Exponential Family Distributions and Conditional Models

Piyush Rai

Topics in Probabilistic Modeling and Inference (CS698X)

Jan 16, 2019

Plan for today

- Exponential family distributions (a very important class of distributions)

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

- Conditional models and parameter estimation for them (our example: Prob. Linear Regression)

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

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": $p(\boldsymbol{x}|\theta)$ <u>as a function of θ</u> 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 <u>cumulant function</u>)
- $h(\boldsymbol{x})$: A constant (doesn't depend on θ)

Expressing a Distribution in Exponential Family Form

- Recall the form of exp-fam distribution: $h(\boldsymbol{x})\exp[\theta^\top\phi(\boldsymbol{x}) - A(\theta)]$
- To write any exp-fam dist $p()$ in the above form, write it as $\exp(\log p())$, e.g., for Binomial

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

- Now compare the resulting expression with the exponential family form

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

.. 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 (already has exp in it, so less work :))

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

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

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

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

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

- 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?

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 (as we'll see shortly), 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.

MLE and Moment Matching

- The likelihood is of the form $p(\mathcal{D}|\theta) = \left[\prod_{i=1}^{N} h(\boldsymbol{x}_i)\right] \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right]$
- The log-likelihood is (ignoring constant w.r.t. θ): $\log p(\mathcal{D}|\theta) = \theta^\top \phi(\mathcal{D}) - NA(\theta)$
- Note: This is concave in θ (since $-A(\theta)$ is concave). Maximization will yield a global maxima of θ
- MLE for exp-fam distributions can also be seen as doing moment-matching. To see this, note that

$$\nabla_\theta \left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right] = \phi(\mathcal{D}) - N\nabla_\theta[A(\theta)] = \phi(\mathcal{D}) - N\mathbb{E}_{p(\boldsymbol{x}|\theta)}[\phi(\boldsymbol{x})] = \sum_{i=1}^{N} \phi(\boldsymbol{x}_i) - N\mathbb{E}_{p(\boldsymbol{x}|\theta)}[\phi(\boldsymbol{x})]$$

- Therefore, at the "optimal" (i.e., MLE) $\hat{\theta}$, where the derivative is 0, the following must hold

$$\boxed{\mathbb{E}_{p(\boldsymbol{x}|\theta)}[\phi(\boldsymbol{x})] = \frac{1}{N}\sum_{i=1}^{N} \phi(\boldsymbol{x}_i)}$$

- This is basically matching the expected moments of the distribution with empirical moments ("empirical" here means what we compute using the observed data)

Moment Matching: An Example

- Given N observations $x_1, \ldots, x_N$ from a univariate Gaussian $\mathcal{N}(x|\mu, \sigma^2)$, doing moment-matching

$$\mathbb{E}[\phi(x)] = \frac{1}{N}\sum_{i=1}^{N}\phi(x_i)$$

- The "true", i.e., expected moments: $\mathbb{E}[\phi(x)] = \mathbb{E}\begin{bmatrix} x \\ x^2 \end{bmatrix}$. Therefore

$$\mathbb{E}\begin{bmatrix} x \\ x^2 \end{bmatrix} = \begin{bmatrix} \frac{1}{N}\sum_{i=1}^{N} x_i \\ \frac{1}{N}\sum_{i=1}^{N} x_i^2 \end{bmatrix}$$

- For a univariate Gaussian, note that $\mathbb{E}[x] = \mu$ and $\mathbb{E}[x^2] = \text{var}[x] + \mathbb{E}[x]^2 = \sigma^2 + \mu^2$
- Thus we have two equations and two unknowns
- From the first equation, we immediately get $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$
- From the second equation, we get $\sigma^2 = \mathbb{E}[x^2] - \mu^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$

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 number of "pseudo-observations" coming from the prior
  - τ_0 is the total sufficient statistics 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

$$\begin{aligned} {\nu_0}' &\leftarrow \nu_0 + N && \text{(no. of pseudo-obs + no. of actual obs)} \\ {\tau_0}' &\leftarrow \tau_0 + \phi(\mathcal{D}) && \text{(total suff-stats from pseudo-obs + total suff-stats from actual obs)} \end{aligned}$$

