Logistic/Softmax Regression, GLM, and Exponential Family

CS772A: Probabilistic Machine Learning

Piyush Rai

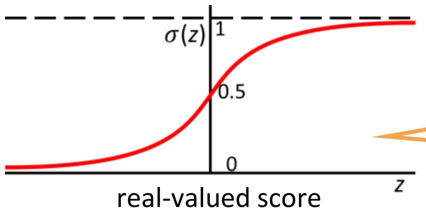
Logistic Regression

There are other ways too that can convert the score into a probability, such as a CDF: $p(y=1|\boldsymbol{x},\boldsymbol{w}) = \mu = \Phi(\boldsymbol{w}^\top\boldsymbol{x})$ where Φ is the CDF of $\mathcal{N}(0,1)$. This model is known as "Probit Regression".

- A discriminative model for binary classification ($y \in \{0,1\}$)
- A linear model with parameters $\boldsymbol{w} \in \mathbb{R}^D$ computes a score $\boldsymbol{w}^\top\boldsymbol{x}$ for input $\boldsymbol{x}$
- A sigmoid function maps this real-valued score into probability of label being 1

Also used as a nonlinear "activation function" in deep neural networks

$$p(y=1|\boldsymbol{x},\boldsymbol{w}) = \mu = \sigma(\boldsymbol{w}^\top\boldsymbol{x})$$



$$\sigma(z) = \frac{1}{1+\exp(-z)} = \frac{\exp(z)}{1+\exp(z)}$$

Large positive score $\boldsymbol{w}^\top\boldsymbol{x}$ means large prob of label being 1, and large negative score means low prob

- Thus conditional distribution of label $y \in \{0,1\}$ given $\boldsymbol{x}$ is the following Bernoulli

Likelihood

$$p(y|\boldsymbol{x},\boldsymbol{w}) = \text{Bernoulli}[y|\mu] = \mu^y(1-\mu)^{1-y} = \left[\frac{\exp(\boldsymbol{w}^\top\boldsymbol{x})}{1+\exp(\boldsymbol{w}^\top\boldsymbol{x})}\right]^y \left[\frac{1}{1+\exp(\boldsymbol{w}^\top\boldsymbol{x})}\right]^{1-y}$$

- NLL is the binary cross-entropy loss: $-[y_n \log \mu_n + (1-y_n)\log(1-\mu_n)]$
- NLL is convex in $\boldsymbol{w}$. Can also use a prior $p(\boldsymbol{w}|\lambda) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{0},\lambda^{-1}\boldsymbol{I})$ if interested in MAP or full posterior on $\boldsymbol{w}$

Logistic Regression: MAP and Posterior

- The posterior will be

Gaussian Bernoulli

$$p(\boldsymbol{w}|\boldsymbol{X},\boldsymbol{y}) = \frac{p(\boldsymbol{w})p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w})}{p(\boldsymbol{y}|\boldsymbol{X})} = \frac{\overbrace{p(\boldsymbol{w})}^{\text{Gaussian}}\prod_{n=1}^{N}\overbrace{p(y_n|\boldsymbol{w},\boldsymbol{x}_n)}^{\text{Bernoulli}}}{\int p(\boldsymbol{w})\prod_{n=1}^{N}p(y_n|\boldsymbol{w},\boldsymbol{x}_n)\,d\boldsymbol{w}}$$

- MAP estimation is easy. $-\log p(\boldsymbol{w}|\boldsymbol{X},\boldsymbol{y})$ is convex for LR. Unique minima
  - Can use first or second order optimization with gradient and Hessian being

$$\mathbf{g} = -\sum_{n=1}^{N}(y_n-\mu_n)\mathbf{x}_n + \lambda\mathbf{I}\mathbf{w} = \mathbf{X}^\top(\boldsymbol{\mu}-\mathbf{y})+\lambda\mathbf{w} \qquad \text{(a } D\times 1 \text{ vector)}$$

$$\mathbf{H} = \sum_{n=1}^{N}\mu_n(1-\mu_n)\mathbf{x}_n\mathbf{x}_n^\top + \lambda\mathbf{I} = \mathbf{X}^\top\mathbf{S}\mathbf{X}+\lambda\mathbf{I} \qquad \text{(a } D\times D \text{ matrix)}$$

