BAYESIAN MODEL CLASS SELECTION

Introduction:

We must first ask a question before starting the topic of model class selection:

What constitutes a good model class?

Option 1: A good model class is a set of models such that there exists a model in the set that can fit the data well (we will state this as "the model class fits data well" for convenient later). Does this sound right? This cannot be right! Consider a case where we would like to fit a curve to 10 data points. We all know that the model class of ninth-order polynomials can fit the data. But this makes no sense since they overfit the data. In fact, to fit the data well is only necessary for a good model class but not sufficient.

Option 2: A good model class should (1) fit the data well and (2) predict unseen testing data well. Now it sounds right.

Notation change: the data is now *D*. Based on the above discussion, the "score" of a model class M should depend on two sub-scores: (1) the score of fitting training data $\hat{D}_1$: $score_1\left(\hat{D}_1 \mid M\right)$ and (2) the score of predicting testing data $\hat{D}_2$ after trained by $\hat{D}_1$: $score_2\left(\hat{D}_2 \mid \hat{D}_1, M\right)$. The final score of M, denoted by $S\left(\hat{D}_1, \hat{D}_2 \mid M\right)$, should monotonically depend on $score_1\left(\hat{D}_1 \mid M\right)$ and $score_2\left(\hat{D}_2 \mid \hat{D}_1, M\right)$.

Score examples:

Example 1:

$$score_1\left(\hat{D}_1 \mid M\right) = -\left(training\ \ error\ \ of\ \ M\ \ to\ \ \hat{D}_1\right)^2$$

$$score_2\left(\hat{D}_2 \mid \hat{D}_1, M\right) = -\left(testing\ \ error\ \ of\ \ "M\ \ trained\ \ by\ \ \hat{D}_1"\ \ to\ \ \hat{D}_2\right)^2$$

$$S\left(\hat{D}_1, \hat{D}_2 \mid M\right) = score_1 + score_2$$

Note: $S\left(\hat{D}_1, \hat{D}_2 \mid M\right) \neq S\left(\hat{D}_2, \hat{D}_1 \mid M\right)$. Does this sound right?

Example 2:

$$score_1\left(\hat{D}_1 \mid M\right) = f\left(\hat{D}_1 \mid M\right)$$

$$score_2\left(\hat{D}_2 \mid \hat{D}_1, M\right) = f\left(\hat{D}_2 \mid \hat{D}_1, M\right)$$

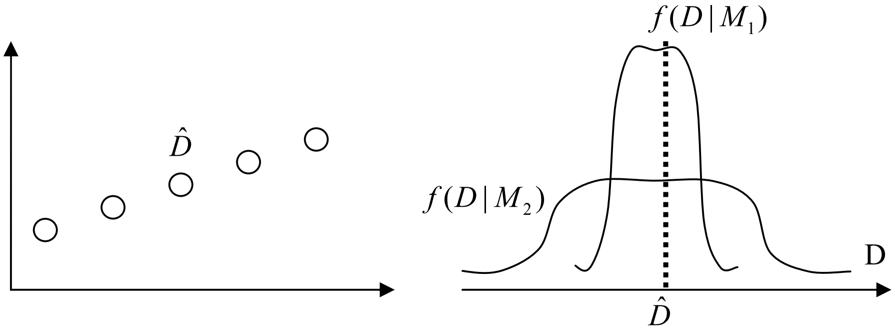
$$S\left(\hat{D}_2 \mid \hat{D}_1, M\right) = f(\hat{D}_1, \hat{D}_2 \mid M) = f(\hat{D} \mid M) = score_1 \times score_2$$

Note that now $S\left(\hat{D}_1,\hat{D}_2 \mid M\right)=S\left(\hat{D}_2,\hat{D}_1 \mid M\right)=f(\hat{D} \mid M)$. In the following, we choose $f(\hat{D} \mid M)$ to quantify how good M is.

Remarks:

1. When computing $f(\hat{D} \mid M)$, in reality we don't need to divide the data $\hat{D}$ into $\hat{D}_1$ and $\hat{D}_2$: the principle of training and testing has been built into $f(\hat{D} \mid M)$ even though we don't really do training and testing explicitly. Moreover, we don't need to worry about how to divide the data $\hat{D}$ into training and testing data because it is clear that no matter how we make the division, the resulting score $f(\hat{D} \mid M)$ will be the same.

2. The score $f(\hat{D} \mid M)$ is called the evidence of M. Sometimes it is called the marginal likelihood of M.
3. Given two model classes that can fit the data $\hat{D}$ well, the simpler model class will often have higher evidence. This can be seen using the following illustration: consider that we would like to fit a data set $\hat{D}$ that looks like a straight line using two model classes: (1) M_1 consists of all straight lines and (2) M_2 consists of all 2nd order polynomials. Note that both model classes can fit the data perfectly but M_2 can fit more data. Observe that the evidence is a PDF in the data space, i.e. in the data space the volume under $f(D \mid M_i)$ is always 1. Since M_2 can fit more data than M_1, that means the flat region in $f(D \mid M_2)$ is wider than that in $f(D \mid M_1)$. This implies that $f(\hat{D} \mid M_1)$ > $f(\hat{D} \mid M_2)$.



4. Posterior probability of model classes and model class averaging:
   One can derive the posterior probability for each model class from the Bayes rule:

$$P\left(M_i \mid \hat{D},\Omega\right) = \frac{f\left(\hat{D} \mid M_i\right) P\left(M_i \mid \Omega\right)}{\sum_{j=1}^{m} f\left(\hat{D} \mid M_j\right) P\left(M_j \mid \Omega\right)}$$

where $\Omega = \{M_1,...,M_m\}$ and also specifying the prior probability of M$_i$.

