldsymbol{x}) - A(\theta)\right]$$

- To do MLE, we need the overall likelihood. This is simply a product of the individual likelihoods

$$p(\mathcal{D}|\theta) = \prod_{i=1}^{N} p(\boldsymbol{x}_i|\theta) = \left[\prod_{i=1}^{N} h(\boldsymbol{x}_i)\right] \exp\left[\theta^\top \sum_{i=1}^{N} \phi(\boldsymbol{x}_i) - NA(\theta)\right] = \left[\prod_{i=1}^{N} h(\boldsymbol{x}_i)\right] \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right]$$

- To estimate θ via MLE (and even MAP), we only need $\phi(\mathcal{D}) = \sum_{i=1}^{N} \phi(\boldsymbol{x}_i)$ and N
- Size of $\phi(\mathcal{D}) = \sum_{i=1}^{N} \phi(\boldsymbol{x}_i)$ does not grow with N (same as the size of each $\phi(\boldsymbol{x}_i)$)
- Only exponential family distributions have finite-sized sufficient statistics
  - No need to store all the data; can simply store and recursively update the sufficient statistics with more and more data
  - Very useful when doing probabilistic/Bayesian inference with large-scale data sets. Also useful in online parameter estimation problems.

Bayesian Inference for Exponential Family Distributions

- We saw that the total likelihood given N i.i.d. observations $\mathcal{D}\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$

$$p(\mathcal{D}|\theta) \propto \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right] \qquad \text{where} \quad \phi(\mathcal{D}) = \sum_{i=1}^{N} \phi(\boldsymbol{x}_i)$$

- Let's choose the following prior (note: it looks similar in terms of θ within the exponent)

$$\boxed{p(\theta|\nu_0, \boldsymbol{\tau}_0) = h(\theta) \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta) - A_c(\nu_0, \boldsymbol{\tau}_0)\right]}$$

- Ignoring the prior's log-partition function $A_c(\nu_0, \boldsymbol{\tau}_0) = \log \int_\theta h(\theta) \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta)\right] d\theta$

$$\boxed{p(\theta|\nu_0, \boldsymbol{\tau}_0) \propto h(\theta) \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta)\right]}$$

- Comparing the prior's form with the likelihood, we notice that
    - ν_0 is like the <u>number of "pseudo-observations"</u> coming from the prior
    - τ_0 is the <u>total sufficient statistics</u> of these ν_0 pseudo-observations

The Posterior Distribution

- As we saw, the likelihood is

$$p(\mathcal{D}|\theta) \propto \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right] \qquad \text{where} \quad \phi(\mathcal{D}) = \sum_{i=1}^{N} \phi(\boldsymbol{x}_i)$$

- And the prior we chose is

$$p(\theta|\nu_0, \boldsymbol{\tau}_0) \propto h(\theta) \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta)\right]$$

- For this form of the prior, the posterior $p(\theta|\mathcal{D}) \propto p(\theta)p(\mathcal{D}|\theta)$ will be

$$\boxed{p(\theta|\mathcal{D}) \propto h(\theta) \exp\left[\theta^\top (\tau_0 + \phi(\mathcal{D})) - (\nu_0 + N)A(\theta)\right]}$$

- Note that the posterior has the same form as the prior; such a prior is called a **conjugate prior** (note: all exponential family distributions have a conjugate prior having a form shown as above)
- Thus posterior hyperparams ${\nu_0}', {\tau_0}'$ are obtained by simply adding "stuff" to prior's hyperparams

$${\nu_0}' \leftarrow \nu_0 + N \qquad \text{(no. of pseudo-obs + no. of actual obs)}$$

$${\tau_0}' \leftarrow \tau_0 + \phi(\mathcal{D}) \qquad \text{(total suff-stats from pseudo-obs + total suff-stats from actual obs)}$$

- Note: Prior's log-partition function $A_c(\nu_0, \boldsymbol{\tau}_0)$ updates to posterior's: $A_c(\nu_0 + N, \boldsymbol{\tau}_0 + \phi(\mathcal{D}))$