$$\mu_n = \sigma(\boldsymbol{w}^\top\boldsymbol{x}_n)$$

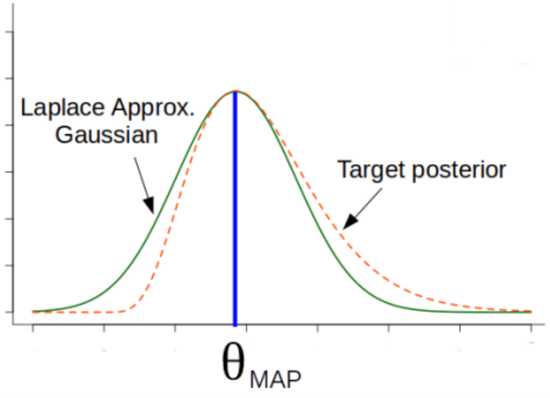
- Full posterior is intractable because of non-conjugacy
  - A popular option is to use the Laplace's approximation (other methods like MCMC and variational inference can also be used; will see them later)

Laplace's (or Gaussian) Approximation

- Consider a posterior distribution that is intractable to compute

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D},\theta)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}$$

- Laplace approximation approximates the above using a Gaussian distribution



$$p(\theta|\mathcal{D}) \approx \mathcal{N}(\theta|\theta_{MAP}, \mathbf{\Lambda}^{-1})$$

Tells us about the space (curvature) of the true posterior around θ_{MAP}

Related to the Fisher Information Matrix (FIM); will see shortly

$$\theta_{MAP} = \operatorname{argmax}_\theta \log p(\theta|\mathcal{D})$$

Negative of the Hessian, i.e., the second derivative of the log joint, at θ_{MAP}

$$\mathbf{\Lambda} = -\nabla_\theta^2 \log p(\theta|\mathcal{D})\Big|_{\theta=\theta_{MAP}} = -\nabla_\theta^2 \log p(\mathcal{D},\theta)\Big|_{\theta=\theta_{MAP}}$$

- Laplace's approx. is based on a second-order Taylor approx. of the posterior (will see the proof and details later)

LR: Posterior Predictive Distribution

- The posterior predictive distribution can be computed as

$$p(y_* = 1|\boldsymbol{x}_*, \boldsymbol{X}, \boldsymbol{y}) = \int p(y_* = 1|\boldsymbol{w}, \boldsymbol{x}_*) p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y}) d\boldsymbol{w}$$

Integral not tractable and must be approximated

sigmoid

Gaussian (if using Laplace approx.)

- Monte-Carlo approximation of this integral is one possible way
  - Draw M samples $\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_M$, from the approx. of posterior
  - Approximate the PPD as follows

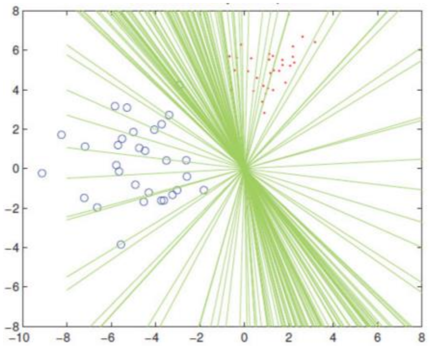
$$p(y_* = 1|\boldsymbol{x}_*, \boldsymbol{X}, \boldsymbol{y}) \approx \frac{1}{M}\sum_{m=1}^{M} p(y_* = 1|\boldsymbol{w}_m, \boldsymbol{x}_*) = \frac{1}{M}\sum_{m=1}^{M} \sigma(\boldsymbol{w}_m^\top \boldsymbol{x}_n)$$

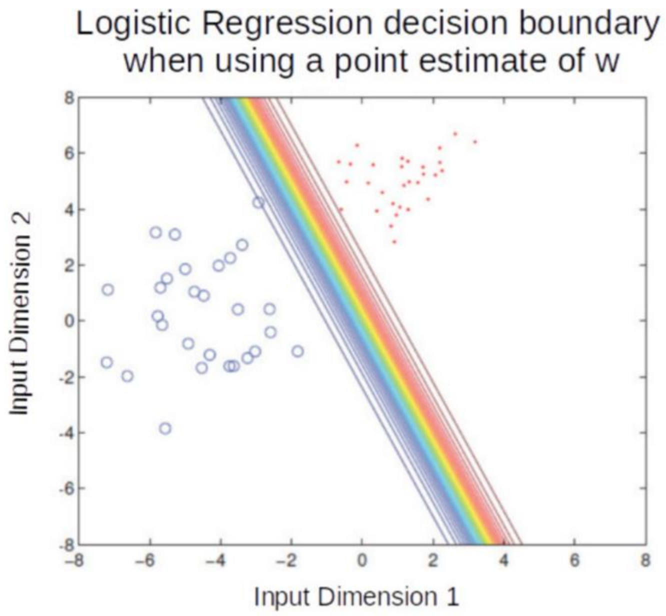
- In contrast, when using MLE/MAP solution $\hat{\boldsymbol{w}}_{opt}$, the plug-in pred. distribution

