Pattern Recognition and Machine Learning Chapter 1 Summary *

CSCI-GA 3033 Bayesian Machine Learning
New York University
Yilun Kuang

Sep 16, 2022

1 Introduction

Convention pattern recognition techniques like minimizing RMS for regression suffers from overfitting due to lack of data or excess of parameters of the models. Techniques like regularization and cross-validation are proposed to reduce the overfitting issues. A more formal treatment of curve fitting comes from probability theory. Given the i.i.d assumption, gaussian noise assumption, and $D = \{(\vec{x}_i, y_i)\}_{i=1}^N$, we have

$$p(\vec{t}|\vec{x}, w, \beta) = \prod_{i=1}^{N} \mathcal{N}(t_i|y(x_i, w), \beta^{-1}).$$

Notice that if we apply log on $p(\vec{t}|\vec{x}, w, \beta)$, maximizing the log likelihood is equivalent to minimizing the sum of squared errors under gaussian assumption. Now if we consider $p(\vec{t}|\vec{x}, w, \beta)$ to be the likelihood, there is also a corresponding posterior $p(w|\vec{x}, \vec{t}, \alpha, \beta) \propto p(\vec{t}|\vec{x}, w, \beta)p(w|\alpha)$. So we can also maximizing the posterior. From a Bayesian perspective, we have

$$p(t|x, \vec{x}, \vec{t}) = \mathcal{N}(t|\beta\phi(x)^\top S \sum_{n=1}^{N} \phi(x_n)t_n, \beta^{-1} + \phi(x)^\top S\phi(x)).$$

which has additional $\phi(x)^\top S\phi(x))$ terms compared to maximum posterior. Also, for decision problems, there are generative models and discriminative models that corresponds to different decision behaviors.

2 Information Theory

To quantify the average amount of information of a given random variable x, we use the notion of entropy $H(x) = -\sum p(x) \ln p(x)$. To further differentiate between the distributions of two random variables, we introduce relative entropy or KL-Divergence:

$$\mathrm{KL}(p||q) = -\int p(x) \ln q(x) dx - \left(-\int p(x) \ln p(x) dx\right) = -\int p(x) \ln \left\{\frac{q(x)}{p(x)}\right\} dx.$$

Notice that if there is a data generating distribution $p(x)$, we can approximate this distribution by $q(x|\theta)$, which gives us $\mathrm{KL}(p||q) = \sum_{n=1}^{N}\{-\ln q(x_n|\theta) + \ln p(x_n)\}$. Then minimizing the KL divergnce is equivalent to maximizing the likelihood. Now, given a joint distribution $p(x, y)$, we can tell whether x and y are independent by calculating the mutual information:

$$I[x, y] = \mathrm{KL}(p(x, y)||p(x)p(y)) = -\int\int p(x, y) \ln \left(\frac{p(x)p(y)}{p(x, y)}\right) dxdy.$$

Notice that $I[x, y] = 0$ if and only if $p(x, y) = p(x)p(y)$ and $I[x, y] = H[x] - H[x|y]$. From a Bayesian perspective, $I[x, y]$ is the measure of uncertainty of x given observations of y.

*Full Version in Development