Bayesian Inference for Some Basic Models

Piyush Rai

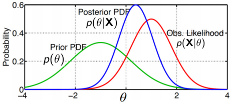
Topics in Probabilistic Modeling and Inference (CS698X)

Jan 12, 2019

Recap: Bayesian Inference

- Given data $\mathbf{X}$ from a model m with parameters θ, the posterior over the parameters θ

$$p(\theta|\mathbf{X}, m) = \frac{p(\mathbf{X}, \theta|m)}{p(\mathbf{X}|m)} = \frac{p(\mathbf{X}|\theta, m)p(\theta|m)}{\int p(\mathbf{X}|\theta, m)p(\theta|m)d\theta} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Marginal likelihood}}$$



- Can use the posterior for various purposes, e.g.,
  - Getting point estimates e.g., mode (though, for this, directly doing point estimation is often easier)
  - Uncertaintly in our estimates of θ (variance, credible intervals, etc)
  - Computing the posterior predictive distribution (PPD) for new data, e.g.,

$$p(x_*|\mathbf{X}, m) = \int p(x_*|\theta, m)p(\theta|\mathbf{X}, m)d\theta$$

- Caveat: Computing the posterior/PPD is in general hard (due to the intractable integrals involved)

Recap: Marginal Likelihood and Its Usefulness

- Likelihood vs Marginal Likelihood: $p(\mathbf{X}|\theta, m)$ vs $p(\mathbf{X}|m)$
  - Prob. of $\mathbf{X}$ for a single θ under model m vs prob. of $\mathbf{X}$ averaged over all θ's under model m
- Can use marginal likelihood $p(\mathbf{X}|m)$ to select the best model from a finite set of models

$$\hat{m} = \arg\max_m p(m|\mathbf{X}) = \arg\max_m p(\mathbf{X}|m)p(m) \quad = \arg\max_m p(\mathbf{X}|m), \text{if } p(m) \text{ is uniform}$$

- Also useful for estimating hyperparam of the assumed model (if we consider m as the hyperparams)
  - Suppose hyperparams of likelihood are α_ℓ and that of prior are α_p (so here $m = \{\alpha_\ell, \alpha_p\}$)
  - Assuming $p(\alpha_\ell, \alpha_p)$ is uniform, hyperparams can be estimated via MLE-II (a.k.a. empirical Bayes)

$$\{\hat{\alpha}_\ell, \hat{\alpha}_p\} = \arg\max_{\alpha_\ell, \alpha_p} p(\mathbf{X}|\alpha_\ell, \alpha_p) = \arg\max_{\alpha_\ell, \alpha_p} \int p(\mathbf{X}|\theta, \alpha_\ell)p(\theta|\alpha_p)d\theta$$

  - Again, note that the integral here may be intractable and may need to be approximated
- Can also compute $p(m|\mathbf{X})$ and do Bayesian Model Averaging: $p(\boldsymbol{x}_*|\mathbf{X}) = \sum_{m=1}^{M} p(\boldsymbol{x}_*|\mathbf{X}, m)p(m|\mathbf{X})$

Recap: Bayesian Inference for a Beta-Bernoulli Model

- Saw the example of estimating the bias $\theta \in (0,1)$ of a coin using Bayesian inference
- Chose a Bernoulli likelihood for each coin toss and a conjugate Beta prior for θ

$$\begin{aligned} p(x_n|\theta) &= \text{Bernoulli}(x_n|\theta) = \theta^{x_n}(1-\theta)^{1-x_n} \\ p(\theta|\alpha,\beta) &= \text{Beta}(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \end{aligned}$$

- Here, prior's hyperparams (assumed fixed here) control its shape; also act as pseudo-observations
- Assuming x_n's as i.i.d. given θ, posterior $p(\theta|\mathbf{X},\alpha,\beta) \propto p(\mathbf{X}|\theta)p(\theta|\alpha,\beta)$ turned out to be Beta

$$p(\theta|\mathbf{X},\alpha,\beta) = \text{Beta}(\theta|\alpha + \sum_{n=1}^{N} x_n, \beta + N - \sum_{n=1}^{N} x_n) = \text{Beta}(\theta|\alpha + N_1, \beta + N_0)$$

