Model Selection and Prior Selection

Khanh Nguyen

June 04, 2015

Consider a general machine learning problem, where we have data D, model m and model parameter θ_m.

Preliminary concepts:

- Likelihood: $P(D|\theta_m, m)$
- Prior: $P(\theta_m|m)$
- Posterior: $P(\theta_m|D, m)$[1]

1 Model selection

To select between different families of models, we compute the posterior over models:

$$P(m|D) = \frac{P(D|m)P(m)}{\sum_m P(m, D)}$$

Then we choose the MAP model $\hat{m} = arg \max_m p(m|D)$.

Assume $P(m) \propto 1$ (uniform), then it becomes picking

$$\hat{m} = arg \max P(D|m)$$

.

To compute $P(D|m)$, we marginalize over all model parameters.

$$P(D|m) = \int P(D|\theta_m)P(\theta_m|m)$$

where $P(D|\theta_m)$ is the likelihood and $P(\theta_m|m)$ is the prior.

$P(D|m)$ is called the *marginal likelihood*. Notice that when we computing the marginal likelihood by integrating over all parameters, we eliminate overfitting. A model with a lot of parameters should not be favored since it fits a large set of data and has to split probabilities among them.

[1] m is sometimes dropped from the condition.

2 Empirical Bayes

Let α_m be the prior of the parameters of the model. The full Bayesian generative model is as follows:

$$P(\theta_m, \alpha_m|D) \propto P(D|\theta_m)P(\theta_m|\alpha_m)P(\theta_m)$$

As an approximation, we use a point estimate of α_m and assume the distribution of α_m is uniform [2]:

$$\hat{\alpha}_m = arg\max P(D|\alpha_m) = arg\max \left[\int P(D|\theta_m)P(\theta_m|\alpha_m)d\theta_m\right]$$

This is called *Empirical Bayes* since we are using a empirical value rather than putting a distribution on the prior. Notice that, in this case, the prior is not chosen independently of the data anymore.

Here is the hierarchy of methods, going from the "most empirical" one to the "most Bayesian" one:

- Maximum likelihood: $\hat{\theta}_m = arg\max_{\theta_m} P(D|\theta_m)$
- MAP estimation: $\hat{\theta}_m = arg\max_{\theta_m} P(D|\theta_m)P(\theta_m|\alpha_m)$
- ML-II (Empirical Bayes): $\hat{\alpha}_m = arg\max_{\alpha_m} \int P(D|\theta_m)P(\theta_m|\alpha_m)d\theta_m$
- MAP-II: $\hat{\alpha}_m = arg\max_{\alpha_m} \int P(D|\theta_m)P(\theta_m|\alpha_m)P(\alpha_m)d\theta_m$
- Full Bayes: $P(\alpha_m, \theta_m|D) = P(D|\theta_m)P(\theta_m|\alpha_m)P(\alpha_m)$

[2] As we go higher in the Bayesian hierarchy, the prior distribution matters less.