Bayesian Machine Learning
Bayesian Model Comparison

1 The Bayesian approach to model comparison

We wish to compare a set of L *models* (or probability distributions over the observed data), that we'll note $\{\mathcal{M}_i\}_{i\in\{1,\dots,L\}}$.

Given a training set $\mathcal{D}$ and a prior belief $p(\mathcal{M}_i)$, Baye's rule writes :

$$p(\mathcal{M}_i \,|\, \mathcal{D}) \propto \underbrace{p(\mathcal{D} \,|\, \mathcal{M}_i)}_{\text{model evidence}} \; p(\mathcal{M}_i) \tag{1}$$

We call the quantity $p(\mathcal{D} \,|\, \mathcal{M}_i)$ the *model evidence* (or marginal likelihood), which is a measure a the preference shown by the data for the model $\mathcal{M}_i$.

Also, one can note that once the posterior distributions of model $p(\mathcal{M}_i \,|\, \mathcal{D})$ are known, the predictive distribution for new inputs writes :

$$p(t \,|\, x, \mathcal{D}) = \sum_{i=1}^{L} p(t \,|\, x, \mathcal{D}, \mathcal{M}_i) p(\mathcal{M}_i \,|\, \mathcal{D}) \tag{2}$$

by marginalizing over the models. Using such a prediction employs the mixture of expert distributions, and is known as *Bayesian model averaging*. We could also seek to pick the most appropriate model, procedure known as *model selection*.

2 The Evidence

For a parametric model, the evidence writes :

$$p(\mathcal{D} \,|\, \mathcal{M}_i) = \int_w p(\mathcal{D} \,|\, \mathcal{M}_i, w) p(w \,|\, \mathcal{M}_i) dw \tag{3}$$

From a sampling perspective, we are looking how well $\mathcal{D}$ could be generated when sampling from the prior distribution of the parameter.

2.1 Intuition

In the following, we'll omit the model $\mathcal{M}_i$ to keep notations uncluttered. To gain some insight on the evidence, let us assume that the posterior distribution is sharply peaked around its maximum value (w_{MAP}) with width Δw_{po}. The prior is supposed to be flat around a Δw_{pr} interval. Therefore the evidence approximates as :

$$p(\mathcal{D}) = p(\mathcal{D} \,|\, w_{MAP}) \frac{\Delta w_{po}}{\Delta w_{pr}} \tag{4}$$

and a first approximation of the log evidence writes :

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D} \,|\, w_{MAP}) + M \ln \frac{\Delta w_{po}}{\Delta w_{pr}} \tag{5}$$

if we have M parameters. This measure is starting to look like the BIC model selection criterion ! Hence the BIC can be related to evidence maximization for selecting a model. Indeed, the BIC simply adds a Gaussian approximation of the posterior to write in its full form.

2.2 Baye's factor

The ratio of two different model evidences $p(\mathcal{M}_i \,|\, D)/p(\mathcal{M}_i \,|\, D)$ is known as *Baye's factor*.

In the Bayesian model comparison framework, one assumes that the true model (i.e from which the data was generated) to be within the set of model under comparison. Provided this assumption, we'd like to show that the Bayesian approach averages models in favor of the true one. If we consider only $\mathcal{M}_1$ and $\mathcal{M}_2$, with the first being the true model. For a finite amount of data, it is possible for the Bayes factor to be largest for the wrong model. However, in the *large data limit*, and if we average the Baye's factor over the distribution of datasets, we obtained that the expected Baye's factor writes (transposing to the log likelihoods) :

$$\int_{\mathcal{D}} p(\mathcal{D} \,|\, \mathcal{M}_1) \frac{p(\mathcal{D} \,|\, \mathcal{M}_1)}{p(\mathcal{D} \,|\, \mathcal{M}_2)} d\mathcal{D} \tag{6}$$

Therefore, the expected Baye's factor writes simply as the *KL divergence* of p to q ! Using Jensen's inequality, one can easily show that this divergence is positive, hence giving most credit to the true model $\mathcal{M}_1$.

For reminders, $KL(p||q)$ is a distance (more precisely a divergence) between two p.d.f. It writes :

KL divergence

$$\begin{aligned} KL(p||q) &= \mathbb{E}_p \left[\ln \frac{p}{q} \right] \\ &= \int_\theta p(\theta) \ln \frac{p(\theta)}{q(\theta)} d\theta \end{aligned} \tag{7}$$

2.3 Evidence approximation or ML2

In the following, we'll consider the design matrix

$$X = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix} \in \mathcal{M}_{n,d}(\mathbb{R})$$

alongside the target vector $\boldsymbol{t} = \begin{pmatrix} t_1 & \dots & t_n \end{pmatrix}^T$.