$$\begin{aligned} p(y_* = 1|\boldsymbol{x}_*, \boldsymbol{X}, \boldsymbol{y}) &= \int p(y_* = 1|\boldsymbol{w}, \boldsymbol{x}_*) p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y}) d\boldsymbol{w} \\ &\approx p(y_* = 1|\hat{\boldsymbol{w}}_{opt}, \boldsymbol{x}_*) = \sigma(\hat{\boldsymbol{w}}_{opt}{}^\top \boldsymbol{x}_n) \end{aligned}$$

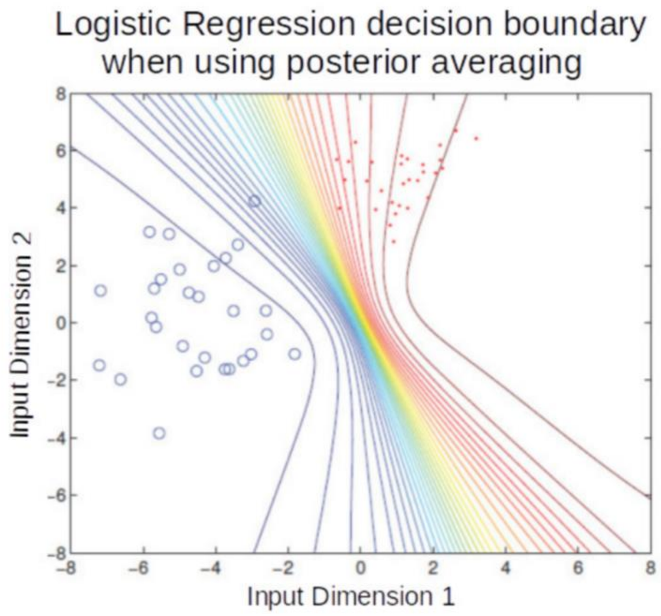
LR: Plug-in Prediction vs Bayesian Averaging

- Plug-in prediction uses a single $\boldsymbol{w}$ (point est) to make prediction
- PPD does an averaging using all possible $\boldsymbol{w}$'s from the posterior





Color transitions (red to blue) in both plots denote how the probability of an input changes from belonging to red class to belonging to blue class. All inputs on a line (or curve on RHS plot) have the same probability of belonging to the red/blue class



Posterior averaging is like using an ensemble of models. In this example, each model is a linear classifier but the ensemble-like effect resulted in nonlinear boundaries

$$p(y_* = 1|\boldsymbol{x}_*, \boldsymbol{X}, \boldsymbol{y}) \approx \sigma(\widehat{\boldsymbol{w}}_{opt}{}^\top \boldsymbol{x}_n)$$

$$p(y_* = 1|\boldsymbol{x}_*, \boldsymbol{X}, \boldsymbol{y}) \approx \frac{1}{M}\sum_{m=1}^{M} \sigma(\boldsymbol{w}_m^\top \boldsymbol{x}_n)$$

Multiclass Logistic (a.k.a. Softmax) Regression

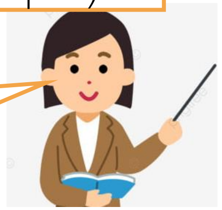
- Also called multinoulli/multinomial regression: Basically, LR for $K > 2$ classes
- In this case, $y_n \in \{1,2,\ldots,K\}$ and label probabilities are defined as

Softmax function

Real-valued scores $\boldsymbol{w}_k^\top \boldsymbol{x}_n$ are also known as "logits" (thus K logits for each input)

$$p(y_n = k|\boldsymbol{x}_n, \boldsymbol{W}) = \frac{\exp(\boldsymbol{w}_k^\top \boldsymbol{x}_n)}{\sum_{\ell=1}^K \exp(\boldsymbol{w}_\ell^\top \boldsymbol{x}_n)} = \mu_{nk}$$

Also note that $\sum_{\ell=1}^K \mu_{n\ell} = 1$ for any input $\boldsymbol{x}_n$