- Note: Here posterior only depends on data $\mathbf{X} = \{x_1, \ldots, x_N\}$ via sufficient statistics N_1 and N_0

$$p(\theta|\mathbf{X},\alpha,\beta) = p(\theta|s(\mathbf{X}))$$

- We will see many other cases where the posterior depends on data only via some sufficient statistics

Recap: Making Predictions in the Beta-Bernoulli Model

- The posterior predictive distribution (averaging over all θ weighted by their posterior probabilities):

$$\begin{aligned} p(x_{N+1}=1|\mathbf{X},\alpha,\beta) &= \int_0^1 p(x_{N+1}=1|\theta)p(\theta|\mathbf{X},\alpha,\beta)d\theta \\ &= \int_0^1 \theta \times \text{Beta}(\theta|\alpha+N_1,\beta+N_0)d\theta \\ &= \mathbb{E}[\theta|\mathbf{X}] \\ &= \frac{\alpha+N_1}{\alpha+\beta+N} \end{aligned}$$

- Therefore the posterior predictive distribution: $p(x_{N+1}|\mathbf{X}) = \text{Bernoulli}(x_{N+1} \mid \mathbb{E}[\theta|\mathbf{X}])$
- In contrast, the plug-in predictive distribution using a point estimate $\hat{\theta}$ (e.g., using MLE/MAP)

$$p(x_{N+1}=1|\mathbf{X},\alpha,\beta) \approx p(x_{N+1}=1|\hat{\theta}) = \hat{\theta} \qquad \underline{\text{or equivalently}} \qquad p(x_{N+1}|\mathbf{X}) \approx \text{Bernoulli}(x_{N+1} \mid \hat{\theta})$$

More Examples..

Bayesian Inference for Multinoulli/Multinomial

- Assume N discrete-valued observations $\{x_1, \dots, x_N\}$ with each $x_n \in \{1, \dots, K\}$, e.g.,
  - x_n represents the outcome of a dice roll with K faces
  - x_n represents the class label of the n-th example (total K classes)
  - x_n represents the identity of the n-th word in a sequence of words
- Assume likelihood to be multinoulli with unknown params $\boldsymbol{\pi} = [\pi_1, \dots, \pi_K]$ s.t. $\sum_{k=1}^{K} \pi_k = 1$

$$p(x_n|\boldsymbol{\pi}) = \text{multinoulli}(x_n|\boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{\mathbb{I}[x_n=k]}$$

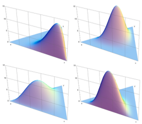
- $\boldsymbol{\pi}$ is a vector of probabilities ("probability vector"), e.g.,
  - Biases of the K sides of the dice
  - Prior class probabilities in multi-class classification
  - Probabilities of observing each words in the vocabulary
- Assume a conjugate Dirichlet prior on $\boldsymbol{\pi}$ with hyperparams $\boldsymbol{\alpha} = [\alpha_1, \dots, \alpha_K]$ (also, $\alpha_k \geq 0, \forall k$)

$$p(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \text{Dirichlet}(\boldsymbol{\pi}|\alpha_1, \dots, \alpha_K) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}$$

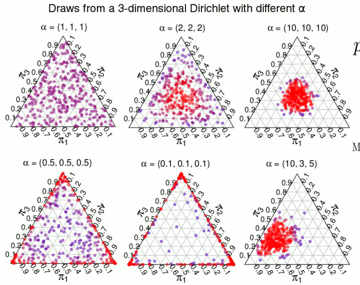
Brief Detour: Dirichlet Distribution

- Very important distribution: Models non-neg. vectors $\boldsymbol{\pi}$ that sum to one (e.g., probability vectors)
- A random draw from Dirichlet will be a point under the probability simplex

