PRML studies(Ch3.4, 3.5)

Ch3.4 Bayesian Model Comparison

3.4 Bayesian Model Comparison

- Main consideration of Model comparison
  1. Compare between different model families (ex. SVM vs CNN)
  2. Compare between same model families with different hyperparameter values/options (ex. Branch number in Decision Tree based models, model architecture in Neural Network models, scale of weight norm penalty in parametric models)

- Primary Model selection methods in frequentist approach:
  1. Criteria based: AIC, BIC, Mallows's Cp… compare models by calculating [(Train data Negative Log Likelihood) – (number of parameters or model complexity penalty)] form criteria for each model. But this may yield results selecting overly simple models.
  2. Holdout Method/Cross Validation: Check the model's generalizability by calculating likelihood or performance metric of separated(independent) data set different from the data which model has trained.

3.4 Bayesian Model Comparison

- In Bayesian approach, we use probabilities to represent uncertainties in the choice of models.
- Suppose we wish to compare a set of L models $\{M_i\}$ where $i = 1, 2, \ldots, L$. Here, model refers to a probability distributions over observed data(training data) D.
- We think that data is generated from one of these models, but we are uncertain which one. And our uncertainty is expressed through a prior probability $p(M_i)$.
- First, we evaluate posterior distribution $p(M_i|D)$.
- Dropping constant term with respect to M_i,

$$p(M_i|D) \propto p(M_i)p(D|M_i)$$

*Note that $p(D|M_i)$ is called “model evidence” or sometimes “marginal likelihood”

*And the ratio of model evidences $p(D|M_i)/p(D|M_j)$ is known as “Bayes factor”

3.4 Bayesian Model Comparison

- Once we know the posterior distribution $p(M_i|D)$, the predictive distribution in supervised tasks for new unobserved target variables can be represented as:

$$p(t|\underline{x}, D) = \sum_{i=1}^{L} p(t|\underline{x}, M_i, D)p(M_i|D)$$

- It is just a weighted average of L distributions weighted by posterior probability.
- However, its distributional form can be multi-modal.
- Simple approximation to this model averaging is to use single most probable model alone to make predictions.

  => *model selection*

- Or, we can combine multiple models to make predictions.

  => *model averaging, simplest forms of model ensemble methods*

3.4 Bayesian Model Comparison

- For a model governed by a set of parameters w, model evidence can be represented as:

$$p(D|M_i) = \int p\left(D\middle|\underline{w}, M_i\right) p\left(\underline{w}\middle|M_i\right) dw$$

- Evaluating posterior distribution over parameters by Bayes formula:

$$p\left(\underline{w}\middle|D, M_i\right) = \frac{p\left(D\middle|\underline{w}, M_i\right) p(\underline{w}|M_i)}{p(D|M_i)}$$

- Thus, from sampling perspective, the marginal likelihood can be viewed as the probability of generating the data set D from a model whose parameters are sampled from prior at random.

3.4 Bayesian Model Comparison

- Also, we can analyze about model evidence to gain insights more by making simple approximation to the integral over parameters.
- If we assume posterior distribution is sharply peaked around the most probable value w_{MAP}, with width $\Delta w_{posterior}$, then we can approximate the integral by the value of the integrand at its maximum times the width of the peak.
- If we further assume that prior is flat with width Δw_{prior} so that $\mathrm{p(w)} = 1/\Delta w_{prior}$,

