Notes on Bayesian inference

Yang Xu

Machine Learning Department @ Carnegie Mellon University

Denoting X as the observed data, M the model and θ the unknown model parameters, the Bayes Theorem describes a mechanism to infer the unknown variables following basic probability theories:

$$\boldsymbol{P}(\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{M}) = \frac{P(X|\theta,M)P(\theta|M)}{P(X|M)} = \frac{P(X|\theta,M)P(\theta|M)}{\int P(X|\theta,M)P(\theta|M)d\theta}$$

Here the term in red is the Likelihood (a function of the parameters; given θ, the data are often assumed independently and identically distributed), the term in green is the Prior (pre-specified distribution supposedly reflecting some degree of belief about θ), the term in blue is the Marginal Likelihood or Model Evidence (a.k.a. the normalizing constant, although its significance is much beyond normalization), and finally the term in bold is the **Posterior** (distribution of unknown variables to be inferred). Specifying the RGB terms and cranking the mechanism via the Bayes Theorem informs one of the hidden quantities–commonly one might wish to find a single set of estimate for θ, e.g. the maximum or the mean of the posterior distribution. To make prediction about the new data X^*, the Bayes Theorem allows one to incorporate the acquired posterior knowledge by integration or marginalization (incorporating uncertainties in θ):

$$P(X^*|X,M) = \int P(X^*|\theta,M)\boldsymbol{P}(\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{M})d\theta$$

Note that the term in red is conceptually identical to the likelihood term we just introduced, whereas the term in bold is exactly the posterior we acquired given the "old" observation X. One advantage in the Bayes predictive distribution is that it implicitly encodes robustness against overfitting while integrating all learnt knowledge about the *a priori* unknown θ. In summary, the Bayes Theorem as described can be understood as an inverse operation, or inductive inference, where the goal is to infer hidden quantities (in this case θ, although the connotation of hidden variables is much more general than model parameters) with incomplete information and prior beliefs. It is worth noting though that the efficacy of Bayesian inference is influenced by the prior as well as the likelihood form.

Last but not least, Bayesian inference offers a coherent way to compare and select models. Suppose that there are N models $M_1, \dots, M_N$, the Bayes Theorem applies equally to the distribution of these models:

$$\boldsymbol{P}(\boldsymbol{M_i}|\boldsymbol{X}) = \frac{P(X|M_i)P(M_i)}{P(X)}$$

Again the terms in green and in bold are the prior and the posterior, only that here they apply to the models (zoom-out from individual models). Note that the term in blue is the model evidence introduced previously. To see why it is so called, assuming that each model is equally likely, i.e. a flat prior, then the best or most "evidential" model would be:

$$M_{best} \leftarrow max[\boldsymbol{P}(\boldsymbol{M_i}|\boldsymbol{X})] = \max\,[P(X|M_i)]$$

In other words, if you have two models M_1 and M_2, the way that you would figure out which one better describes the data is simply considering the ratio, called the Bayes Factor:

$$\frac{P(X|M_1)}{P(X|M_2)}$$