PDF for a 3-dim Dirichlet



Red dots denote regions of high probability denity



$$p(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})}\prod_{k=1}^{K}\pi_k^{\alpha_k-1}$$

$$\text{Mean} = \left[\frac{\alpha_1}{\sum_{k=1}^K \alpha_k}, \ldots, \frac{\alpha_K}{\sum_{k=1}^K \alpha_k}\right]$$

$$\text{Mode} = \left[\frac{\alpha_1 - 1}{\sum_{k=1}^K \alpha_k - K}, \ldots, \frac{\alpha_K - 1}{\sum_{k=1}^K \alpha_k - K}\right] \quad (\alpha_k > 1)$$

$$\text{var}(\pi_k) = \frac{\alpha_k(\alpha_0 - \alpha_k)}{\alpha_0^2(\alpha_0 + 1)}$$

$$\alpha_0 = \sum_{k=1}^K \alpha_k$$

- Hyperparams $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_K]$ control the shape of Dirichlet (akin to Beta's hyperparams)
- Can also be thought of as a multi-dimensional Beta distribution
- Note: Can also be seen as normalized version of K independent gamma random variables

Bayesian Inference for Multinoulli/Multinomial

- The posterior over $\boldsymbol{\pi}$ is easy to compute in this case due to conjugacy b/w multinoulli and Dirichlet

$$p(\boldsymbol{\pi}|\mathbf{X},\boldsymbol{\alpha}) = \frac{p(\mathbf{X}|\boldsymbol{\pi},\boldsymbol{\alpha})p(\boldsymbol{\pi}|\boldsymbol{\alpha})}{p(\mathbf{X}|\boldsymbol{\alpha})} = \frac{p(\mathbf{X}|\boldsymbol{\pi})p(\boldsymbol{\pi}|\boldsymbol{\alpha})}{p(\mathbf{X}|\boldsymbol{\alpha})}$$

- Assuming x_n's are i.i.d. given π, $p(\mathbf{X}|\boldsymbol{\pi}) = \prod_{n=1}^{N} p(x_n|\boldsymbol{\pi})$, therefore

$$p(\boldsymbol{\pi}|\mathbf{X},\boldsymbol{\alpha}) \propto \prod_{n=1}^{N}\prod_{k=1}^{K}\pi_k^{\mathbb{I}[x_n=k]}\prod_{k=1}^{K}\pi_k^{\alpha_k-1} = \prod_{k=1}^{K}\pi_k^{\alpha_k+\sum_{n=1}^{N}\mathbb{I}[x_n=k]-1}$$

- Even without computing the normalization constant $p(\mathbf{X}|\boldsymbol{\alpha})$, we can see that it's a Dirichlet! :-)
- Denoting $N_k = \sum_{n=1}^{N}\mathbb{I}[x_n = k]$, i.e., number of observations with value k, the posterior will be

$$p(\boldsymbol{\pi}|\mathbf{X},\boldsymbol{\alpha}) = \text{Dirichlet}(\boldsymbol{\pi}|\alpha_1 + N_1, \ldots, \alpha_K + N_K)$$

- Note: $N_1, \ldots, N_K$ are the sufficient statistics in this case
- Note: If we want, we can also get the MAP estimate of $\boldsymbol{\pi}$ (mode of the above Dirichlet)
  - MAP estimation via standard way will require solving a constraint opt. problem (via Lagrangian)

Bayesian Inference for Multinoulli/Multinomial

- Finally, let's also look at the posterior predictive distribution (i.e., the probability distribution of a new observation $x_* \in \{1, \ldots, K\}$ given the previous observations $\mathbf{X} = \{x_1, \ldots, x_N\}$)

$$p(x_*|\mathbf{X}, \boldsymbol{\alpha}) = \int p(x_*|\boldsymbol{\pi})p(\boldsymbol{\pi}|\mathbf{X}, \boldsymbol{\alpha})d\boldsymbol{\pi}$$

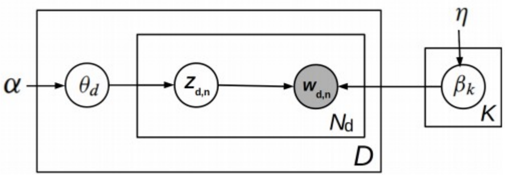
- Note that $p(x_*|\boldsymbol{\pi}) = \text{multinoulli}(x_*|\boldsymbol{\pi})$ and $p(\boldsymbol{\pi}|\mathbf{X}, \boldsymbol{\alpha}) = \text{Dirichlet}(\boldsymbol{\pi}|\alpha_1 + N_1, \ldots, \alpha_K + N_K)$
- We can compute the posterior predictive for each possible outcome (K possibilities)

$$\begin{aligned} p(x_* = k|\mathbf{X}, \boldsymbol{\alpha}) &= \int p(x_* = k|\boldsymbol{\pi})p(\boldsymbol{\pi}|\mathbf{X}, \boldsymbol{\alpha})d\boldsymbol{\pi} \\ &= \int \pi_k \times \text{Dirichlet}(\boldsymbol{\pi}|\alpha_1 + N_1, \ldots, \alpha_K + N_K)d\boldsymbol{\pi} \\ &= \frac{\alpha_k + N_k}{\sum_{k=1}^K \alpha_k + N} \quad \text{(expectation of } \pi_k \text{ under the Dirichlet posterior)} \end{aligned}$$