Robust estimation/prediction (model class averaging):

$$E\left[g(X) \mid \hat{D},\Omega\right] = \sum_{i=1}^{m} P\left(M_i \mid D,\Omega\right) \cdot E\left[g(X) \mid \hat{D},M_i\right]$$

- **How to evaluate** $f(D \mid M_i)$

Option 1: asymptotic approximation

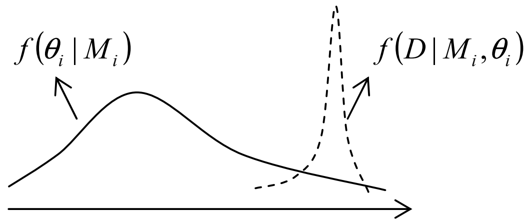
$$f\left(\hat{D} \mid M_i\right) \approx (2\pi)^{\frac{n_i}{2}} \frac{f\left(\hat{D} \mid M_i, x_i^*\right) f\left(x_i^* \mid M_i\right)}{\sqrt{\left|-\nabla_{x_i}^2 \log\left[f\left(\hat{D} \mid M_i, x_i^*\right) f\left(x_i^* \mid M_i\right)\right]\right|}}$$

Good for asymptotic cases. Requires solving optimization problems.

Option 2: sample directly from prior PDF

$$f\left(\hat{D} \mid M_i\right) = \int f\left(\hat{D}, x_i \mid M_i\right) dx_i = \int f\left(\hat{D} \mid M_i, x_i\right) f\left(x_i \mid M_i\right) dx_i$$

$$= E_{f(x_i|M_i)}\left[f\left(\hat{D} \mid M_i, X_i\right)\right] \approx \frac{1}{N}\sum_{k=1}^{N} f\left(\hat{D} \mid M_i, \hat{X}_{i,k}\right) \qquad \hat{X}_{i,k} \sim f\left(x_i \mid M_i\right)$$



Not a good choice since the main support regions of the prior and likelihood can be very different. Also, the likelihood is usually quite peaked. Sampling just from the prior can have high chance of missing the important region of the likelihood.

Option 3: importance sampling

Let $q(x)$ be the importance sampling PDF,

$$f\left(\hat{D} \mid M_i\right) = \int \frac{f\left(\hat{D} \mid M_i, x_i\right) f\left(x_i \mid M_i\right)}{q\left(x_i\right)} q\left(x_i\right) dx_i = E_q\left[\frac{f\left(\hat{D} \mid M_i, X_i\right) f\left(X_i \mid M_i\right)}{q\left(X_i\right)}\right]$$

$$\approx \frac{1}{N}\sum_{k=1}^{N} \frac{f\left(\hat{D} \mid M_i, \hat{X}_{i,k}\right) f\left(\hat{X}_{i,k} \mid M_i\right)}{q\left(\hat{X}_{i,k}\right)} \qquad \hat{X}_{i,k} \sim q(x_i)$$

Note: good for low dimensional *X*. May be inefficient for high dimensional *X*. The optimal choice is $q(x_i) = f(x_i \mid \hat{D}, M_i) \propto f(\hat{D} \mid M_i, x_i) f(x_i \mid M_i)$. But we don't know how to evaluate it since

$$f(x_i \mid \hat{D}, M_i) = \frac{f(\hat{D}, \mid M_i, x_i) f(x_i \mid M_i)}{f(\hat{D} \mid M_i)}$$

Option 4: entropy approach

Observe that

$$\log f(\hat{D} \mid M_i) = \log\left[f(\hat{D} \mid M_i, x_i) f(x_i \mid M_i)\right] - \log f(x_i \mid M_i, \hat{D})$$

$$\begin{aligned} \log f(\hat{D} \mid M_i) &= E_{f(x_i \mid M_i, \hat{D})}\left[\log\left[f(\hat{D} \mid M_i, x_i) f(x_i \mid M_i)\right] - \log f(x_i \mid M_i, \hat{D})\right] \\ &\approx \frac{1}{N}\sum_{k=1}^{N} \log\left[f(D \mid M_i, \hat{X}_{i,k}) f(\hat{X}_{i,k} \mid M_i)\right] + \underbrace{H\left[f(x_i \mid M_i, \hat{D})\right]}_{\substack{\textit{differential entropy:} \\ \textit{estimated from samples}}} \qquad \hat{X}_{i,k} \sim f(x_i \mid M_i, \hat{D}) \end{aligned}$$

where the posterior samples can be drawn from $f(x_i \mid \hat{D}, M_i)$ using MCMC. The differential entropy of $f(x_i \mid \hat{D}, M_i)$ can also be estimated from the posterior samples (see [1] for estimating entropy from samples).

From our experience, we found that the behavior of the estimator $\frac{1}{N}\sum_{k=1}^{N} \log\left[f(D \mid M_i, \hat{X}_{i,k}) f(\hat{X}_{i,k} \mid M_i)\right]$ is quite nice. The possible reasons may include the following: (1) the posterior samples will not miss the central region of $\log\left[f(\hat{D} \mid M_i, x_i) f(x_i \mid M_i)\right]$ and (2) although the posterior samples may miss the tail region of $\log\left[f(\hat{D} \mid M_i, x_i) f(x_i \mid M_i)\right]$, where the function tends to minus infinity, but the contribution of that region is theoretically zero anyway.

Option 5: transitional MCMC (later)

⌖ References:

[1] Beirlant, J., Dudewicz, E.J., Gyorfi, L., and van der Meulen, E.C. (2001). "Nonparametric entropy estimation: an overview."