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

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

SCRIBE

2

Parameter Estimation in Probabilistic Models

1. Point Estimate

Point estimate is a computationally cheaper alternative to fully Bayesian Inference. It gives the single best estimate to an optimization problem. There are two types of Point Estimates – Maximum Likelihood Estimation (MLE) and Maximum-a-Posteriori (MAP)

1.1 Point Estimation via MLE

MLE finds the estimate that maximizes the log-Likelihood $\left(\log\left(\mathbb{P}\left[\mathbf{X} \mid \boldsymbol{\theta}\right]\right)\right)$. We can write the maximization function as

$$\mathcal{L}\left[\boldsymbol{\theta}\right] \quad = \quad \log\left(\mathbb{P}\left[\mathbf{X} \mid \boldsymbol{\theta}\right]\right)$$

If the data points are independent, we can write this the joint probability as

$$\begin{aligned} \mathbb{P}\left[\mathbf{X} \mid \boldsymbol{\theta}\right] \quad &= \quad \prod_{n=1}^{N} \mathbb{P}\left[\mathbf{x}^n \mid \boldsymbol{\theta}\right] \\ \implies \mathcal{L}\left[\boldsymbol{\theta}\right] \quad &= \quad \sum_{n=1}^{N} \log\left(\mathbb{P}\left[\mathbf{x}^n \mid \boldsymbol{\theta}\right]\right) \end{aligned}$$

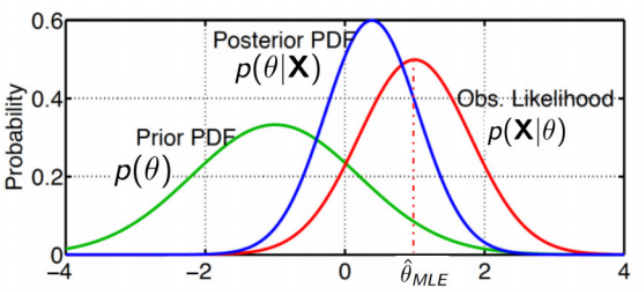

Figure 1: Visual Look at MLE Estimate

Note. MLE is consistent, i.e. as N $\longrightarrow \infty$ the value of $\widehat{\boldsymbol{\theta}}_{MLE}$ converges to the real $\boldsymbol{\theta}$

1.2 Point Estimate via MAP

MAP finds the estimate that maximizes the log-posterior-probability $\left(\log\left(\mathbb{P}\left[\boldsymbol{\theta} \mid \mathbf{X}\right]\right)\right)$. We can write the maximization function as

$$\begin{aligned}\mathcal{L}[\boldsymbol{\theta}] &= \log\left(\mathbb{P}\left[\boldsymbol{\theta} \mid \mathbf{X}\right]\right) \\ &= \log\left(\mathbb{P}\left[\mathbf{X} \mid \boldsymbol{\theta}\right]\right) + \log\left(\mathbb{P}\left[\boldsymbol{\theta}\right]\right)\end{aligned}$$

Again, if the data points are independent, we can write

$$\begin{aligned}\mathbb{P}\left[\mathbf{X} \mid \boldsymbol{\theta}\right] &= \prod_{n=1}^{N} \mathbb{P}\left[\mathbf{x}^n \mid \boldsymbol{\theta}\right] \\ \implies \mathcal{L}[\boldsymbol{\theta}] &= \sum_{n=1}^{N} \left(\log\left(\mathbb{P}\left[\mathbf{x}^n \mid \boldsymbol{\theta}\right]\right)\right) + \log\left(\mathbb{P}\left[\boldsymbol{\theta}\right]\right)\end{aligned}$$

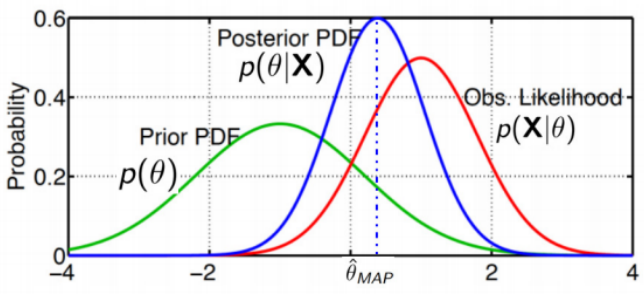

Figure 2: Visual Look at MAP Estimate

Note. When the prior is uniform, the MLE and MAP estimates are the same
MAP estimate is not Bayesian, as it still gives a point estimate (can be considered as optimization with regularization). We are still not modelling the uncertainty in the parameter estimation.

1.3 Point Estimate via Loss Minimization Function

Instead of maximizing the log-probability, we can also minimize the negative-log-probability. This is the loss function.