The Posterior Distribution

- Assuming the prior $p(\theta|\nu_0, \boldsymbol{\tau}_0) \propto h(\theta)\exp\left[\theta^\top \tau_0 - \nu_0 A(\theta)\right]$, the posterior was

$$p(\theta|\mathcal{D}) \propto h(\theta)\exp\left[\theta^\top(\tau_0 + \phi(\mathcal{D})) - (\nu_0 + N)A(\theta)\right]$$

- Assuming $\tau_0 = \nu_0\bar{\tau}_0$, we can also write the prior as $p(\theta|\nu_0, \bar{\tau}_0) \propto \exp\left[\theta^\top \nu_0\bar{\tau}_0 - \nu_0 A(\theta)\right]$
- Can think of $\bar{\tau}_0 = \tau_0/\nu_0$ as the average sufficient statistics per pseudo-observation
- The posterior can be written as

$$p(\theta|\mathcal{D}) \propto h(\theta)\exp\left[\theta^\top(\nu_0 + N)\frac{\nu_0\bar{\tau}_0 + \phi(\mathcal{D})}{\nu_0 + N} - (\nu_0 + N)A(\theta)\right]$$

- Denoting $\bar{\phi} = \frac{\phi(D)}{N}$ as the average suff-stats per real observation, the posterior updates are

$$\begin{aligned} {\nu_0}' &\leftarrow \nu_0 + N \\ \bar{\tau}_0' &\leftarrow \frac{\nu_0\bar{\tau}_0 + N\bar{\phi}}{\nu_0 + N} \end{aligned}$$

- Note that the posterior hyperparam $\bar{\tau}_0'$ is a convex combination of the average suff-stats $\bar{\tau}_0$ of the ν_0 pseudo-observations and the average suff-stats $\bar{\phi}$ of the N actual observations

Posterior Predictive Distribution

- Assume some past (training) data $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ generated from an exp. family distribution
- Assme some test data $\mathcal{D}' = \{\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_{N'}\}$ from the same distribution ($N' \geq 1$)
- The posterior predictive distribution of $\mathcal{D}'$ (probability distribution of new data given old data)

$$p(\mathcal{D}'|\mathcal{D}) = \int p(\mathcal{D}'|\theta)p(\theta|\mathcal{D})d\theta$$

- We've already seen some specific examples of computing the posterior predictive dist., e.g.,
  - Beta-Bernoulli case: Posterior predictive distribution of next coin toss
  - Bayesian linear regression: Posterior predictive distribution of the response y_* of test input $\boldsymbol{x}_*$
- Nice Property: If the likelihood is an exponential family distribution, prior is conjugate (and thus is the posterior), the posterior predictive always has a closed form expression (shown next)

Posterior Predictive Distribution

- Recall the form of the likelihood $p(\mathcal{D}|\theta)$ for exp. family dist.

$$p(\mathcal{D}|\theta) = \left[\prod_{i=1}^{N} h(\boldsymbol{x}_i)\right] \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right]$$

- The conjugate prior was

$$p(\theta|\nu_0, \boldsymbol{\tau}_0) = h(\theta) \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta) - A_c(\nu_0, \boldsymbol{\tau}_0)\right]$$

- For this choice of the conjugate prior, the posterior was shown to be

$$p(\theta|\mathcal{D}) = h(\theta) \exp\left[\theta^\top (\tau_0 + \phi(\mathcal{D})) - (\nu_0 + N)A(\theta) - A_c(\nu_0 + N, \boldsymbol{\tau}_0 + \phi(\mathcal{D}))\right]$$

- For the test data $\mathcal{D}'$, the likelihood will be

$$p(\mathcal{D}'|\theta) = \left[\prod_{i=1}^{N'} h(\tilde{\boldsymbol{x}}_i)\right] \exp\left[\theta^\top \phi(\mathcal{D}') - N'A(\theta)\right] \qquad \text{where} \quad \phi(\mathcal{D}') = \sum_{i=1}^{N'} \phi(\tilde{\boldsymbol{x}}_i)$$