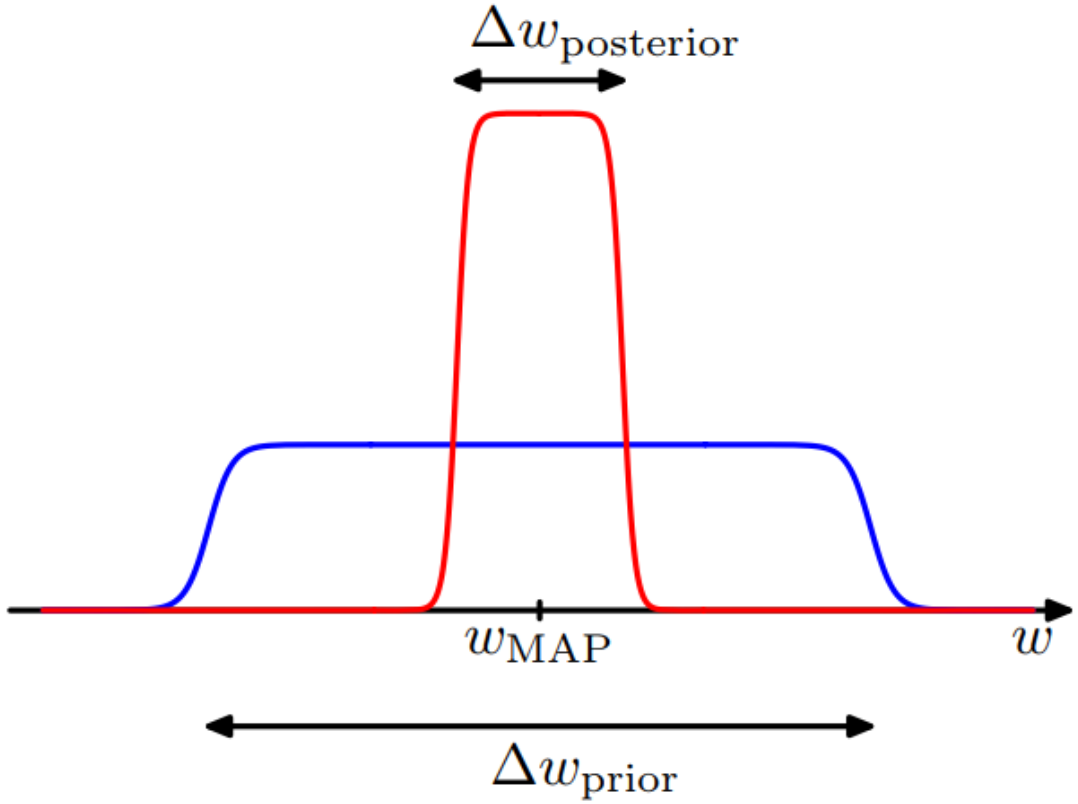
$$p(D) = \int p(D|w)p(w)dw \cong p(D|w_{MAP})\frac{\Delta w_{posterior}}{\Delta w_{prior}}$$

$$\ln p(D) \cong \ln p(D|w_{MAP}) + \ln \frac{\Delta w_{posterior}}{\Delta w_{prior}}$$

- First term(log likelihood) represents the fit to data given by the most probable parameter value and second term penalizes the model according to its complexity.

3.4 Bayesian Model Comparison

Figure 3.12 We can obtain a rough approximation to the model evidence if we assume that the posterior distribution over parameters is sharply peaked around its mode w_{MAP}.



3.4 Bayesian Model Comparison

- For models which has M parameters, by using same procedures, we can get:

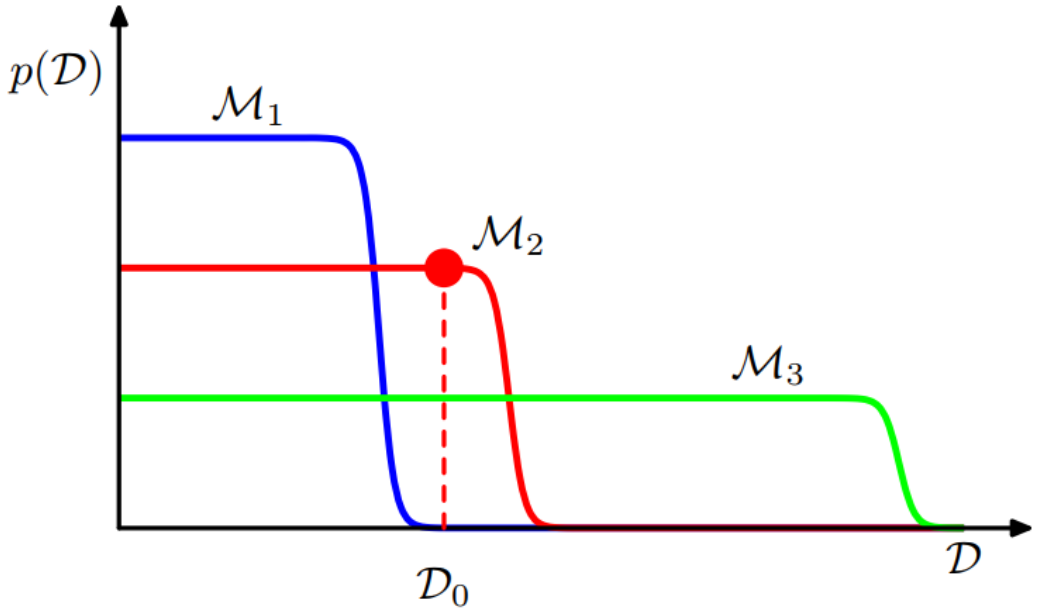
$$\ln p(D) \cong \ln p\left(D\middle|\underline{w}_{MAP}\right) + M \ln \frac{\Delta \underline{w}_{posterior}}{\Delta \underline{w}_{prior}}$$

- If we increase the complexity of model, the first term will typically decrease, because a more complex model is better able to fit the data, where second term will increase due to the dependence on M.

- Therefore, the optimal model complexity, as determined by the maximum evidence will be given by a trade-off between these two competing terms.

3.4 Bayesian Model Comparison

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.



