[Paper review 9]

Priors for Infinite Networks

(Radford M. Meal, 1994)

[Contents]

0. Abstract

Prior over weights

- "Priors over functions reach reasonable limits" as the number of hidden units in the network "goes to infinity"

1. Introduction

meaning of the weights in NN = "obscure" $\rightarrow$ hard to design a prior

focus on the limit, as the number of hidden units $H \rightarrow \infty$

(infinite network = "non-parametric" model)

this approach does not restrict the size of training dataset

(only on the size of the computer used for training)

over fitting does not occur in Bayesian Learning!

structure of NN

- I input values
- H sigmoidal hidden units
- O output values

$$f_k(x) = b_k + \sum_{j=1}^{H} v_{jk} h_j(x)$$

$$h_j(x) = \tanh\left(a_j + \sum_{i=1}^{I} u_{ij} x_i \right)$$

2. Priors Converging to Gaussian Process

(past works) prior : Gaussian distribution

as H increases, the prior over functions (implied by such priors) "converge to GP(Gaussian Process)"

2.1 Limits for Gaussian Priors

prior distribution of $f_k(x^{(1)})$

(= prior distribution for the weights and biases)

$f_k(x) = b_k + \sum_{j=1}^{H} v_{jk} h_j(x)$

(= sum of bias & weighted contributions of H hidden untis)

(1) Bias's contribution

- variance : σ_b^2

(2) Each hidden units' contributions'

- mean : $E\left[v_{jk} h_j\left(x^{(1)}\right)\right] = E\left[v_{jk}\right] E\left[h_j\left(x^{(1)}\right)\right] = 0$
- variance : $E\left[\left(v_{jk} h_j\left(x^{(1)}\right)^2\right] - 0^2 = E\left[v_{jk}^2\right] E\left[h_j\left(x^{(1)}\right)^2\right] = \sigma_v^2 E\left[h_j\left(x^{(1)}\right)^2\right] = \sigma_v^2 V(x^{(1)})$

  (where $V(x^{(1)}) = E\left[h_j\left(x^{(1)}\right)^2\right]$ for all j)

(1) + (2) Total contribution of $f_k(x^{(1)})$:

- variance = $\sigma_b^2 + H\sigma_v^2 V(x^{(1)})$
  $= \sigma_b^2 + \omega_v^2 V\left(x^{(1)}\right)$

  (if we set $\sigma_v = \omega_v H^{-1/2}$)

Joint distribution of $f_k\left(x^{(1)}\right), \ldots, f_k\left(x^{(n)}\right)$

as $H \to \infty$, prior joint distribution converges to "MVN" with

- mean : 0
- variance :

  $$E\left[f_k\left(x^{(p)}\right) f_k\left(x^{(q)}\right)\right] = \sigma_b^2 + \sum_j \sigma_v^2 E\left[h_j\left(x^{(p)}\right) h_j\left(x^{(q)}\right)\right]$$
  $$= \sigma_b^2 + \omega_v^2 C\left(x^{(p)}, x^{(q)}\right)$$

  (where $C\left(x^{(p)}, x^{(q)}\right) = E\left[h_j\left(x^{(p)}\right) h_j\left(x^{(q)}\right)\right]$)

(Gaussian Process : dist'n over function in which " the joint distribution of the values of the function at any finite number of point is MVN")