- Therefore the posterior predictive distribution is multinoulli with posterior mean given as above
- Note that the predicted probabilities are smoothed (the effect of averaging over all possible $\boldsymbol{\pi}$'s)
- Recall that the PPD for the Beta-Bernoulli model also had a similar form!

Applications?

- Both, Beta-Bernoulli and Dirichlet-Multinoulli/Multinomial models are widely used
- We now know how to do fully Bayesian inference if parts of our model have such components



- Some popular examples are
  - Models for text data: Each document can be modeled as a bag-of-words (Beta-Bernoulli) or a sequence of token (Dirichlet-Multinoulli)
  - Bayesian inference for class probabilities in classification models: Class labels of training examples are observations and class proabilities are to be estimated
  - Bayesian inference for mixture models: Cluster ids are our (latent) "observations" of Dir-Mult model and mixing proportions are to be estimated
  - .. and several others, which we will see later..

Some More Examples..

Bayesian Inference for Mean of a Gaussian

- Consider N i.i.d. observations $\mathbf{X} = \{x_1, \ldots, x_N\}$ drawn from a one-dim Gaussian $\mathcal{N}(x|\mu, \sigma^2)$

$$p(x_n|\mu, \sigma^2) = \mathcal{N}(x|\mu, \sigma^2) \propto \exp\left[-\frac{(x_n - \mu)^2}{2\sigma^2}\right]$$

$$p(\mathbf{X}|\mu, \sigma^2) = \prod_{n=1}^{N} p(x_n|\mu, \sigma^2)$$

- Assume the mean $\mu \in \mathbb{R}$ of the Gaussian is unknown and assume variance σ^2 to be known/fixed
- We wish to estimate the unknown μ given the data $\mathbf{X}$
- Let's do fully Bayesian inference for μ (not MLE/MAP)
- We first need a prior distribution for the unknown param. μ
- Let's choose a Gaussian prior on μ, i.e., $p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)$ with μ_0, σ_0^2 as fixed
- The prior basically says that the mean μ is close to μ_0 (with some uncertainty depending on σ_0^2)

Bayesian Inference for Mean of a Gaussian

- The posterior distribution for the unknown mean parameter μ

$$p(\mu|\mathbf{X}) = \frac{p(\mathbf{X}|\mu)p(\mu)}{p(\mathbf{X})} \quad \propto \quad \prod_{n=1}^{N} \exp\left[-\frac{(x_n-\mu)^2}{2\sigma^2}\right] \times \exp\left[-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right]$$

- Simplifying the above (using completing the squares trick) gives $p(\mu|\mathbf{X}) \propto \exp\left[-\frac{(\mu-\mu_N)^2}{2\sigma_N^2}\right]$ with

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$$

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\bar{x} \qquad \left(\text{where } \bar{x} = \frac{\sum_{n=1}^{N} x_n}{N}\right)$$

- Posterior and prior have the same form (not surprising; the prior was conjugate to the likelihood)
- Consider what happens as N (number of observations) grows very large?
  - The posterior's variance σ_N^2 approaches σ^2/N (and goes to 0 as $N \to \infty$)
  - The posterior's mean μ_N approaches $\bar{x}$ (which is also the MLE solution)

Bayesian Inference for Mean of a Gaussian

- What is the posterior predictive distribution $p(x_*|\mathbf{X})$ of a new observation x_*?
- Using the inferred posterior $p(\mu|\mathbf{X})$, we can find the posterior predictive distribution

$$p(x_*|\mathbf{X}) = \int p(x_*|\mu,\sigma^2)p(\mu|\mathbf{X})d\mu = \int \mathcal{N}(x_*|\mu,\sigma^2)\mathcal{N}(\mu|\mu_N,\sigma_N^2)d\mu = \mathcal{N}(x_*|\mu_N,\sigma^2+\sigma_N^2)$$