- 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
    - Dirichlet-Multinoulli case: Posterior predictive distribution of next dice roll
    - Gaussian-Gaussian, Gaussian-IG, Gaussian-Gamma, Gaussian-NIG, Gaussian-NG case: Posterior predictive distribution of the next observation
- 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)

Estimating Conditional Models, e.g., $p(y|\boldsymbol{x})$

Our Example: Probabilistic/Bayesian Linear Regression

Estimating Conditional Models

- Conditional models of the form $p(y|\boldsymbol{x})$ are commonly used in supervised learning problems
  - But more broadly applicable (basically any problem where data y depends on another quantity $\boldsymbol{x}$)
- Conditional models can be estimated using one of the following two ways
  1. Estimate the joint distribution $p(\boldsymbol{x}, y)$ and then use Bayes rule to get $p(y|\boldsymbol{x})$

$$p(y|\boldsymbol{x}, \theta) = \frac{p(\boldsymbol{x}, y|\theta)}{p(\boldsymbol{x}|\theta)}$$

  2. Estimate the conditional $p(y|\boldsymbol{x})$ directly (used when we don't care about modeling $\boldsymbol{x}$), e.g.

$$p(y|\boldsymbol{x}) = \mathcal{N}(y|f_\mu(\boldsymbol{x}), f_{\sigma^2}(\boldsymbol{x})) \qquad \text{(params of } p(y|\boldsymbol{x}) \text{ will be functions of } \boldsymbol{x})$$

- Approach 1 is called generative approach, approach 2 is called discriminative approach
- For pros/cons, refer to CS771 lecture slides and readings
- For now, we will focus on learning (2) using fully Bayesian inference
- Today's focus will be on regression problems (y is real-valued response for the input $\boldsymbol{x}$)

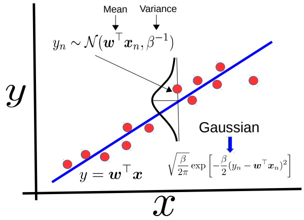
Linear Regression: A Probabilistic Setup

- Given: N training examples $\{\boldsymbol{x}_n, y_n\}_{n=1}^N$, features: $\boldsymbol{x}_n \in \mathbb{R}^D$, response $y_n \in \mathbb{R}$
- Assume a "noisy" linear model with regression weight vector $\boldsymbol{w} = [w_1, w_2, \ldots, w_D] \in \mathbb{R}^D$

$$y_n = \boldsymbol{w}^\top \boldsymbol{x}_n + \epsilon_n$$

  where $\epsilon_n \sim \mathcal{N}(0, \beta^{-1})$, β: precision (inverse variance) of Gaussian (assumed known)

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



- Note: Some books (e.g., PRML) use $\phi(\boldsymbol{x}_n)$ to denote the features where ϕ is some transformation of the original features $\boldsymbol{x}_n$ (we will only use this notation when talking about nonlinear regression)

The Likelihood Model

- Notation: $\mathbf{X} = [\boldsymbol{x}_1 \dots \boldsymbol{x}_N]^\top$: $N \times D$ feature matrix, $\boldsymbol{y} = [y_1 \dots y_N]^\top$: $N \times 1$ response vector
- Assuming independent observations, the likelihood model

$$
\begin{aligned}
p(\boldsymbol{y}|\boldsymbol{w}, \mathbf{X}, \beta) = \prod_{n=1}^{N} p(y_n|\boldsymbol{w}, \boldsymbol{x}_n, \beta) &= \prod_{n=1}^{N} \mathcal{N}(y_n|\boldsymbol{w}^\top \boldsymbol{x}_n, \beta^{-1}) \\
&= \prod_{n=1}^{N} \sqrt{\frac{\beta}{2\pi}} \exp\left[-\frac{\beta}{2}(y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2\right] \\
&= \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} \exp\left[-\frac{\beta}{2}\sum_{n=1}^{N}(y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2\right]
\end{aligned}
$$