Posterior Predictive Distribution

- Therefore the posterior predictive distribution will be

$$
\begin{aligned}
p(\mathcal{D}'|\mathcal{D}) &= \int p(\mathcal{D}'|\theta)p(\theta|\mathcal{D})d\theta \\
&= \int \underbrace{\left[\prod_{i=1}^{N'} h(\tilde{\mathbf{x}}_i)\right]}_{\text{constant w.r.t. } \theta} \exp\left[\theta^\top \phi(\mathcal{D}') - N'A(\theta)\right] h(\theta) \exp\left[\theta^\top(\tau_0 + \phi(\mathcal{D})) - (\nu_0 + N)A(\theta) - \underbrace{A_c(\nu_0 + N, \boldsymbol{\tau}_0 + \phi(\mathcal{D}))}_{\text{constant w.r.t. } \theta}\right] d\theta
\end{aligned}
$$

- The above gets simplified further into

$$
\begin{aligned}
p(\mathcal{D}'|\mathcal{D}) &= \left[\prod_{i=1}^{N'} h(\tilde{\mathbf{x}}_i)\right] \frac{\int h(\theta)\exp\left[\theta^\top(\tau_0 + \phi(\mathcal{D}) + \phi(\mathcal{D}')) - (\nu_0 + N + N')A(\theta)\right] d\theta}{\exp\left[A_c(\nu_0 + N, \boldsymbol{\tau}_0 + \phi(\mathcal{D}))\right]} \\
&= \left[\prod_{i=1}^{N'} h(\tilde{\mathbf{x}}_i)\right] \frac{Z_c(\nu_0 + N + N', \boldsymbol{\tau}_0 + \phi(\mathcal{D}) + \phi(\mathcal{D}'))}{\exp\left[A_c(\nu_0 + N, \boldsymbol{\tau}_0 + \phi(\mathcal{D}))\right]}
\end{aligned}
$$

where $Z_c(\nu_0 + N + N', \boldsymbol{\tau}_0 + \phi(\mathcal{D}) + \phi(\mathcal{D}')) = \int h(\theta)\exp\left[\theta^\top(\tau_0 + \phi(\mathcal{D}) + \phi(\mathcal{D}')) - (\nu_0 + N + N')A(\theta)\right] d\theta$

Posterior Predictive Distribution

- Since $A_c = \log Z_c$ or $Z_c = \exp(A_c)$, we can write the posterior predictive distribution as

$$
\begin{aligned}
p(\mathcal{D}'|\mathcal{D}) &= \left[\prod_{i=1}^{N'} h(\tilde{\mathbf{x}}_i)\right] \frac{Z_c(\nu_0 + N + N', \boldsymbol{\tau}_0 + \phi(\mathcal{D}) + \phi(\mathcal{D}'))}{Z_c(\nu_0 + N, \boldsymbol{\tau}_0 + \phi(\mathcal{D}))} \\
&= \left[\prod_{i=1}^{N'} h(\tilde{\mathbf{x}}_i)\right] \exp\left[A_c(\nu_0 + N + N', \boldsymbol{\tau}_0 + \phi(\mathcal{D}) + \phi(\mathcal{D}')) - A_c(\nu_0 + N, \boldsymbol{\tau}_0 + \phi(\mathcal{D}))\right]
\end{aligned}
$$

- Therefore the posterior predictive is proportional to ..
  - .. the ratio of two partition functions of two "posterior distributions" (one with $N + N'$ examples and the other with N examples)
  - .. or exponential of the difference of the corresponding log-partition functions
- Note that the form of Z_c (and A_c) will simply depend on the chosen conjugate prior
- Very useful result. Also holds for $N = 0$
  - In the $N = 0$ case, $p(\mathcal{D}') = \int p(\mathcal{D}'|\theta)p(\theta)d\theta$ is simply the marginal likelihood of $\mathcal{D}'$

Summary

- Exp. family distributions are very useful for modeling diverse types of data/parameters
- Conjugate priors to exp. family distributions make parameter updates very simple
- Other quantities such as posterior predictive can be computed in closed form
- Useful in designing generative classification models. Choosing class-conditional from exponential family with conjugate priors helps in parameter estimation
- Useful in designing generative models for unsupervised learning
- Uses in designing Generalized Linear Models (GLM): Model $p(y|\boldsymbol{x})$ using exp. family distribution
  - Linear regression (with Gaussian likelihood) and logistic regression are GLMs
- We will see several use cases when we discuss approximate inference algoritms (e.g., Gibbs sampling, and especially variational inference)