Contents

- If we hold the learning rate fixed and increase the batch size, the test accuracy usually falls.

1 Bayesian Model Comparison [1]

1.1 problem settings

- use probabilities to represent uncertainty in the choice of model.
- suppose we wish to compare a set of models: $\{\mathcal{M}_i\}_i^L$
  - model refers to a probability distribution over the observed data $\mathcal{D}$.
- data is generated from one of these models, but we are uncertain which one.
- the uncertainty is expressed through a prior distribution $p(\mathcal{M}_i)$
- given a training set $\mathcal{D}$, we then hope to evaluate the posterior distribution:

$$p(\mathcal{M}_i|\mathcal{D}) \propto p(\mathcal{M}_i)p(\mathcal{D}|\mathcal{M}_i) \tag{1}$$

- the term $p(\mathcal{D}|\mathcal{M}_i)$ is called model evidence, sometimes also called marginal likelihood, which expressed the preference shown by the data for different models. It can be viewed as a likelihood function over the space of models.

1.2 a simple approximation to the evidence

- for a model governed by a set of parameters $\mathbf{w}$, from the sum and product rule of probability, the model evidence is given by:

$$p(\mathcal{D}|\mathcal{M}_i) = \int p(\mathcal{D}|\mathbf{w}, \mathcal{M}_i)p(\mathbf{w}|\mathcal{M}_i)d\mathbf{w} \tag{2}$$

  from a sampling perspective, the evidence can be viewed as the probability of generating the data set $\mathcal{D}$ from a model whose parameters are sampled from the prior.

- Let omit the dependence on model $\mathcal{M}_i$ to keep the notation uncluttered, then we have:

$$p(\mathcal{D}) = \int p(\mathcal{D}|\mathbf{w})p(\mathbf{w})d\mathbf{w} \tag{3}$$

- Let's consider a simple approximation to gain some insights:
  1. the model has one parameter w
  2. assume $p(\mathcal{D}|\mathbf{w})$ is sharply peaked around the most probable value at w_{MAP} with width $\triangle w_{posterior}$
  3. assume $p(\mathbf{w})$ is flat with width $\triangle w_{prior}$, so that $p(\mathbf{w}) = \frac{1}{\triangle w_{prior}}$
- then, the integral canbe approximated by the value of the integrand at its maximum times the width of the peak, we get:

$$p(\mathcal{D}) \backsimeq p(\mathcal{D}|w_{MAP})\frac{\triangle w_{posterior}}{\triangle w_{prior}}$$

$$\ln p(\mathcal{D}) \backsimeq \ln p(\mathcal{D}|w_{MAP}) + \ln\frac{\triangle w_{posterior}}{\triangle w_{prior}}$$

1.3 insights from Bayesian model comparison

- for a model has M parameter, we can make a similar approximation. Suppose each parameter has the same ratio $\frac{\triangle w_{posterior}}{\triangle w_{prior}}$, then we can get:

$$\ln p(\mathcal{D}) \backsimeq \ln p(\mathcal{D}|\mathbf{w}_{MAP}) + M \ln \frac{\triangle w_{posterior}}{\triangle w_{prior}} \tag{4}$$

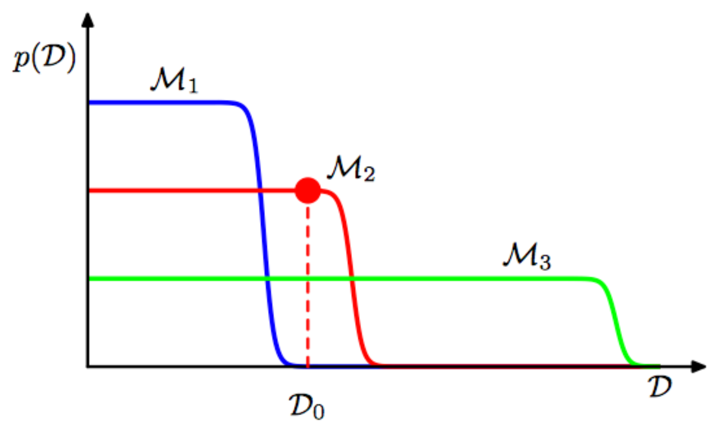
1. the first term represents the fit the data given by the most probable value.
2. the second penalizes the model according to its complexity.
   - $\triangle \mathbf{w}_{posterior} < \mathbf{w}_{prior}$, the second term is negative.

1.4 how the evidence favors intermediate complexity

Let's take an example: imagine running the models generatively to produce example datasets:

1. step1: choose the values of parameters from their prior distribution.
2. step2: for these parameter values, sample data from $p(\mathcal{D}|\mathbf{w})$

Figure 3.13 Schematic illustration of the distribution of data sets for three models of different complexity, in which $\mathcal{M}_1$ is the simplest and $\mathcal{M}_3$ is the most complex. Note that the distributions are normalized. In this example, for the particular observed data set $\mathcal{D}_0$, the model $\mathcal{M}_2$ with intermediate complexity has the largest evidence.



from the above figure.

- A simple model:
  - has little variability, the data generated are very similar to each other.
  - its distribution $p(\mathcal{D})$ is confined to a small region of the horizontal axis.
- A complex model:
  - can generate a variety of different data.
  - its distribution $p(\mathcal{D})$ is spread over a large region of the horizontal axis.

Essentially:

1. **the simple model cannot fit data well.**
2. **the complex model spreads its predictive probability over too broad a range of data sets and so assigns a relatively small probability to any one of them.**

- Some notes
  - the Bayesian framework assumes that the true distribution from which the data generated are contained in within the set of models under consideration.
  - provided this, the **Bayesian model comparison will on average favor the correct model.**.

1.5 Reference

1. chapter 3.4 of Pattern recognition and machine learning.

2. Wilson D R, Martinez T R. The general inefficiency of batch training for gradient descent learning[J]. Neural Networks, 2003, 16(10): 1429-1451.
3. Hardt M, Recht B, Singer Y. Train faster, generalize better: Stability of stochastic gradient descent[J]. arXiv preprint arXiv:1509.01240, 2015.
4. Smith S L, Kindermans P J, Le Q V. Don't Decay the Learning Rate, Increase the Batch Size[J]. arXiv preprint arXiv:1711.00489, 2017.
5. Smith S L, Le Q V. A Bayesian perspective on generalization and stochastic gradient descent[C]//International Conference on Learning Representations. 2018.