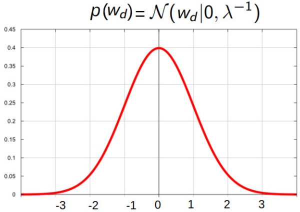
- Note that NLL = sum of squared errors! Minimizing w.r.t. $\boldsymbol{w}$ will give MLE/least squares solution!
- For brevity, can also write the likelihood $p(\boldsymbol{y}|\boldsymbol{w}, \mathbf{X})$ as an N-dim multivariate Gaussian

$$
p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w}, \beta) = \mathcal{N}(\boldsymbol{y}|\mathbf{X}\boldsymbol{w}, \beta^{-1}\mathbf{I}_N) = \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} \exp\left[-\frac{\beta}{2}(\boldsymbol{y} - \mathbf{X}\boldsymbol{w})^\top(\boldsymbol{y} - \mathbf{X}\boldsymbol{w})\right]
$$

The Prior

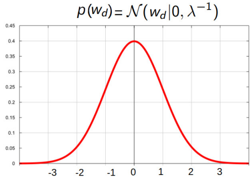
- Assume the entries in $\boldsymbol{w}$ are i.i.d. with zero mean Gaussian priors. Therefore

$$p(\boldsymbol{w}) = \prod_{d=1}^{D} p(w_d) = \prod_{d=1}^{D} \mathcal{N}(w_d|0,\lambda^{-1}) = \mathcal{N}(\boldsymbol{w}|\mathbf{0},\lambda^{-1}\mathbf{I}_D) = \left(\frac{\lambda}{2\pi}\right)^{\frac{D}{2}} \exp\left[-\frac{\lambda}{2}\boldsymbol{w}^\top \boldsymbol{w}\right]$$



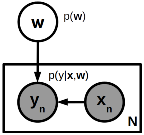
- This prior promotes the entries in $\boldsymbol{w}$ to be small (close to zero)
  - Also, the negative of log-prior is the same as an ℓ_2 regularizer on $\boldsymbol{w}$
- This prior is conjugate to the likelihood (Gaussian) which makes posterior inference easy

The Prior



- The role of the precision hyperparam λ in the prior is important
- Large values of λ would more aggressively encourage w_d to be close to zero
- Can think of λ as the regularization hyperparam for the weights
- Important: Can infer λ as well (will see later how to do this)
- Can even have different λ for each w_d, i.e., $p(\boldsymbol{w}|\{\lambda_d\}_{d=1}^D) = \prod_{d=1}^D \mathcal{N}(w_d|0, \lambda_d^{-1})$
  - Useful in sparse regression/classification models in which very few features are relevant which can be identified by inferring $\{\lambda_d\}_{d=1}^D$. Popularly known as sparse Bayesian learning (more on this later).

Inference Tasks for Bayesian Linear Regression



(Hyperparameters λ,β not shown as they are fixed/known)

- Want to infer the posterior distribution over $\boldsymbol{w}$ (for now, assume β and λ to be known)

$$p(\boldsymbol{w}|\boldsymbol{y}, \mathbf{X}, \beta, \lambda) = \frac{p(\boldsymbol{w}|\lambda)p(\boldsymbol{y}|\boldsymbol{w}, \mathbf{X}, \beta)}{p(\boldsymbol{y}|\mathbf{X}, \beta, \lambda)}$$

- Want to infer the posterior predictive distribution

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

- Likelihood $p(y|\boldsymbol{w}, \boldsymbol{x}, \beta)$ and prior $p(\boldsymbol{w}|\lambda)$ are Gaussians, so above computations are easy!
- Also note that it's also like a noisy linear Gaussian model: $\boldsymbol{y} = \mathbf{X}\boldsymbol{w} + \boldsymbol{\epsilon}$ with noise $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_N]$
  - $D \times 1$ Gaussian r.v. $\boldsymbol{w}$ transformed via $N \times D$ matrix $\mathbf{X}$ to produce $N \times 1$ vector $\boldsymbol{y}$