- K weight vecs $\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_K$ (one per class), each D-dim, and $\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_K]$
- Each likelihood $p(y_n|\boldsymbol{x}_n, \boldsymbol{W})$ is a multinoulli distribution. Therefore total likelihood

$$p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{W}) = \prod_{n=1}^N \prod_{\ell=1}^K \mu_{n\ell}^{y_{n\ell}}$$

Notation: $y_{n\ell} = 1$ if true class of $\boldsymbol{x}_n$ is ℓ and $y_{n\ell'} = 0 \ \forall \ \ell' \neq \ell$

- Can do MLE/MAP/fully Bayesian estimation for $\boldsymbol{W}$ similar to LR model

Generalized Linear Models

- (Probabilistic) Linear Regression: when response $\boldsymbol{y}$ is real-valued

$$p(y|\boldsymbol{x},\boldsymbol{w}) = \mathcal{N}(y|\boldsymbol{w}^\top\boldsymbol{x},\beta^{-1})$$

- Logistic Regression: when response $\boldsymbol{y}$ is binary (0/1)

$$p(y|\boldsymbol{x},\boldsymbol{w}) = \text{Bernoulli}[y|\sigma(\boldsymbol{w}^\top\boldsymbol{x})] = \left[\frac{\exp(\boldsymbol{w}^\top\boldsymbol{x})}{1+\exp(\boldsymbol{w}^\top\boldsymbol{x})}\right]^y \left[\frac{1}{1+\exp(\boldsymbol{w}^\top\boldsymbol{x})}\right]^{1-y}$$

- Both are examples of a Generalized Linear Model (GLM)
  - The model depends on the inputs $\boldsymbol{x}$ via a linear model $\boldsymbol{w}^\top\boldsymbol{x}$
- GLM is defined using an exponential family distribution

$$p(y|\boldsymbol{x},\boldsymbol{w}) = \text{ExpFam}[y|f(\boldsymbol{w}^\top\boldsymbol{x})]$$

MLE/MAP of $\boldsymbol{w}$ is easy for GLMs (due to convex objective, thanks to exp-family). Posterior usually requires approximations if likelihood and prior are not conjugate pairs (Laplace approximation or other methods used)

- ExpFam can be any suitable distribution depending on the nature of outputs, e.g.,
  - Gaussian for reals, Bernoulli for binary, Poisson for Count, gamma for positive reals
- ExpFam distributions are more generally useful in other contexts as well

Exp. Family (Pitman, Darmois, Koopman, 1930s)

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

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

- $\boldsymbol{x} \in \mathcal{X}^m$ is the r.v. being modeled ($\mathcal{X}$ denotes some space, e.g., $\mathbb{R}$ or $\{0,1\}$)
- $\theta \in \mathbb{R}^d$: Natural parameters or canonical parameters defining the distribution
- $\boldsymbol{\phi}(\boldsymbol{x}) \in \mathbb{R}^d$: Sufficient statistics (another random variable)
  - Knowing this quantity suffices to estimate parameter $\boldsymbol{\theta}$ from $\boldsymbol{x}$
- $Z(\theta) = \int h(\boldsymbol{x}) \exp[\theta^\top \boldsymbol{\phi}(\boldsymbol{x})] d\boldsymbol{x}$: Partition Function
- $A(\theta) = \log Z(\theta)$: Log-partition function (also called cumulant function)
- $\boldsymbol{h}(\boldsymbol{x})$: A constant (doesn't depend on $\boldsymbol{\theta}$)

Expressing a Distribution in Exp. Family Form

- Recall the form of exp-fam distribution $p(x|\theta) = h(x)\exp[\theta^\top \phi(x) - A(\theta)]$
- To write any exp-fam dist $p()$ in the above form, write it as $\exp(\log p())$

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

- Now compare the resulting expression with the exponential family form

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

.. 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 PDF of a univar Gaussian (already has exp, so less work needed :))

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

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

$$\theta = \begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix} = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} \qquad \phi(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix} \quad \text{, and } \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix} = \begin{bmatrix} -\frac{\theta_1}{2\theta_2} \\ -\frac{1}{2\theta_2} \end{bmatrix}$$

$$h(x) = \frac{1}{\sqrt{2\pi}} \qquad A(\theta) = \frac{\mu^2}{2\sigma^2} + \log\sigma = \frac{-\theta_1^2}{4\theta_2} - \frac{1}{2}\log(-2\theta_2) - \frac{1}{2}\log(2\pi)$$

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)

Log-Partition Function

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

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

- Above result also holds when $\boldsymbol{\theta}$ and $\boldsymbol{\phi}(x)$ are vector-valued (the “var” will be “covar”)
- Important: $A(\boldsymbol{\theta})$ is a convex function of $\boldsymbol{\theta}$. Why?