Recall the Bayesian inference approach for linear regression :

- likelihood :
$$p(t \,|\, x, w) = \mathcal{N}\left(t \,\middle|\, w^T \phi(x), \beta^{-1}\right) \tag{8}$$
- prior :
$$p(w \,|\, \alpha) = \mathcal{N}\left(w \,\middle|\, 0, \alpha^{-1}\mathbb{I}\right) \tag{9}$$

In a fully Bayesian approach, we should provide α and β with prior distribution and make predictions by marginalizing over them. This however leads to intractable computations. Indeed, if we introduce hyperpriors for α and β, the predictive distribution would write :

$$p(t \,|\, \boldsymbol{t}, x, X) = \iiint_{\alpha,\beta,w} p(t \,|\, w, x, \beta) p(w \,|\, \alpha, X, \boldsymbol{t}) p(\alpha, \beta \,|\, \boldsymbol{t}, X) \mathrm{d}w \mathrm{d}\alpha \mathrm{d}\beta \tag{10}$$

The most logical approximation to simplify this expression is to set hyper-parameters to values maximizing the marginal likelihood. This framework is known as *empirical Bayes*, *type-II maximum likelihood* or *the evidence approximation*.
If we now assume that $p(\alpha, \beta \,|\, \boldsymbol{t}, X)$ - which is the posterior distribution over the hyper-parameters - is sharply peaked around $\hat{\alpha}, \hat{\beta}$ (maximal values), the predictive distribution will now write :

$$p(t \,|\, \boldsymbol{t}, x, X) = \int_w p(t \,|\, w, \hat{\beta}) p(w \,|\, \boldsymbol{t}, \hat{\alpha}) dw \tag{11}$$

Since according to Baye's rule :

$$p(\alpha, \beta \,|\, \boldsymbol{t}) \propto p(\alpha, \beta) p(\mathbf{t} \,|\, \alpha, \beta)$$

if we assume a flat prior $p(\alpha, \beta)$, we therefore define :

$$(\hat{\alpha}, \hat{\beta}) = \underset{\alpha,\beta}{\operatorname{argmax}} \{p(\boldsymbol{t} \,|\, \alpha, \beta)\} \tag{12}$$

In the end, we choose to assign the hyper-parameters to the values that **maximize the evidence**.

Let's evaluate the model evidence for the linear model. We have that :

$$p(\boldsymbol{t} \,|\, \alpha, \beta) = \int_w p(\boldsymbol{t} \,|\, w, \beta) p(w \,|\, a) dw \tag{13}$$

Using (8) and (9) we have that :

$$p(\boldsymbol{t} \,|\, \alpha, \beta) = \left(\frac{\beta}{2\pi}\right)^{n/2} \left(\frac{\alpha}{2\pi}\right)^{d/2} \int_w \exp(-E(w)) dw \tag{14}$$

where we have that

$$\begin{aligned} E(w) &= \frac{\beta}{2} \|\boldsymbol{t} - \boldsymbol{\phi} w\|^2 + \frac{\alpha}{2} w^T w \\ &= \frac{\beta}{2} \boldsymbol{t}^T \boldsymbol{t} + \frac{1}{2} w^T \left(\alpha + \beta \boldsymbol{\phi}^T \boldsymbol{\phi}\right) w + \beta^{-1} \boldsymbol{t}^T \boldsymbol{\phi} w \end{aligned} \tag{15}$$

Let us write :

$$\begin{cases} \Sigma^{-1} = \alpha \mathbb{I}_d + \beta \boldsymbol{\phi}^T \boldsymbol{\phi} \\ \mu = \beta \Sigma \boldsymbol{\phi}^T \boldsymbol{t} \end{cases} \tag{16}$$

Then :

$$\begin{aligned} E(w) &= \frac{1}{2}(w - \mu)^T \Sigma^{-1} (w - \mu) + \frac{\beta}{2} \boldsymbol{t}^T \boldsymbol{t} - \frac{1}{2} \mu^T \Sigma^{-1} \mu \\ &= E(\mu) + \frac{1}{2}(w - \mu)^T \Sigma^{-1} (w - \mu) \end{aligned} \tag{17}$$

where we defined :

$$E(\mu) = \frac{\beta}{2} \|\boldsymbol{t} - \boldsymbol{\phi} \mu\|^2 + \frac{\alpha}{2} m_N^T m_N \tag{18}$$

In the end, we obtain for the full expression of the evidence :

Evidence (linear model)

$$\ln p(\boldsymbol{t} \,|\, \alpha, \beta) = \frac{n}{2} \ln \beta + \frac{d}{2} \ln \alpha - E(\mu) - \frac{1}{2} \ln |\Sigma^{-1}| - \frac{n}{2} \ln 2\pi \tag{19}$$