- Note; Can also get the above result by thinking of x_* as $x_* = \mu + \epsilon$ where $\mu \sim \mathcal{N}(\mu_N, \sigma_N^2)$, and $\epsilon \sim \mathcal{N}(0,\sigma^2)$ is independently added observation noise
- Note that, as per the above, the uncertainty in distribution of x_* now has two components
  - σ^2: Due to the noisy observation model, σ_N^2: Due to the uncertainty in μ
- In contrast, the plug-in predictive posterior, given a point estimate $\hat{\mu}$ (e.g., MLE/MAP) would be

$$p(x_*|\mathbf{X}) = \int p(x_*|\mu,\sigma^2)p(\mu|\mathbf{X})d\mu \approx p(x_*|\hat{\mu},\sigma^2) = \mathcal{N}(x_*|\hat{\mu},\sigma^2)$$

.. which doesn't incorporate the uncertainty in our estimate of μ (since we used a point estimate)

- Note that as $N \to \infty$, both approaches would give the same $p(x_*|\mathbf{X})$ since $\sigma_N^2 \to 0$

Bayesian Inference for Variance of a Gaussian

- Again consider N i.i.d. observations $\mathbf{X} = \{x_1, \ldots, x_N\}$ drawn from a one-dim Gaussian $\mathcal{N}(x|\mu, \sigma^2)$

$$p(x_n|\mu, \sigma^2) = \mathcal{N}(x|\mu, \sigma^2) \quad \text{and} \quad p(\mathbf{X}|\mu, \sigma^2) = \prod_{n=1}^{N} p(x_n|\mu, \sigma^2)$$

- Assume the variance $\sigma^2 \in \mathbb{R}_+$ of the Gaussian is unknown and assume mean μ to be known/fixed
- Let's estimate σ^2 given the data $\mathbf{X}$ using fully Bayesian inference (not MLE/MAP)
- We first need a prior distribution for σ^2. What prior $p(\sigma^2)$ to choose in this case?
- If we want a conjugate prior, it should have the same form as the likelihood

$$p(x_n|\mu, \sigma^2) \propto (\sigma^2)^{-1/2} \exp\left[-\frac{(x_n - \mu)^2}{2\sigma^2}\right]$$

- An inverse-gamma prior $IG(\alpha, \beta)$ has this form (α, β are shape and scale hyperparams, resp)

$$p(\sigma^2) \propto (\sigma^2)^{-(\alpha+1)} \exp\left[-\frac{\beta}{\sigma^2}\right] \qquad \text{(note: mean of } IG(\alpha, \beta) = \frac{\beta}{\alpha - 1}\text{)}$$

- (Verify) The posterior $p(\sigma^2|\mathbf{X}) = IG(\alpha + \frac{N}{2}, \beta + \frac{\sum_{n=1}^{N}(x_n - \mu)^2}{2})$. Again IG due to conjugacy.

Working with Gaussians: Variance vs Precision

- Often, it is easier to work with the precision (=1/variance) rather than variance

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

- If mean is known, for precision Gamma(α,β) is a conjugate prior to Gaussian likelihood

$$p(\tau) \propto (\tau)^{(\alpha-1)}\exp\left[-\beta\tau\right] \qquad \text{(note: mean of Gamma}(\alpha,\beta) = \frac{\alpha}{\beta}\text{)}$$

.. where α and β are the shape and rate hyperparamers, respectively, for the Gamma

- (Verify) The posterior $p(\tau|\mathbf{X})$ will also be Gamma$(\alpha + \frac{N}{2}, \beta + \frac{\sum_{n=1}^{N}(x_n-\mu)^2}{2})$
- Note: Gamma distribution can be defined in terms of shape and scale or shape and rate parametrization (scale = 1/rate). Likewise, inverse Gamma can also be defined both shape and scale (which we saw) as well as shape and rate parametrizations.

Bayesian Inference for Both Parameters of a Gaussian!

