Bayesian Machine Learning Review

Akhil Vasvani

March 2019

1 Questions

Exercise 1. What are the differences between "Bayesian" and "Frequentist" approach for Machine Learning?

Proof. **Frequentist statistics** and approaches are based on estimating a single value of $\boldsymbol{\theta}$, then making all predictions thereafter based on that one estimate. Another approach is to consider all possible values of $\boldsymbol{\theta}$ when making a prediction. The latter is the domain of **Bayesian statistics**.

The frequentist perspective is that the true parameter value $\boldsymbol{\theta}$ is fixed but unknown, while the point estimate $\hat{\boldsymbol{\theta}}$ is a random variable on account of it being a function of the dataset (which is seen as random). The Bayesian perspective on statistics is quite different. The Bayesian uses probability to reflect degrees of certainty of states of knowledge. The dataset is directly observed and so is not random. On the other hand, the true parameter $\boldsymbol{\theta}$ is unknown or uncertain and thus is represented as a random variable.

Now consider that we have a set of data samples $\{x^{(1)}, ..., x^{(m)}\}$. We can recover the effect of data on our belief about $\boldsymbol{\theta}$ by combining the data likelihood $p(x^{(1)}, ..., x^{(m)} \,|\, \boldsymbol{\theta})$ with the prior via Bayes' rule:

$$p(\boldsymbol{\theta} \,|\, x^{(1)}, ..., x^{(m)}) = \frac{p(x^{(1)}, ..., x^{(m)} \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(x^{(1)}, ..., x^{(m)})} \tag{1}$$

In the scenarios where Bayesian estimation is typically used, the prior—the **prior probability distribution**, $p(\boldsymbol{\theta})$—begins as a relatively uniform or Gaussian distribution with high entropy, and the observation of the data usually causes the posterior to lose entropy and concentrate around a few highly likely values of the parameters.

Relative to maximum likelihood estimation, Bayesian estimation offers two important differences. First, unlike the maximum likelihood approach that makes predictions using a point estimate of $\boldsymbol{\theta}$, the Bayesian approach is to make predictions using a full distribution over $\boldsymbol{\theta}$. For example, after observing m examples, the predicted distribution over the next data sample, $x^{(m+1)}$, is given by

$$p(x^{(m+1)} \,|\, x^{(1)}, ..., x^{(m)}) = \int p(x^{(m+1)} \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta} \,|\, x^{(1)}, ..., x^{(m)}) \; d\boldsymbol{\theta}. \tag{2}$$

Here each value of $\boldsymbol{\theta}$ with positive probability density contributes to the prediction of the next example, with the contribution weighted by the posterior density itself. After having observed $\{x^{(1)}, ..., x^{(m)}\}$, if we are still quite uncertain about the value of $\boldsymbol{\theta}$, then this uncertainty is incorporated directly into any predictions we might make.

The frequentist approach addresses the uncertainty in a given point estimate of by evaluating its variance. The variance of the estimator is an assessment of how the estimate might change with alternative samplings of the observed data. The Bayesian answer to the question of how to deal with the uncertainty in the estimator is to simply integrate over it, which tends to protect well against overfitting. This integral is of course just an application of the laws of probability, making the Bayesian approach simple to justify, while the frequentist machinery for constructing an estimator is based on the rather ad hoc decision to summarize all knowledge contained in the dataset with a single point estimate.

The second important difference between the Bayesian approach to estimation and the maximum likelihood approach is due to the contribution of the Bayesian prior distribution. The prior has an influence by shifting probability mass density towards regions of the parameter space that are preferred *a priori*. In practice, the prior often expresses a preference for models that are simpler or more smooth. Critics of the Bayesian approach identify the prior as a source of subjective human judgment impacting the predictions. □

Exercise 2. Compare and contrast maximum likelihood and maximum a posteriori estimation.