In case of both MLE and MAP, the loss function is the NLL (Negative Log Likelihood). For MLE, we minimize the following function

$$\mathrm{L}\ (\boldsymbol{\theta}) \quad = \quad \mathrm{NLL}\ (\boldsymbol{\theta}) \quad = \quad -\log\left(\mathbb{P}\left[\mathbf{X} \mid \boldsymbol{\theta}\right]\right)$$

In case of MAP, we minimize the loss function with a regularizer term

$$\mathrm{L}\,(\boldsymbol{\theta}) \quad = \quad \mathrm{NLL}\,(\boldsymbol{\theta}) - \log\left(\mathbb{P}\left[\boldsymbol{\theta}\right]\right)$$

Note. NLL is the 'loss function' and $-\log\left(\mathbb{P}\left[\boldsymbol{\theta}\right]\right)$ is the regularizer term

Thus, MLE is like emperical risk / loss minimization (ERM) and MAP is like regularized ERM.

1.4 Point Estimate Example: Coin-Toss

Consider a simple problem. We toss a coin N times. The task is to predict the next outcome. Assume consecutive coin tosses to be independent.

Say the outcome of toss is oru random variable x. Since there are two options, the logical choice is a Bernoulli distribution. Hence we can write the probability function of the random variable $x \in \{0, 1\}$ (where 1 represents 'Head' and 0 represents 'Tail') conditioned on the bias θ as follows

$$\mathbb{P}[x] \quad = \quad \theta^x (1-\theta)^{1-x}$$

Since this is the probability of our data points, we can write the log-likelihood as follows

$$\log\left(\mathbf{X} \mid \theta\right) \quad = \quad \sum_{n=1}^{N} x^n \log(\theta) + (1-x^n) \log(1-\theta)$$

Maximizing the log-likelihood, we get the MLE estimate $\left(\widehat{\theta}_{MLE}\right)$

$$\widehat{\theta}_{MLE} \quad = \quad \frac{\sum_{n=1}^{N} x^n}{N}$$

In this case, MLE is simply the fraction of heads. The problem with MLE is that it does not express our prior beliefs about θ. This can be problematic if the number of samples is very less (the number of heads might be zero)

To avoid this, we use the MAP estimate. Here, we are faced with another problem of inductive bias. We need to choose the prior distribution. In our case, we need to find a probability function for the random variable θ.

Since $\theta \in (0,1)$, we can assume a Beta prior

$$\mathbb{P}\left[\theta \mid \alpha, \beta\right] \quad = \quad \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}$$

Here, α and β are hyperparameters and are generally assumed to be known, however can also be estimated either using MLE-II or EM Method (Discussed Later). Now, for the MAP esimtation, we have the following minimizing term

$$L \quad = \quad \sum_{n=1}^{N} x^n \left(\log(\theta) + (1-x^n) \log(1-\theta)\right) + (\alpha-1) \log(\theta) + (\beta-1) \log(1-\theta)$$

Therefore, we have the MAP estimate as follows

$$\widehat{\theta}_{MAP} \quad = \quad \frac{\sum_{n=1}^{N} x^n + \alpha - 1}{N + \alpha + \beta - 2}$$

Intuitively, the hyperparameters represent the pseudo-observations *i.e.* $\alpha - 1$ and $\beta - 1$ are the number of heads and tails, respectively, before actually seeing any data.

2. Bayesian Inference

Although harder, for certain problems, we can compute the posterior probability of the parameters. This gives us a more complete picture than point estimates, as it allows us to model the uncertainty in the predicted value of the paramters.

The cases in which it is possible to do Bayesian Infernce are when the prior and the likelihood probabilities are *conjugate*. In other cases, we might be able to approximate the posterior using approximate Bayesian Inference methods such as MCMC and Variational Bayes.

2.1 Parameter Estimation via Bayesian Inference

We can use the Bayes Rule to compute the posterior distribution over the parameters

$$\mathbb{P}\left[\boldsymbol{\theta} \mid \mathbf{X}\right] \quad = \quad \frac{\mathbb{P}\left[\mathbf{X} \mid \boldsymbol{\theta}\right]\mathbb{P}\left[\boldsymbol{\theta}\right]}{\mathbb{P}\left[\mathbf{X}\right]}$$

The denominator *i.e.* $\mathbb{P}\left[\mathbf{X}\right]$ is the marginal likelihood. Since the marginal likelihood is independent of $\boldsymbol{\theta}$, we generally only compute the numerator and then integrate (either exact or approximate) it to compute the marginal likelihood.

