Learning Hyperparameters via MLE-II, and Introduction to Multiparameter Models

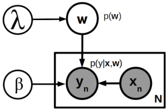
Piyush Rai

Topics in Probabilistic Modeling and Inference (CS698X)

Jan 18, 2018

Learning Hyperparameters

- Let's (re-)consider the Bayesian linear regression model



- Ideally, we would like to learn the hyperparams as well (rather than using cross-validation)
- When also learning the hyperparams, the set of unknowns will be $(\boldsymbol{w}, \beta, \lambda)$
- Suppose we have a prior $p(\boldsymbol{w}, \beta, \lambda) = p(\boldsymbol{w}|\lambda)p(\beta)p(\lambda)$
- Ideally, would like to get the joint posterior over all the unknowns

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

where $p(\boldsymbol{y}|\mathbf{X}) = \int p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w}, \beta)p(\boldsymbol{w}|\lambda)p(\beta)p(\lambda)\, d\boldsymbol{w}\, d\lambda d\beta$ (ouch! intractable)

- Note: Computing posterior predictive distribution would also require integrating out hyperparams

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

Learning Hyperparameters

- Intractability in hyperparameter estimation can usually be handled via
  - Doing point estimation for hyperparams and learning full posterior for parameters (today's lecture)
    - This procedure is commonly known as Type-II MLE, or evidence approximation, or empirical Bayes
    - Note: This won't exactly be 100% Bayesian but nevertheless very widely used
  - Using general approximate inference methods, e.g.,
    - MCMC, Variational Bayesian Inference (we will look at these later)
- In either case, the resulting algorithms work in an alternating fashion
  - Basically, alternate between estimating the parameters and hyperparameters (until convergence)
  - Many algorithms have such a flavor (EM, Gibbs sampling, variational inference, etc.)
- Note: This approach to learning hyperparams is not limited to only learning "hyperparams" (instead of hyperparams, we might have a large number of parameters in a multiparameter model)

Learning Hyperparameters

- For Bayesian linear regression, the posterior over the unknowns can be written as

$$p(\boldsymbol{w}, \beta, \lambda | \mathbf{X}, \boldsymbol{y}) = p(\boldsymbol{w} | \mathbf{X}, \boldsymbol{y}, \beta, \lambda) p(\beta, \lambda | \mathbf{X}, \boldsymbol{y}) \qquad \text{(from product rule)}$$

- Computing $p(\boldsymbol{w} | \mathbf{X}, \boldsymbol{y}, \beta, \lambda)$ is easy for Bayesian linear regression if β and λ are "known"
- However, computing $p(\beta, \lambda | \mathbf{X}, \boldsymbol{y})$ is NOT easy. Note that it is given by

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

.. where the denominator $p(\boldsymbol{y} | \mathbf{X}) = \int p(\boldsymbol{y} | \mathbf{X}, \beta, \alpha) p(\beta) p(\lambda) d\beta d\lambda$ is intractable in general

- One work-around to this intractability is to do point estimation for these hyperparameters

Learning Hyperparameters via Point Estimation

- Let's approximate $p(\beta, \lambda|\mathbf{X}, \boldsymbol{y}) = \frac{p(\boldsymbol{y}|\mathbf{X},\beta,\alpha)p(\beta)p(\lambda)}{p(\boldsymbol{y}|\mathbf{X})}$ by a point function δ at its mode

$$p(\beta, \lambda|\mathbf{X}, \boldsymbol{y}) \approx \delta(\hat{\beta}, \hat{\lambda}) \quad \text{where} \quad \hat{\beta}, \hat{\lambda} = \arg\max_{\beta,\lambda} p(\beta, \lambda|\mathbf{X}, \boldsymbol{y})$$

- Moreover, if $p(\beta)$, $p(\lambda)$ are uniform/uninformative priors then

$$\hat{\beta}, \hat{\lambda} = \arg\max_{\beta,\lambda} p(\boldsymbol{y}|\mathbf{X}, \beta, \lambda)$$

- This is point estimation of hyperparams by doing MLE on the marginal likelihood
  - Also called Type-II MLE and Evidence based method for hyperparam estimation
- With the above point approximation $p(\beta, \lambda|\mathbf{X}, \boldsymbol{y}) \approx \delta(\hat{\beta}, \hat{\lambda})$

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

- Note: $p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \hat{\beta}, \hat{\lambda})$ is also called the conditional posterior of $\boldsymbol{w}$ given $\lambda = \hat{\lambda}$ and $\beta = \hat{\beta}$ (and for Bayesian linear regression, we know what the form of this conditional posterior is!)