Let's maximize the evidence with respect to α. We'll write $\{u_i\}_i \in \mathbb{R}^d$ the eigen values of $\beta \boldsymbol{\phi}^T \boldsymbol{\phi}$:

$$\beta \boldsymbol{\phi}^T \boldsymbol{\phi} u_i = \lambda_i u_i, \quad \forall i \in \{1, \dots, d\} \tag{20}$$

Therefore, the precision matrix $A = \Sigma^{-1}$ has eigen values $(\alpha + \lambda_i)_i$. Since

$$\ln |A| = \sum_{i=1}^{d} \ln (\alpha + \lambda_i)$$

we'll have that

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} \ln |A| = \sum_{i=1}^{d} \frac{1}{\alpha + \lambda_i} \tag{21}$$

Therefore the stationary points of the evidence function satisfy :

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\alpha} p(\boldsymbol{t} \,|\, \alpha, \beta) \quad &= 0 \\ \Leftrightarrow \quad \frac{d}{2\alpha} - \frac{1}{2} \mu^T \mu - \frac{1}{2} \sum_{i=1}^{d} \frac{1}{\lambda_i + \alpha} &= 0 \end{aligned} \tag{22}$$

Therefore we have :

$$\begin{aligned}\alpha\mu^T\mu &= d - \sum_{i=1}^{d} \frac{\alpha}{\lambda_i + \alpha} \\ &= \sum_{i=1}^{d} \frac{\lambda_i}{\lambda_i + \alpha} \\ &= \gamma\end{aligned} \tag{23}$$

which we'll sum um in the *implicit* equation :

$$\boxed{\alpha = \frac{\gamma}{\mu^T\mu}} \tag{24}$$

This equation is of course not solvable in place (both μ and γ are functions of α). It however leads to a iterative solving of the evidence maximization (EM like algorithm). The same reasoning can be held for the determination of β.

2.3.1 Effective number of parameters

In directions where $\lambda_i >> \alpha$, the parameter w_i will be close to its maximal likelihood value. It is said to be *well-determined* (i.e tightly constrained by the data). Conversely, in directions for which $\lambda_i << \alpha$, we'll have that the likelihood would be relatively insensitive to α.

Actually,

$$\gamma = \sum_{i=1}^{d} \frac{\lambda_i}{\alpha + \lambda_i}$$

is a measure of the effective total number of well determined parameters. If we now consider the large data limit ($n >> d$), it is likely that all the parameters will be determined by the data as :

$$\boldsymbol{\phi}^T\boldsymbol{\phi} = \sum_i \phi_i\phi_i^T$$

which is a sum of rank one matrixes.

3 Model comparison and BIC

Consider the Laplace approximation for incomputable posteriors : we wish to have an approximation of the normalizing constant Z. We will approximate the posterior by a Gaussian centered around its maximum-a-posteriori value.

$$\begin{aligned}Z &= \int_z f(z)dz \\ &\simeq f(z_0)\int_z \exp\left(-\frac{1}{2}(z-z_0)^T A(z-z_0)\right)dz \\ &\simeq f(z_0)\frac{(2\pi)^{d/2}}{|A|^{1/2}}\end{aligned} \tag{25}$$

Now if we consider a dataset $\mathcal{D}$ with a set of models $\{\mathcal{M}_i\}_i$ having parameters $\{\theta_i\}$. Each model evidence writes :

$$p(D\,|\,\mathcal{M}_i) = \int_{\theta_i} p(D\,|\,\theta_i, \mathcal{M}_i)p(\theta_i\,|\,\mathcal{M}_i)d\theta_i \tag{26}$$

If we take its Laplace approximation we obtain (omitting the model reference for keeping uncluttered notations) :

$$\ln p(\mathcal{D}) = \ln p(\mathcal{D}\,|\,\theta_{MAP}) + \underbrace{\ln p(\theta_{MAP}) + \frac{M}{2}\ln 2\pi - \frac{1}{2}\ln|A|}_{\text{Occam's factor}} \tag{27}$$

where we took

$$A = -\nabla\nabla \ln p(\mathcal{D}\,|\,\theta_{MAP})p(\theta_{MAP}) \tag{28}$$

The equation (27) is often roughly approximated by considering a flat prior and a full-rank Hessian. This gives rise to the **Bayesian Information Criterion** :

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D} \,|\, \theta_{MAP}) - \frac{1}{2} M \ln N \tag{29}$$

which is often used when it comes to easily comparing models, based on their performance (likelihood) and their complexity (Occam's factor).