Indian Institute of Technology, Kanpur
Probabilistic Machine Learning and Inference, 18–19

Instructors: Piyush Rai
Authors: Gurpreet Singh
Date: December 2, 2017

SCRIBE

4

Hyper-Parameter Estimation using MLE-II

In the previous scribes, we have discussed how to estimate paramters for probabilistic models, and looked at parameter estimation for the coin-toss example as well as the regression problem. In this scribe, we will look at estimating the hyperparameters for Bayesian Regression, using the MLE-II approach.

1. Learning Hyperparameters

Let us reconsider the Bayesian Linear Regression model.

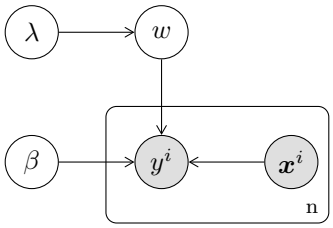

Figure 1: Plate Notation of a Bayesian Linear Regression Model

As discussed earlier, we estimate the hyperparameters using methods like Cross-Validation, however we wish to probabilitically estimate a good value for the hyperparameters. There are various methods to do this, such as

(i) Doing point estimation of hyperparameters and learning full posterior for the parameters (MLE-II)

(ii) Using general inference methods such as MCMC and Variational Bayes (Discussed in later scribes)

Before we move on to learning these methods, we must first understand why we need them. Let us try a simple probabilistic approach to estimating hyperparameters (determining the full posterior).

Let us first take a look at the form of the prior for the hyperparameters and the parameters (all can be considered as the parameters of the model) $\mathbb{P}[\mathbf{w}, \beta, \lambda]$.

From the plate notation, we can say that β and λ are independent of other parameters, whereas $\mathbf{w}$ is dependent on λ (which makes intuitive sense as well since λ is the precision value for the assumed gaussian that explains the distribution of $\mathbf{w}$). Therefore,

$$\mathbb{P}[\mathbf{w}, \beta, \lambda] \quad = \quad \mathbb{P}\left[\mathbf{w} \mid \lambda\right] \cdot \mathbb{P}[\beta] \cdot \mathbb{P}[\lambda]$$

Now, let us try to compute the posterior —

$$\mathbb{P}\left[\mathbf{w}, \beta, \lambda \mid \mathcal{X}, \mathcal{Y}\right] \quad = \quad \frac{\mathbb{P}\left[\mathcal{Y} \mid \mathcal{X}, \mathbf{w}, \beta, \lambda\right] \mathbb{P}[\mathbf{w}, \beta, \lambda]}{\mathbb{P}\left[\mathcal{Y} \mid \mathcal{X}\right]}$$

where

$$\mathbb{P}\left[\mathcal{Y} \mid \mathcal{X}\right] \quad = \quad \int_{\mathbf{w},\beta,\lambda} \mathbb{P}\left[\mathbb{P}\left[\mathcal{Y} \mid \mathcal{X}, \mathbf{w}, \beta\right]\right] \mathbb{P}[\mathbf{w}, \beta, \lambda] \tag{1}$$

It can be shown that the above term is intractable. Therefore, we cannot perform Bayesian Inference on this directly. This is the reason we require different approaches to predicting the hyperparameters.

Exercise 4.1. Prove that the integral in Equation 1 is indeed intractable.

Note. The approaches to estimate hyperparamters are not just limited to hyperparameters, but can also be extended for multi-parameter models

2. Learning Hyperparameters

As mentioned in the previous section, we have the prior over $\mathbf{w}, \beta, \lambda$ with the form

$$\mathbb{P}\left[\mathbf{w}, \beta, \lambda\right] \quad = \quad \mathbb{P}\left[\mathbf{w} \mid \lambda\right] \cdot \mathbb{P}\left[\beta\right] \cdot \mathbb{P}\left[\lambda\right] \tag{2}$$

We can also write the posterior (using Bayes Rule) as

$$\mathbb{P}\left[\mathbf{w}, \beta, \lambda \mid \mathcal{X}, \mathcal{Y}\right] \quad = \quad \mathbb{P}\left[\mathbf{w} \mid \mathcal{X}, \mathcal{Y}, \beta, \lambda\right] \cdot \mathbb{P}\left[\beta, \lambda\right] \tag{3}$$

Now, we already know how to compute $\mathbb{P}\left[\mathbf{w} \mid X, \mathbf{y}, \beta, \lambda\right]$ if β and λ are known. However, computing $\mathbb{P}\left[\beta, \lambda \mid X, \mathbf{Y}\right]$ is intractable.

$$\mathbb{P}\left[\beta, \lambda \mid X, \mathbf{Y}\right] \quad = \quad \frac{\mathbb{P}\left[\mathbf{y} \mid X, \beta, \lambda\right] \cdot \mathbb{P}\left[\beta\right] \cdot \mathbb{P}\left[\lambda\right]}{\mathbb{P}\left[\mathbf{y} \mid X\right]}$$

From the arguments in the discussion in Section 1, we know that this quantity is intractable, therfore, we need a work-around for this.