An Iterative Scheme

We can now follow an iterative scheme for estimating $\boldsymbol{w}$, λ, and β

- Initialize $\hat{\lambda}$ and $\hat{\beta}$
- Repeat until convergence
  - Estimate the posterior over $\boldsymbol{w}$ as

$$p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \hat{\lambda}, \hat{\beta}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

    where $\boldsymbol{\Sigma} = (\hat{\beta}\sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top + \hat{\lambda}\mathbf{I}_D)^{-1}$ and $\boldsymbol{\mu} = \boldsymbol{\Sigma}(\hat{\beta}\sum_{n=1}^{N} y_n \boldsymbol{x}_n)$
  - Re-estimate $\hat{\lambda}$ as

$$\hat{\lambda} = \frac{\gamma}{\boldsymbol{\mu}^\top \boldsymbol{\mu}} \qquad \text{(note: } \gamma \text{ depends on } \boldsymbol{\Sigma}\text{)}$$

  - Re-estimate $\hat{\beta}$ as

$$\hat{\beta} = \frac{N - \gamma}{\sum_{n=1}^{N}(y_n - \boldsymbol{\mu}^\top \boldsymbol{x}_n)^2}$$

- Note: This is akin to an Expectation-Maximization (EM) scheme. More on this later.

Using Type-II MLE Approximations

- With the Type-II MLE approximation $p(\beta,\lambda|\mathbf{X},\boldsymbol{y}) \approx \delta(\hat{\beta},\hat{\lambda})$, the posterior over unknowns

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

- Therefore the overall posterior over the unknowns $\boldsymbol{w},\beta,\lambda$ is the same as the conditional posterior of $\boldsymbol{w}$ treating the hyperparams $\hat{\beta},\hat{\lambda}$ fixed at their Type-II MLE based estimates
- The posterior predictive distribution can also be approximated as

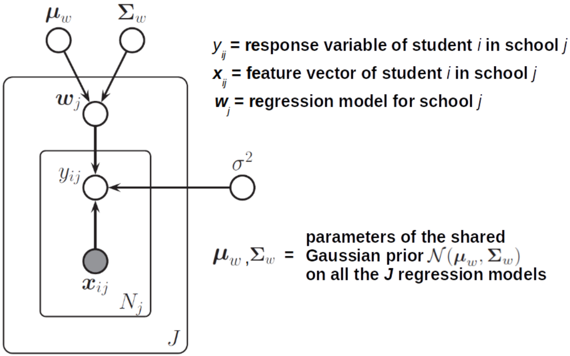
$$\begin{aligned}
p(y_*|\boldsymbol{x}_*,\mathbf{X},\boldsymbol{y}) &= \int p(y_*|\boldsymbol{x}_*,\boldsymbol{w},\beta)p(\boldsymbol{w},\beta,\lambda|\mathbf{X},\boldsymbol{y})\, d\boldsymbol{w}\; d\beta\; d\lambda \\
&= \int p(y_*|\boldsymbol{x}_*,\boldsymbol{w},\beta)p(\boldsymbol{w}|\mathbf{X},\boldsymbol{y},\beta,\lambda)p(\beta,\lambda|\mathbf{X},\boldsymbol{y}) d\beta\; d\lambda\; d\boldsymbol{w} \\
&\approx \int p(y_*|\boldsymbol{x}_*,\boldsymbol{w},\beta)p(\boldsymbol{w}|\mathbf{X},\boldsymbol{y},\hat{\beta},\hat{\lambda})\, d\boldsymbol{w}
\end{aligned}$$

- This is also the same as the usual posterior predictive distribution we have seen earlier, except we are treating the hyperparams $\hat{\beta},\hat{\lambda}$ fixed at their Type-II MLE based estimates

Multiparameter Models

Multiparameter Models

- Multiparameter models arise in many situations, e.g.,
  - Probabilistic models with unknown hyperparameters (e.g., Bayesian linear regression we just saw)
  - Joint analysis of data from multiple (and possibly related) groups



- .. and in fact, pretty much in any non-toy example of probablistic model :)

Multiparameter Models

- A simple two-parammeter model: Consider a model with two unknowns θ_1 and θ_2, e.g.,
  - Gaussian $\mathcal{N}(\mu, \tau^{-1})$ observation model with both unknown mean $(\mu = \theta_1)$ and precision $(\tau = \theta_2)$
- Assume a likelihood model $p(\boldsymbol{y}|\theta_1, \theta_2)$ for observed data $\boldsymbol{y}$
- Assume a joint prior distribution $p(\theta_1, \theta_2)$ over both unknowns
- The joint posterior distribution for (θ_1, θ_2) would be

$$p(\theta_1, \theta_2|\boldsymbol{y}) \propto p(\boldsymbol{y}|\theta_1, \theta_2)p(\theta_1, \theta_2)$$

- If the joint prior $p(\theta_1, \theta_2)$ is conjugate to $p(\boldsymbol{y}|\theta_1, \theta_2)$, inference would be direct and easy!
  - The joint posterior $p(\theta_1, \theta_2|\boldsymbol{y})$ will have the same form as the joint prior $p(\theta_1, \theta_2)$
  - Such cases are more of exception than a norm!
  - Coming with with jointly conjugate priors is not always possible/easy