Bayesian Linear Regression: The Posterior

- The posterior over $\boldsymbol{w}$ (for now, assume hyperparams β and λ to be known)

$$p(\boldsymbol{w}|\boldsymbol{y},\mathbf{X},\beta,\lambda) = \frac{p(\boldsymbol{w}|\lambda)p(\boldsymbol{y}|\boldsymbol{w},\mathbf{X},\beta)}{p(\boldsymbol{y}|\mathbf{X},\beta,\lambda)} \propto p(\boldsymbol{w}|\lambda)p(\boldsymbol{y}|\boldsymbol{w},\mathbf{X},\beta)$$

- Computing $p(\boldsymbol{w}|\mathbf{X},\boldsymbol{y},\beta,\lambda)$

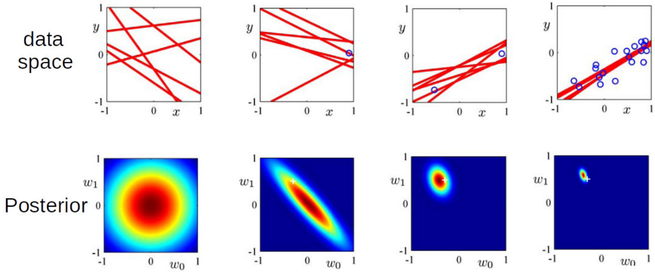
$$p(\boldsymbol{w}|\boldsymbol{y},\mathbf{X},\beta,\lambda) \propto \mathcal{N}(\boldsymbol{w}|\mathbf{0},\lambda^{-1}\mathbf{I}_D) \times \mathcal{N}(\boldsymbol{y}|\mathbf{X}\boldsymbol{w},\beta^{-1}\mathbf{I}_N)$$

- Using the "completing the squares" trick (or directly using Gaussian conditioning formula)

$$\begin{aligned} p(\boldsymbol{w}|\boldsymbol{y},\mathbf{X},\beta,\lambda) &= \mathcal{N}(\boldsymbol{\mu}_N,\boldsymbol{\Sigma}_N) \\ \text{where} \quad \boldsymbol{\Sigma}_N &= (\beta\sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{x}_n^\top + \lambda\mathbf{I}_D)^{-1} = (\beta\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_D)^{-1} \quad \text{(posterior's covariance matrix)} \\ \boldsymbol{\mu}_N &= \boldsymbol{\Sigma}_N\left[\beta\sum_{n=1}^{N}y_n\boldsymbol{x}_n\right] = \boldsymbol{\Sigma}_N\left[\beta\mathbf{X}^\top\boldsymbol{y}\right] = (\mathbf{X}^\top\mathbf{X} + \frac{\lambda}{\beta}\mathbf{I}_D)^{-1}\mathbf{X}^\top\boldsymbol{y} \quad \text{(posterior's mean)} \end{aligned}$$

The Posterior: A Visualization

- Assume a linear regression problem with ground truth $\boldsymbol{w} = [w_0, w_1]$ with $w_0 = -0.3, w_1 = 0.5$
- Assume data generated by a linear regression model $y = w_0 + w_1 x +$ "noise"
  - Note: It's actually 1-D regression (w_0 is just a bias term), or 2-D reg. with feature $[1, x]$
- Figures below show the "data space" and posterior of $\boldsymbol{w}$ for different number of observations (note: with no observations, the posterior = prior)



Try to minimize noise

- The "data space" (red lines) shown above denotes various possible linear regression datasets with data of the form $y = w_0 + w_1 x$ generated using $\boldsymbol{w}$ drawn from the current posterior of $\boldsymbol{w}$

Bayesian Linear Regression: Posterior Predictive Distribution

- Given the posterior $p(\boldsymbol{w}|\boldsymbol{y}, \mathbf{X}, \beta, \lambda) = \mathcal{N}(\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N)$, how to make prediction y_* for a new input $\boldsymbol{x}_*$?
- The posterior predictive distribution will be

