Bayesian Model Selection, the Marginal Likelihood, and Generalization

Sanae Lotfi Pavel Izmailov Gregory Benton Micah Goldblum Andrew Gordon Wilson

NYU

Paper Code

Overview

How do we select between hypotheses that are entirely consistent with any observations?

The marginal likelihood, which represents the probability of generating our data from a prior, provides an answer that encodes Occam's razor.

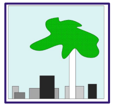
$$p(D\,|\mathcal{M}) = \int p(D|w,\mathcal{M})\, p(w|\mathcal{M})\, dw$$

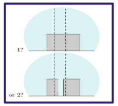
We fundamentally re-evaluate whether the log marginal likelihood (LML) is the right metric for predicting the generalization of trained models and hyper learning, and pursue a conditional marginal likelihood alternative.

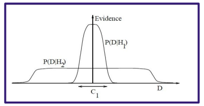
The marginal likelihood encodes Occam's razor

Occam's Razor: *"we should accept the **simplest** explanation that fits the data" [MacKay 2003].*

The marginal likelihood is a **normalized** probability density: the **most constrained** model covering the dataset wins *[MacKay 2003]*.



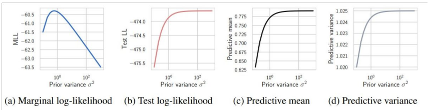




The marginal likelihood is NOT generalization

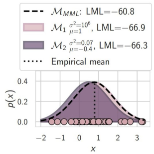
Density estimation example: $x \sim \mathcal{N}(u,1), u \sim \mathcal{N}(\mu,\sigma^2)$

The marginal likelihood can have a strong preference between models with identical predictive distributions, due to its **sensitivity to the prior.**

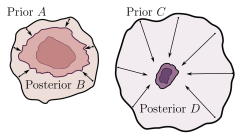


(a) Marginal log-likelihood (b) Test log-likelihood (c) Predictive mean (d) Predictive variance

Therefore the marginal likelihood can easily favour a model with a worse generalization performance:



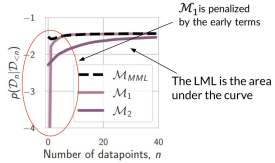
The marginal likelihood penalizes diffuse priors



The marginal likelihood heavily penalizes models where the posterior after observing data is much more concentrated than the prior, even if the posterior generalizes well.

The LML can be decomposed as follows: $\log p(D\,|\mathcal{M}) = \sum_{i=1}^{n} \log p(D_i|D_{<i},\mathcal{M})$

The first terms in this decomposition are irrelevant to generalization but have a large effect on the LML.

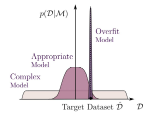


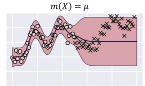
The marginal likelihood can overfit

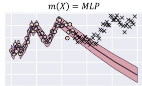
The LML can favor a prior around a **severely overfit** maximum likelihood solution.

The LML can overfit by ignoring uncertainty

$f(X) \sim \mathcal{N}(m(X), k(X,X));\ k(x,x') = \exp\left(-\frac{1}{2\ell^2}\|x-x'\|^2\right)$





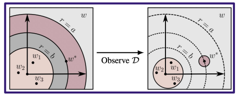


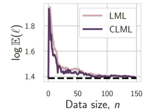
The marginal likelihood can underfit

Underfitting to avoid supporting bad solutions.

Underfitting in the function space

$f(X) \sim \mathcal{N}(m(X), k(X,X));\ k(x,x') = \exp\left(-\frac{1}{2\ell^2}\|x-x'\|^2\right)$





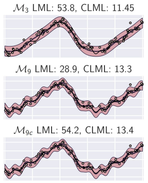
The conditional marginal likelihood

We form a posterior over a subset of the data and use it as a prior to compute LML for the rest of the data, resulting in the conditional log marginal likelihood (CLML):

$$\log p(D_{\geq m}|D_{<m},\mathcal{M}) = \sum_{i=m}^{n} \log p(D_i|D_{<i},\mathcal{M})$$

- Equivalent to removing the early terms in the LML decomposition.
- Has not been used for hyper learning, approx. inference, or underfitting.

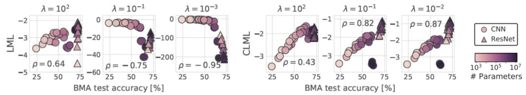
The CLML is more aligned with generalization



- Fourier model: $f(x,a,b) = \sum_{d=1}^{D} a_d \sin(d\cdot x) + b_d \cos(d \cdot x)$
- LML prefers degree-3 $\mathcal{M}_3$ model that underfits and heavily penalizes the degree-9 model
- LML assigns a high likelihood to $\mathcal{M}_{9c}$, a degree-9 model with a modified prior which makes identical predictions to $\mathcal{M}_9$.
- CLML correctly prefers $\mathcal{M}_9$ to $\mathcal{M}_3$.

The CLML for neural architecture search

- The LML is not always aligned with generalization.
- The CLML is aligned with generalization for all prior precisions!



The CLML for large-scale hyperparameter learning

The CLML is more effective for deep kernel hyperparameter learning, especially in the low data regime.