- Simple models tends to have little variability and complex model have large variability in data generation.
- In the figures, for a specific data set D_0, simple models might not generate D_0.
- By contrast, complex models might generate D_0, but probabilities might be relatively low.

3.4 Bayesian Model Comparison

- Implicit Bayesian model comparison framework is the assumption that true data generating distribution is contained within the set of models under consideration.
- Provided that, we can show Bayesian model comparison framework will on average favor the correct model.
- Consider two models M_1 and M_2 in which M_1 correspond to truth.
- For a given finite data set, it is possible for the Bayes factor to be larger for the incorrect model.
- However, if we average the Bayes factor over the true distribution of data sets, we obtain the expected Bayes factor in the form:

$$\int p(D|M_1) \ln \frac{p(D|M_1)}{p(D|M_2)} dD$$

3.4 Bayesian Model Comparison

$$\int p(D|M_1) \ln \frac{p(D|M_1)}{p(D|M_2)} dD$$

- This quantity is an example of “Kullback-Leibler Divergence”
- And it satisfies the mathematical property of always being positive, unless the two distributions are equal in which case it is zero.
- Thus, the Bayes factor will always favor the correct model on average.

3.4 Bayesian Model Comparison

- In conclusion, we have seen Bayesian framework avoids over-fitting and allows models to be compared with the training data alone.
- However, like any other approach to pattern recognition, Bayesian approach needs to make assumption about the form of the model.
- If the assumptions are invalid, then the result can be misleading.

- Therefore, In practical application, it will be wise to keep aside an independent test set of data on which to evaluate the overall performance of the final system.

Ch3.5 The Evidence Approximation

3.5 The Evidence Approximation

- In fully Bayesian treatment, we would introduce prior distributions over the hyperparameters and make predictions by marginalizing with respect to these hyperparameters as well as parameters.
- However, although we can integrate over parameters or hyperparameters individually, the complete marginalization over all these variables is analytically intractable.
- By setting hyperparameters to specific values determined by maximizing the "marginal likelihood function" obtained by first integrating over the parameters, we can make approximation of the predictive distribution.

*This framework is known in the statistics literature as "Empirical Bayes" or "Generalized maximum likelihood".

*In Machine Learning literature, it is also called "Evidence Approximation"

3.5 The Evidence Approximation

- If we introduce hyperpriors $p(\underline{a})$ over hyperparameters $\underline{a}$, the predictive distribution $p(t^*|x^*, D)$ is obtained by marginalizing over parameters $\underline{w}$ and hyperparameters $\underline{a}$. (* marks new data)

$$p(t^*|\underline{x^*}, D) = \iint p(t^*|\underline{x^*}, \underline{w})p(\underline{w}|D, \underline{a})p(\underline{a}|D)d\underline{w}d\underline{a}$$

*Note that there might be some dependencies in some hyperparameters. For example, in linear basis models, noise precision β can be introduced and $p(t^*|\underline{x^*}, \underline{w})$ can be replaced with $p(t^*|\underline{x^*}, \underline{w}, \beta)$.

- From the Bayes theorem, since model evidence is posterior distribution for hyperparameter $\underline{a}$ and it is given by:

$$p(\underline{a}|D) \propto p(D|\underline{a})p(\underline{a})$$

3.5 The Evidence Approximation

- If the posterior distribution $p(\underline{a}|D)$ is sharply peaked around values $\underline{\hat{a}}$, then the predictive distribution is obtained by marginalizing over $\underline{w}$ in which $\underline{a}$ are fixed to the values $\underline{\hat{a}}$, so that

$$p(t^*|\underline{x}^*, D) \cong p(t^*|\underline{x}^*, D, \underline{\hat{a}}) = \int p(t^*|\underline{w}, D)\, p(\underline{w}|D, \underline{\hat{a}}) d\underline{w}$$

- If the prior is relatively flat, then in the evidence framework the values of $\underline{\hat{a}}$ are obtained by maximizing the marginal likelihood function $p(D|\underline{a})$.
- This will allow as to determine values for hyperparameters from training data alone, without recourse to Holdout methods or Cross-Validation.
- Note that there are two approaches that we can take to the maximization of the log evidence, which is:
  1. Evaluate evidence function analytically and then set its derivative equal to zero to obtain re-estimation equations for hyperparameters.
  2. Use a technique called Expectation Maximization algorithm.

  *These two approaches converge to the same solution.

3.5 The Evidence Approximation

- However, there might exists a practical alternative to the evidence framework called Laplace approximation.
- Laplace approximation is based on the local Gaussian approximation centered on the mode of the posterior distribution.
- However, the integrand as a function of parameters typically has a strong skewed mode so that the Laplace approximation fails to capture the bulk of probability mass, leading to poorer results than those obtained by maximizing the evidence.

Reference

- Figures and contents are from Pattern Recognition and Machine Learning, Bishop