MLE for Exponential Family Distributions

- Assume data $\mathcal{D} = \{x_1, \ldots, x_N\}$ drawn i.i.d. from an exp. family distribution

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

- To do MLE, we need the overall likelihood -- a product of the individual likelihoods

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

- To estimate $\boldsymbol{\theta}$ (as we'll see shortly), we only need $\phi(\mathcal{D}) = \sum_{i=1}^N \phi(\mathbf{x}_i)$ and N
- Size of $\phi(\mathcal{D}) = \sum_{i=1}^N \phi(x_i)$ does not grow with N (same as the size of each $\boldsymbol{\phi}(\boldsymbol{x}_i)$)
- Only exponential family distributions have finite-sized sufficient statistics
  - No need to store all the data; can simply update the sufficient statistics as data comes
  - Useful in probabilistic inference with large-scale data sets and "online" parameter estimation

Bayesian Inference for Expon. Family Distributions

- Already saw that the total likelihood given N i.i.d. observations $\mathcal{D} = \{x_1, \ldots, x_N\}$

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

- Let's choose the following prior (note: looks similar in terms of $\boldsymbol{\theta}$ within exp)

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

- 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, note that
  - ν_0 is like the <u>number of "pseudo-observations"</u> coming from the prior
  - τ_0 is the <u>total sufficient statistics of the pseudo-observations</u> (τ_0 / ν_0 per pseudo-obs)

The Posterior

- The likelihood and prior were

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

Assume its log partition function denoted as $A_c(\nu_0, \tau_0)$

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

- The posterior $p(\theta|\mathcal{D}) \propto p(\theta)p(\mathcal{D}|\theta)$ therefore will be

Posterior is also from the same family as the prior

Happens when the prior is conjugate to the likelihood

Its log partition function will be $A_c(\nu_0 + N, \tau_0 + \phi(\mathcal{D}))$

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

- Every exp family likelihood has a conjugate prior having the form above
- Posterior's hyperparams $\boldsymbol{\tau}_0', \boldsymbol{\nu}_0'$ obtained by adding "stuff" to prior's hyperparams

Number of pseudo-observations plus number of actual observations

$$\nu_0' \leftarrow \nu_0 + N$$

Suff-stats of pseudo-obervations plus suff-stats of actual observations

$$\tau_0' \leftarrow \tau_0 + \phi(\mathcal{D})$$

Another equivalent form

$\bar{\tau}_0 = \tau_0/\nu_0$

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

$$\nu_0' \leftarrow \nu_0 + N$$

$$\bar{\tau}_0' \leftarrow \frac{\nu_0\bar{\tau}_0 + N\bar{\phi}}{\nu_0 + N}$$

$\bar{\phi} = \frac{\phi(D)}{N}$

Convex comb of avg suff-stats of pseudo obs and actual obs

Posterior Predictive Distribution

- Assume some training data $\mathcal{D} = \{x_1, \ldots, x_N\}$ from some exp-fam distribution
- Assume some test data $\mathcal{D}' = \{\tilde{x}_1, \ldots, \tilde{x}_{N'}\}$ from the same distribution
- The posterior pred. distr. of $\mathcal{D}'$

Exp. Fam. likelihood w.r.t. test data

Posterior (same form as the prior due to conjugacy)

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

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

- This gets further simplified into

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

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

Posterior Predictive Distribution

Thus PPD as well as marginal likelihood has closed form expression when working with exp-family distributions

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

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

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

- Therefore the posterior predictive is proportional to
  - Ratio of two partition functions of two "posterior distributions" (one with $N + N'$ examples and the other with N examples)
  - 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 this case $p(\mathcal{D}') = \int p(\mathcal{D}'|\theta)p(\theta)d\theta$ is simply the marginal likelihood of test data $\mathcal{D}'$

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
- Used in designing Generalized Linear Models: Model $p(y|x)$ using exp. fam distribution
  - Linear regression (with Gaussian likelihood) and logistic regression are GLMs
- Will see several use cases when we discuss approx inference algorithms (e.g., Gibbs sampling, and especially variational inference)