- Gaussian with unknown scalar mean and unknown scalar precision (two parameters)
- Consider N i.i.d. observations $\mathbf{X} = \{x_1, \ldots, x_N\}$ drawn from a one-dim Gaussian $\mathcal{N}(x|\mu, \lambda^{-1})$
- Assume both mean μ and precision λ to be unknown. The likelihood will be

$$\begin{aligned} p(\mathbf{X}|\mu,\lambda) &= \prod_{n=1}^{N} \sqrt{\frac{\lambda}{2\pi}} \exp\left[-\frac{\lambda}{2}(x_n-\mu)^2\right] \\ &\propto \left[\lambda^{1/2}\exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^N \exp\left[\lambda\mu\sum_{n=1}^{N}x_n - \frac{\lambda}{2}\sum_{n=1}^{N}x_n^2\right] \end{aligned}$$

- If we want a conjugate joint prior $p(\mu, \lambda)$, it must have the same form as likelihood. Suppose

$$p(\mu,\lambda) \propto \left[\lambda^{1/2}\exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^{\kappa_0} \exp\left[\lambda\mu c - \lambda d\right]$$

- What's this prior? A normal-gamma (Gaussian-gamma) distribution! (will see its form shortly)
  - Can be used when we wish to estimate the unknown mean and unknown precision of a Gaussian
  - Note: Its multivariate version is the Normal-Wishart (for multivariate mean and precision matrix)

Normal-gamma (Gaussian-gamma) Distribution

- We saw that the conjugate prior needed to have the form

$$
\begin{aligned}
p(\mu,\lambda) &\propto \left[\lambda^{1/2}\exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^{\kappa_0}\exp\left[\lambda\mu c-\lambda d\right] \\
&= \underbrace{\exp\left[-\frac{\kappa_0\lambda}{2}(\mu-c/\kappa_0)^2\right]}_{\text{prop. to a Gaussian}}\underbrace{\lambda^{\kappa_0/2}\exp\left[-\left(d-\frac{c^2}{2\kappa_0}\right)\lambda\right]}_{\text{prop. to a gamma}} \qquad \text{(re-arranging terms)}
\end{aligned}
$$

- The above is product of a normal and a gamma distribution[1]

$$p(\mu,\lambda)=\mathcal{N}(\mu|\mu_0,(\kappa_0\lambda)^{-1})\text{Gamma}(\lambda|\alpha_0,\beta_0)=\text{NG}(\mu_0,\kappa_0,\alpha_0,\beta_0)$$

where $\mu_0 = c/\kappa_0$, $\alpha_0 = 1+\kappa_0/2$, $\beta_0 = d - c^2/2\kappa_0$ are prior's hyperparameters

- $p(\mu,\lambda) = \text{NG}(\mu_0,\kappa_0,\alpha_0,\beta_0)$ is a conjugate for the mean-precision pair (μ,λ)
  - A useful prior in many problems involving Gaussians with unknown mean and precision

[1] shape-rate parametrization assumed for the gamma

Joint Posterior

- Due to conjugacy, the joint posterior $p(\mu, \lambda|\mathbf{X})$ will also be normal-gamma

$$p(\mu, \lambda|\mathbf{X}) \propto p(\mathbf{X}|\mu, \lambda)p(\mu, \lambda)$$

- Plugging in the expressions for $p(\mathbf{X}|\mu, \lambda)$ and $p(\mu, \lambda)$, we get

$$p(\mu, \lambda|\mathbf{X}) = \mathrm{NG}(\mu_N, \kappa_N, \alpha_N, \beta_N) = \mathcal{N}(\mu|\mu_N, (\kappa_N\lambda)^{-1})\mathrm{Gamma}(\lambda|\alpha_N, \beta_N)$$

where the updated posterior hyperparameters are given by[2]

$$\begin{aligned}
\mu_N &= \frac{\kappa_0\mu_0 + N\bar{x}}{\kappa_0 + N} \\
\kappa_N &= \kappa_0 + N \\
\alpha_N &= \alpha_0 + N/2 \\
\beta_N &= \beta_0 + \frac{1}{2}\sum_{n=1}^{N}(x_n - \bar{x})^2 + \frac{\kappa_0 N(\bar{x} - \mu_0)^2}{2(\kappa_0 + N)}
\end{aligned}$$

[2] For full derivation, refer to "Conjugate Bayesian analysis of the Gaussian distribution" - Murphy (2007)