Proof. While the most principled approach is to make predictions using the full Bayesian posterior distribution over the parameter $\boldsymbol{\theta}$, it is still often desirable to have a single point estimate. One common reason for desiring a point estimate is that most operations involving the Bayesian posterior for most interesting models are intractable, and a point estimate offers a tractable approximation. Rather than simply returning to the maximum likelihood estimate, we can still gain some of the benefit of the Bayesian approach by allowing the prior to influence the choice of the point estimate. One rational way to do this is to choose the **maximum a posteriori** (MAP) point estimate. The MAP estimate chooses the point of maximal posterior probability (or maximal probability density in the more common case of continuous $\boldsymbol{\theta}$):

$$\boldsymbol{\theta}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta} \,|\, \boldsymbol{x}) = \arg\max_{\boldsymbol{\theta}} \log p(\boldsymbol{x} \,|\, \boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) \tag{3}$$

We recognize, above on the right hand side, $\log p(\boldsymbol{x} \,|\, \boldsymbol{\theta})$, i.e. the standard log-likelihood term, and $\log p(\boldsymbol{\theta})$, corresponding to the prior distribution.

As with full Bayesian inference, MAP Bayesian inference has the advantage of leveraging information that is brought by the prior and cannot be found in the training data. This additional information helps to reduce the variance in the MAP point estimate (in comparison to the ML estimate). However, it does so at the price of increased bias. □

Exercise 3. How does Bayesian methods do automatic feature selection?

Proof. Bayesian methods require the use of a prior distribution in order to estimate the posterior distribution. As a consequence, sometimes the posterior distribution might overfit (or underfit) the data and so a regularization term is added to the cost function in order to control the weights (penalizes certain weights depending on how strict you are with the regularization).

Feature selection is called "automatic" in the sense that the regularization term may drive to zero some weights when you minimize the cost function, without the need of selecting the features in a previous step. (In practice you may need to do it anyway.) □

Exercise 4. What do you mean by Bayesian regularization?

Proof. Many regularized estimation strategies, such as maximum likelihood learning regularized with weight decay, can be interpreted as making the MAP approximation to Bayesian inference. This view applies when the regularization consists of adding an extra term to the objective function that corresponds to $\log p(\boldsymbol{\theta})$.

Example. As an example, consider a linear regression model with a Gaussian prior on the weights $\boldsymbol{w}$. If this prior is given by $\mathcal{N}(\boldsymbol{w}; \boldsymbol{0}, \frac{1}{\lambda}\boldsymbol{I}^2)$, then the log-prior term in is proportional to the familiar $\lambda\boldsymbol{w}^\top\boldsymbol{w}$ weight decay penalty, plus a term that does not depend on $\boldsymbol{w}$ and does not affect the learning process. MAP Bayesian inference with a Gaussian prior on the weights thus corresponds to weight decay.

In particular, L^2 regularization is equivalent to MAP Bayesian inference with a Gaussian prior on the weights. For L^1 regularization, the penalty $\alpha\Omega(\boldsymbol{w}) = \alpha\sum_i |w_i|$ used to regularize a cost function is equivalent to the log-prior term that is maximized by MAP Bayesian inference when the prior is an isotropic Laplace distribution over $\boldsymbol{w} \in \mathbb{R}^n$:

$$\log p(\boldsymbol{w}) = \sum_i \log \text{Laplace}(w_i; 0, \frac{1}{\alpha}) = -\alpha||\boldsymbol{w}||_1 + n\log\alpha - n\log 2. \tag{4}$$

From the point of view of learning via maximization with respect to $\boldsymbol{w}$, we can ignore the $\log\alpha - \log 2$ terms because they do not depend on $\boldsymbol{w}$.

Note. Not all regularization penalties correspond to MAP Bayesian inference. For example, some regularizer terms may not be the logarithm of a probability distribution. Other regularization terms depend on the data, which of course a prior probability distribution is not allowed to do.

□

Exercise 5. When will you use Bayesian methods instead of Frequentist methods?

Proof. The key difference between Bayesian and frequentist approaches lies in the definition of a probability, so if it is necessary to treat probabilities strictly as a long run frequency then frequentist approaches are reasonable, if it is not then you should use a Bayesian approach.

Another way of putting it, is if you want to know what inferences you can draw from a particular experiment, you probably want to be Bayesian; if you want to draw conclusions about some population of experiments (e.g. quality control) then frequentist methods are well suited.

More here.

Note. Bayesian methods typically generalize much better when limited training data is available (or an large feature set), but typically suffer from high computational cost when the number of training examples is large.

□