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

- Using Gaussian predictive/marginal formula, the posterior predictive will be another Gaussian

$$\boxed{p(y_*|\boldsymbol{x}_*, \mathbf{X}, \boldsymbol{y}, \beta, \lambda) = \mathcal{N}(\boldsymbol{\mu}_N^\top \boldsymbol{x}_*, \beta^{-1} + \boldsymbol{x}_*^\top \boldsymbol{\Sigma}_N \boldsymbol{x}_*)}$$

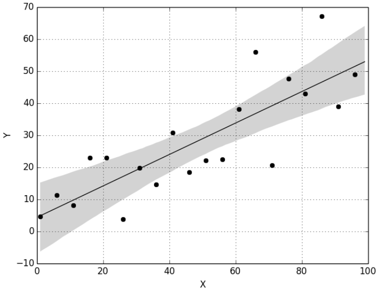
- So we get a predictive mean $\boldsymbol{\mu}_N^\top \boldsymbol{x}_*$ and an input-specific predictive variance $\beta^{-1} + \boldsymbol{x}_*^\top \boldsymbol{\Sigma}_N \boldsymbol{x}_*$
- In contrast, MLE and MAP make "plug-in" predictions (using the point estimate of $\boldsymbol{w}$)

$$\begin{aligned} p(y_*|\boldsymbol{x}_*, \boldsymbol{w}_{MLE}) &= \mathcal{N}(\boldsymbol{w}_{MLE}^\top \boldsymbol{x}_*, \beta^{-1}) && \text{- MLE prediction} \\ p(y_*|\boldsymbol{x}_*, \boldsymbol{w}_{MAP}) &= \mathcal{N}(\boldsymbol{w}_{MAP}^\top \boldsymbol{x}_*, \beta^{-1}) && \text{- MAP prediction} \end{aligned}$$

- Important: Unlike MLE/MAP, the variance of y_* also depends on the input $\boldsymbol{x}_*$ (this, as we will see later, will be very useful in sequential decision-making problems such as active learning)

Posterior Predictive Distribution: An Illustration

Black dots are training examples

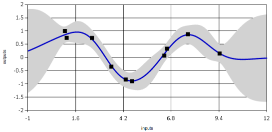


Try to reduce error:
distance between a point vs the fitted line

Width of the shaded region at any x denotes the predictive uncertainty at that x (+/- one std-dev)

Regions with more training examples have smaller predictive variance

Nonlinear Regression?



Gaussian Process:
Gaussian Distribution fitting all those points.

- Can extend the linear regression model to handle nonlinear regression problems
- One way is to replace the feature vectors $\boldsymbol{x}$ by a nonlinear mapping $\phi(\boldsymbol{x})$

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

- The nonlinear mapping can be defined directly, e.g., for a one-dimensional feature x

$$\phi(x) = [1, x, x^2]$$

- Alternatively, a kernel function can be used to implicitly define the nonlinear mapping
- More on nonlinear regression when we discuss Gaussian Processes

What about the hyperparameters of the regression model?

- If hyperparameters are to be estimated, we will have a hierarchical/multiparameter model
- Posterior inference in slightly more involved in this case
- Iterative methods required to learn the weight vector and the hyperparameters, e.g.,
  - Marginal likelihood maximization for hyperparameter estimation
  - Expectation maximization (EM)
  - MCMC or variational inference
- We will discuss more when we talk about inference in hierarchical/multiparameter models

Summary and What Lies Ahead..

- Seen Bayesian inference for several models with a single unknown parameter (and another simple case where we had two unknown parameters - Gaussian with unknown mean and precision)
- Focused on the cases where the likelihood and prior are conjugate
- Both posterior as well as posterior predictive are computable easily in such cases
- Saw various nice properties of exponential family distributions and parameter estimation for such distributions. Also saw estimation in a conditional model (linear regression)
- Things become more challenging/interesting for more complex models, e.g.,
  - Multiple unknown parameters (e.g., hyperparameters, latent variables, hierarchical models etc)
  - Likelihood and prior are not conjugate
- The basic ideas we have seen will turn out to be useful in more complex models as well
  - Conditionally-conjugate models
  - Approximate inference methods (e.g., EM, Gibbs sampling, etc) that resemble alternating optimization techniques