Other Quantities of Interest[3]

- Already saw that joint post. $p(\mu,\lambda|\mathbf{X}) = \text{NG}(\mu_N,\kappa_N,\alpha_N,\beta_N) = \mathcal{N}(\mu|\mu_N,(\kappa_N\lambda)^{-1})\text{Gamma}(\lambda|\alpha_N,\beta_N)$
- Marginal posteriors for μ and λ

$$\begin{aligned} p(\lambda|\mathbf{X}) &= \int p(\mu,\lambda|\mathbf{X})d\mu = \text{Gamma}(\lambda|\alpha_N,\beta_N) \\ p(\mu|\mathbf{X}) &= \int p(\mu,\lambda|\mathbf{X})d\lambda = \int p(\mu|\lambda,\mathbf{X})p(\lambda|\mathbf{X})d\lambda = \underbrace{t_{2\alpha_N}(\mu|\mu_N,\beta_N/(\alpha_N\kappa_N))}_{\text{t distribution}} \end{aligned}$$

- Exercise: What will be the conditional posteriors $p(\mu|\lambda,\mathbf{X})$ and $p(\lambda|\mu,\mathbf{X})$?
- Marginal likelihood of the model

$$p(\mathbf{X}) = \frac{\Gamma(\alpha_N)}{\Gamma(\alpha_0)}\frac{\beta_0^{\alpha_0}}{\beta_N^{\alpha_N}}\left(\frac{\kappa_0}{\kappa_N}\right)^{\frac{1}{2}}(2\pi)^{-N/2}$$

- Posterior predictive distribution of a new observation x_*

$$p(x_*|\mathbf{X}) = \int \underbrace{p(x_*|\mu,\lambda)}_{\text{Gaussian}}\ \underbrace{p(\mu,\lambda|\mathbf{X})}_{\text{Normal-Gamma}}\ d\mu d\lambda = t_{2\alpha_N}\left(x_*|\mu_N,\frac{\beta_N(\kappa_N+1)}{\alpha_N\kappa_N}\right)$$

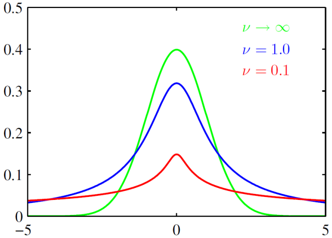
[3] For full derivations, refer to "Conjugate Bayesian analysis of the Gaussian distribution" - Murphy (2007)

An Aside: general-t and Student-t distribution

- Equivalent to an infinite sum of Gaussian distributions, with same means but different precisions

$$\begin{aligned} p(x|\mu, a, b) &= \int \mathcal{N}(x|\mu, \lambda^{-1})\text{Gamma}(\lambda|a, b)d\lambda \\ &= t_{2a}(x|\mu, b/a) = t_\nu(x|\mu, \sigma^2) \quad \text{(general-t distribution)} \end{aligned}$$

- $\mu = 0, \sigma^2 = 1$ gives the Student-t distribution (t_ν). Note: If $x \sim t_\nu(\mu, \sigma^2)$ then $\frac{x-\mu}{\sigma} \sim t_\nu$
- An illustration of student-t



- t distribution has a "fatter" tail than a Gaussian and also sharper around the mean
    - Also a useful prior for sparse modeling

Inferring Parameters of Gaussian: Some Other Cases

- We only considered the simple 1-D Gaussian distribution
- The approach also extends to inferring parameters of a multivariate Gaussian
  - For the unknown mean and precision matrix, normal-Wishart distribution can be used as prior
- Posterior updates have forms similar to that in the 1-D case
- When working with mean-variance, we can use normal-inverse gamma as conjugate prior (or normal-inverse Wishart when working with mean-covariance matrix in case of multivariate Gaussian distribution)
- Other priors can also be used as well when inferring parameters of Gaussians, e.g.,
  - normal-Inverse χ^2 distribution is commonly used in Statistics community for scalar mean-variance
  - Uniform priors can also be used
  - Look at BDA Chapter 3 for such examples
- Also refer to "Conjugate Bayesian analysis of the Gaussian distribution" - Murphy (2007) for various examples and more detailed derivations

Next Class: More examples of Bayesian inference with Gaussians