Inference in Multiparameter Models

- There exist several methods to perform inference in multiparameter models
- Some of the inference methods (which we will study) include
  - Specialized inference methods
    - Apply only for some specific types of multiparameter models (will see some examples)
  - Several general-purpose inference methods
    - Markov Chain Monte Carlo (MCMC) methods such as Gibbs Sampling
    - Variational Bayesian Inference (abbreviated as VB or VI) methods
    - Expectation-Maximization (is actually a special case of VB/VI)
    - Bayesian inference with MLE-II for hyperparameters (we already saw this)
- Note: Some of these methods perform hybrid-style inference
  - Full posteriors for some unknowns, point estimates for others (e.g., EM, MLE-II based methods)
- Most (in fact, all) of these methods perform (usually approximate) inference in an iterative fashion
  - Basically, iteratively infer one unknown given all others fixed at the current value (until convergence)

Marginal Posterior

- Often, instead of (or in addition to) joint posterior, we may be interested in marginal posterior
- Marginal posterior is the posterior of one unknown given the data
- In the two-param case, given joint post. $p(\theta_1, \theta_2|\boldsymbol{y})$, we can get the marginal posteriors as

$$p(\theta_1|\boldsymbol{y}) = \int p(\theta_1, \theta_2|\boldsymbol{y})d\theta_1 \qquad \text{and} \qquad p(\theta_2|\boldsymbol{y}) = \int p(\theta_1, \theta_2|\boldsymbol{y})d\theta_2$$

- In some cases, the marginal posterior can be computed directly using other ways too, e.g.,

$$p(\theta_1|\boldsymbol{y}) \propto p(\boldsymbol{y}|\theta_1)p(\theta_1)$$

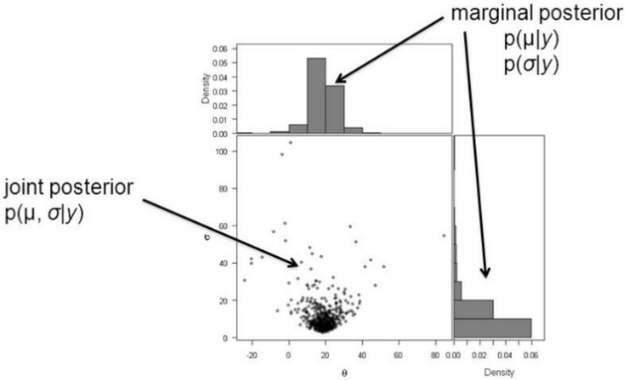
.. but this requires the marginal likelihood $p(\boldsymbol{y}|\theta_1)$ and the marginal prior $p(\theta_1)$ where

$$\begin{aligned} p(\boldsymbol{y}|\theta_1) &= \int p(\boldsymbol{y}|\theta_1, \theta_2)p(\theta_2)d\theta_2 \\ p(\theta_1) &= \int p(\theta_1|\theta_2)p(\theta_2)d\theta_2 \end{aligned}$$

.. getting the marginal likelihood/prior may or may not be easy (depends on distributions involved)

Marginal Posterior: An Illustration

- Often the joint posterior may be hard to visualize/interpret
- Each marginal posterior may provide a better view of the individual parameter's posterior



Conditional Posterior

- Conditional posterior is another very useful quantity when doing Bayesian inference
- E.g.: For a joint posterior $p(\theta_1, \theta_2|\boldsymbol{y})$, can define conditional posteriors $p(\theta_1|\theta_2, \boldsymbol{y})$ or $p(\theta_2|\theta_1, \boldsymbol{y})$
- Note that $p(\theta_1|\theta_2, \boldsymbol{y})$ assumes that θ_2 is fixed at some value (likewise for $p(\theta_2|\theta_1, \boldsymbol{y})$)
- Conditional posteriors are usually easier to derive as compared to the joint $p(\theta_1, \theta_2|\boldsymbol{y})$
  - Recall our example of Gaussian mean inference with variance known (and vice-versa)
- Many inference algorithms for joint posterior are based on inferring the conditional posteriors, e.g.,
  - e.g., Gibbs Sampling (an MCMC based sampling algorithm)

Gibbs Sampling

- Gibbs sampler iteratively draws random samples from conditional posteriors
- When run long enough, the sampler produces samples from the joint posterior
- For the simple two-parameter case $\theta = (\theta_1, \theta_2)$, the Gibb sampler looks like this
  - Initialize $\theta_2^{(0)}$
  - For $s = 1, \ldots, S$
    - Draw a random sample for θ_1 as $\theta_1^{(s)} \sim p(\theta_1 | \theta_2^{(s-1)}, \boldsymbol{y})$
    - Draw a random sample for θ_2 as $\theta_2^{(s)} \sim p(\theta_2 | \theta_1^{(s)}, \boldsymbol{y})$
- The set of S random samples $\{\theta_1^{(s)}, \theta_2^{(s)}\}_{s=1}^{S}$ represent the joint posterior distribution $p(\theta_1, \theta_2 | \boldsymbol{y})$
- More on Gibbs sampling when we discuss MCMC sampling algorithm (above is the high-level idea)
- Many other inference algorithms for inferring the joint posterior (e.g., variational inference) also use conditional posteriors is a similar manner