$$\mathbb{P}\left[\mathbf{X}\right] \quad = \quad \int_{\boldsymbol{\theta}} \mathbb{P}\left[\mathbf{X} \mid \boldsymbol{\theta}\right]\mathbb{P}\left[\boldsymbol{\theta}\right] d\boldsymbol{\theta}$$

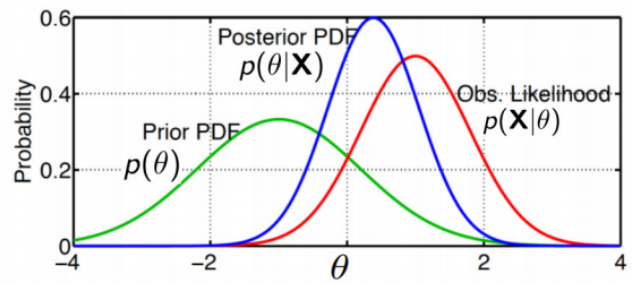

Figure 3: Visual Look at Bayesian Inference

In general, the posterior is hard to compute, as $\mathbb{P}\left[\mathbf{X}\right]$ can be intractable. It is, however, easier if the prior and the likelihood terms and conjugate.

2.2 Inferring the Full Posterior: Coin-Toss Example

The problem is the same as described earlier, and we use the same terms for the likelihood $\left(\text{Bernoulli}\left(x^n \mid \theta\right)\right)$ and the prior $\left(\text{Beta}\left(\theta \mid \alpha, \beta\right)\right)$.

The posterior can then be computed as follows

$$\begin{aligned}
\mathbb{P}\left[\theta \mid \mathbf{X}\right] \quad &= \quad \frac{\mathbb{P}\left[\mathbf{X} \mid \theta\right]\mathbb{P}\left[\theta\right]}{\mathbb{P}\left[\mathbf{X}\right]} \\
&\propto \quad \mathbb{P}\left[\mathbf{X} \mid \theta\right]\mathbb{P}\left[\theta\right] \\
&\propto \quad \theta^{\sum_{n=1}^{N} x^n + \alpha - 1}\left(1-\theta\right)^{\beta + N - 1 - \sum_{n=1}^{N} x^n}
\end{aligned}$$

After normalizing (or through normal inspection), we can say that the posterior is also Beta

$$\mathbb{P}\left[\theta \mid \mathbf{X}\right] \quad = \quad \text{Beta}\left(\theta \mid \sum_{n=1}^{N} x^n + \alpha - 1, \beta + N - 1 - \sum_{n=1}^{N} x^n\right)$$

Here, the posterior has the same form as the prior (both Beta): *property of conjugate priors*

2.3 Using the Posterior for Prediction

We need to predict the probability of getting a head in the next try. Essentially we need to predict the value of $\mathbb{P}\left[x = 1 \mid \mathbf{X}\right]$

In order to make predictions, we can use posterior averaging, *i.e.* compute the predicted probabilities (of getting a head) over all values of θ weighted on their posterior probability (Expected Value)

$$\begin{aligned}
\mathbb{P}\left[\, x = 1 \;\middle|\; \mathbf{X} \,\right] &= \int_{\theta} \mathbb{P}\left[\, x = 1 \;\middle|\; \theta, \mathbf{X} \,\right] \mathbb{P}\left[\, \theta \;\middle|\; \mathbf{X} \,\right] d\theta \\
&= \int_{\theta} \theta \cdot \mathbb{P}\left[\, \theta \;\middle|\; \mathbf{X} \,\right] d\theta \\
&= \mathop{\mathbb{E}}_{\mathbb{P}[\,\theta \,|\, \mathbf{X}\,]}\left[\, \theta \,\right] \\
&= \frac{\sum_{n=1}^{N} x^n + \alpha}{\alpha + \beta + N}
\end{aligned}$$

Hence, $x \sim \text{Bernoulli}\left(x \;\middle|\; \mathop{\mathbb{E}}_{\mathbb{P}[\,\theta \,|\, \mathbf{X}\,]}\left[\, \theta \,\right] \right)$

Note. The predicted distribution doesn't depend on only one point estimate, and is an expectation over the complete distribution

<++>

2.4 Examples of Conjugate Priors

Many pairs of distributions are conjugate to each other such as

- Bernoulli (likelihood) + Beta (prior) $\longrightarrow$ Beta posterior
- Binomial (likelihood) + Beta (prior) $\longrightarrow$ Beta posterior
- Multinomial (likelihood) + Dirichlet (prior) $\longrightarrow$ Dirichlet posterior
- Poisson (likelihood) + Gamma (prior) $\longrightarrow$ Gamma posterior
- Gaussian (likelihood) + Gaussian (prior) $\longrightarrow$ Gaussian posterior