A possible solution is (as already stated) take only the point estimate of β and λ. Therfore, we only need to find an estimate of β and λ and plug that into the posterior for simple Bayesian Linear Regression (discussed in the previous scribe) *i.e.* assume α and β to be known.

$$\mathbb{P}\left[\mathbf{w} \mid \mathbf{y}, X, \widehat{\beta}, \widehat{\lambda}\right] \quad = \quad \mathcal{N}\left(\mathbf{w} \mid \boldsymbol{\mu}, \Sigma\right) \tag{4}$$

where

$$\begin{aligned} \Sigma \quad &= \quad \left(\widehat{\lambda}\mathbf{I} + \widehat{\beta} X X^{\mathrm{T}}\right)^{-1} \\ \boldsymbol{\mu} \quad &= \quad \Sigma \left(\widehat{\beta} \sum_{i \in [n]} y^i \mathbf{x}^i\right) \end{aligned}$$

Therefore, we now find the point estimate of β and λ

$$\widehat{\beta}, \widehat{\lambda} \quad = \quad \arg\max_{\beta, \lambda} \mathbb{P}\left[\beta, \lambda \mid X, \mathbf{y}\right]$$

Note that this is MAP estimation. Therefore, we need to assume some prior on β and λ. However, we simply assume that the priors on β and λ are uniform and therefore uninformative. Thus, the MAP estimation boils down to MLE estimation.

$$\widehat{\beta}, \widehat{\lambda} \quad = \quad \arg\max_{\beta, \lambda} \mathbb{P}\left[\mathbf{y} \mid X, \beta, \lambda\right] \tag{5}$$

Exercise 4.2. Prove that if the prior is a uniform distribution, then MAP estimate is equal to the MLE estimate.

Remark. We can also do MAP estimation on β and λ, if we assume some informative prior for them. The method could be called MAP-II

At this point, it may seem like we are done with the estimation of the hyperparameters. However, we are missing a point. The RHS in Equation 5 does not have a closed form. This is evident from the following equations.

$$\mathbb{P}\left[\mathbf{y} \mid X, \beta, \lambda\right] \quad = \quad \int_{\mathbf{w}} \mathbb{P}\left[\mathbf{y} \mid X, \mathbf{w}, \beta\right] \cdot \mathbb{P}\left[\mathbf{w} \mid \lambda\right] \, d\mathbf{w}$$

We can integrate this, using properties of Gaussian Distributions, to get the following term, the intermediate steps for which are left as an exercise.

$$\mathbb{P}\left[\mathbf{y} \mid X, \beta, \lambda\right] \quad = \quad \mathcal{N}\left(\mathbf{y} \mid \mathbf{0}, \beta^{-1}\mathbf{I} + \lambda^{-1} X X^{\mathrm{T}}\right) \tag{6}$$

Now MLE solution of this is generally not in closed form, hence, we require an alternate tecnique, known as Alternating Minimization, discussed in the next section.

3. Alternating Optimization

Suppose we already have an estimate of β and λ. Then we know we can easily compute $\mathbb{P}\left[\mathbf{w} \mid X, \mathbf{y}, \widehat{\beta}, \widehat{\lambda}\right]$. Also, if we have this probability, then we can easily perform MLE estimation on the term shown in Equation 6.

The above suggests that we can have have an alternating approach to solving for the optimal β and λ and at the same time computing the posterior $\mathbf{w} \mid \mathbf{y}, X$ *i.e.* assume some α^0 and β^0, compute posterior in Equation 4 and then solve for Equation 6 iteratively. We give the algorithm for the alternating optimization below

Algorithm 1: Alternating Optimization for MLE-II

1. Initialize $\left\{\widehat{\beta}\right\}^0$ and $\left\{\widehat{\lambda}\right\}^0$
2. Repeat until convergence for t = 1, 2 ...
   (a) Estimate the posterior over $\{\mathbf{w}\}^t$ as

$$\mathbb{P}\left[\{\mathbf{w}\}^t \mid \mathbf{y}, X, \left\{\widehat{\beta}\right\}^{t-1}, \left\{\widehat{\lambda}\right\}^{t-1}\right] = \mathcal{N}\left(\{\mathbf{w}\}^t \mid \boldsymbol{\mu}, \Sigma\right)$$

   where $\boldsymbol{\mu}$ and Σ are as defined in Equation 4

   (b) Estimate $\widehat{\lambda}^t$ and $\widehat{\beta}^t$ as

$$\left\{\widehat{\lambda}\right\}^t = \frac{\gamma}{\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu}}$$
$$\left\{\widehat{\beta}\right\}^t = \frac{N-\gamma}{\sum_{i\in[n]}\left(y^i - \boldsymbol{\mu}^{\mathrm{T}}\mathbf{x}^i\right)^2}$$

   where γ is a function of the eigenvalues of Σ

Note that we do not discuss the intricacies of the algorithm or even the exact form of the algorithm, however the general idea of Alternating Optimization should be clear from this example.

Remark. The optimization function in Equation 4 is a convex function, and hence, we can expect that function to always converge. Therefore, we say that Alternating Optimization for this example will always converge to a minima (local or global)

<++>