[Paper review 10]

Computing with Infinite Networks

(Christopher K. I. Williams, 1997)

[Contents]

0. Abstract

when $H \rightarrow \infty$: single layer NN with a prior = GP (Neal, 1994)

contribution: "Analytic forms"are derived for the "Covariance Function" of the GP corresponding to networks with "Sigmoidal & Gaussian " hidden units

1. Introduction

In practice, will not use ∞ hidden units $\rightarrow$ overfitting!

BUT, in Bayesian, no worry!

Neal(1994) : Infinite NN = GP, but does not give the covariance function

this paper shows...

- for "certain weight priors" and "transfer functions" in NN,

  the covariance function of GP can be calculated "ANALYTICALLY"

calculating analytically allows...

- 1) predictions to be made in $O(n^3)$ (n = number of training examples)
- 2) facilitates the comparison of the properties of NN with ∞ hidden units, as compared to other GP priors
- 3) dramatically reduces the dimensionality of MCMC integrals, thus improve speed of convergence

1.1 From "prior on WEIGHTS" to "prior on FUNCTIONS"

(original)

usually specified "hierarchically"

($P(w)=\int P(w \mid \theta) P(\theta) d \theta$ ... integrate out hyperprior)

(our case)

do not do as above

(since it introduces weight correlations, which prevents convergence to GP)

weight posterior : $P(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{\theta})$

predictive distribution for y_{*} : $P(y_{*} \mid \boldsymbol{t}, \boldsymbol{\theta})$

predictive distribution

(1) $P\left(y_{*} \mid \boldsymbol{t}, \boldsymbol{\theta}\right)=\int \delta\left(y_{*}-f_{*}(\boldsymbol{w})\right) P(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{\theta}) d \boldsymbol{w}$

((1) can be viewed as making prediction, using "priors over functions" rather than "prior over weights")

(by using Bayes Theorem, $P(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{\theta})=P(\boldsymbol{t} \mid \boldsymbol{w}) P(\boldsymbol{w} \mid \boldsymbol{\theta}) / P(\boldsymbol{t} \mid \boldsymbol{\theta})$, and $P(\boldsymbol{t} \mid \boldsymbol{w})=\int P(\boldsymbol{t} \mid \boldsymbol{y}) \delta(\boldsymbol{y}-\boldsymbol{f}(\boldsymbol{w})) d \boldsymbol{y}$)

(2) $P\left(y_{*} \mid \boldsymbol{t}, \boldsymbol{\theta}\right)=\frac{1}{P(\boldsymbol{t} \mid \boldsymbol{\theta})} \iint P(\boldsymbol{t} \mid \boldsymbol{y}) \delta\left(y_{*}-f_{*}(\boldsymbol{w})\right) \delta(\boldsymbol{y}-\boldsymbol{f}(\boldsymbol{w})) P(\boldsymbol{w} \mid \boldsymbol{\theta}) d \boldsymbol{w} d \boldsymbol{y}$

(since $P\left(y_{*}, \boldsymbol{y} \mid \theta\right)=P\left(y_{*} \mid \boldsymbol{y}, \boldsymbol{\theta}\right) P(\boldsymbol{y} \mid \theta)=\int \delta\left(y_{*}-f_{*}(\boldsymbol{w}) \delta(\boldsymbol{y}-\boldsymbol{f}(\boldsymbol{w})) P(\boldsymbol{w} \mid \theta) d \boldsymbol{w}\right.$)

(3) $P\left(y_{*} \mid \boldsymbol{t}, \boldsymbol{\theta}\right)=\frac{1}{P(\boldsymbol{t} \mid \boldsymbol{\theta})} \int P(\boldsymbol{t} \mid \boldsymbol{y}) P\left(y_{*} \mid \boldsymbol{y}, \boldsymbol{\theta}\right) P(\boldsymbol{y} \mid \boldsymbol{\theta}) d \boldsymbol{y}=\int P\left(y_{*} \mid \boldsymbol{y}, \boldsymbol{\theta}\right) P(\boldsymbol{y} \mid \boldsymbol{t}, \boldsymbol{\theta}) d \boldsymbol{y}$

$\rightarrow$ Result : view of "priors over functions" (= $P\left(y_{*} \mid \boldsymbol{y}, \boldsymbol{\theta}\right)$)

In general, we can use

- 1) weight space view
- 2) function space view

For infinite NN, more useful to use 2) function space view

2. Gaussian Process

widely used covariance functions

- stationary : $C\left(x, x^{\prime}\right)=C\left(x-x^{\prime}\right)$
- isotropic : $C\left(h^{*}\right)=C(h)$ where $h^{*}=x-x^{\prime}$ and $h=\mid h^{*} \mid$

2-1. Prediction with GP

data : generated from "prior" stochastic process + independent Gaussian "noise" added

- 1) prior covariance function : $C_P(x_i, x_j)$
- 2) noise process : $C_N(x_i, x_j) = \sigma_\nu^2 \delta_{ij}$

as both 1) and 2) are Gaussian, the integral can be done analytically!

$P(y_* \mid \boldsymbol{t}, \boldsymbol{\theta})$

- mean :$\hat{y}(\boldsymbol{x}_*) = \boldsymbol{k}_P^T(\boldsymbol{x}_*)(K_P + K_N)^{-1}\boldsymbol{t}$
- variance : $\sigma_{\hat{y}}^2(\boldsymbol{x}_*) = C_P(\boldsymbol{x}_*, \boldsymbol{x}_*) - \boldsymbol{k}_P^T(\boldsymbol{x}_*)(K_P + K_N)^{-1}\boldsymbol{k}_P(\boldsymbol{x}_*)$

where $[K_\alpha]_{ij} = C_\alpha(x_i, x_j)$ for $\alpha = P, N$, $k_P(x_*) = (C_P(x_*, x_1), \ldots, C_P(x_*, x_n))^T$

and $\sigma_{\hat{y}}^2(x_*)$ gives the "error bars" of the prediction.

3. Covariance Functions for Neural Network

input-to-hidden weights : u

$f(x) = b + \sum_{j=1}^H v_j h(\boldsymbol{x}; \boldsymbol{u}_j)$

- mean : $E\boldsymbol{w}[f(\boldsymbol{x})] = 0$
- variance :

$$\begin{aligned} E\boldsymbol{w}[f(\boldsymbol{x})f(\boldsymbol{x}')] &= \sigma_b^2 + \sum_j \sigma_v^2 E\boldsymbol{u}[h_j(\boldsymbol{x};\boldsymbol{u})h_j(\boldsymbol{x}';\boldsymbol{u})] \\ &= \sigma_b^2 + H\sigma_v^2 E\boldsymbol{u}[h(\boldsymbol{x};\boldsymbol{u})h(\boldsymbol{x}';\boldsymbol{u})] \\ &= \omega^2 E_{\boldsymbol{u}}[h(\boldsymbol{x};\boldsymbol{u})h(\boldsymbol{x}';\boldsymbol{u})] \end{aligned}$$

(letting ω^2/H as a scale of σ_v^2)

obtain covariance function by calculating $E_{\boldsymbol{u}}[h(\boldsymbol{x};\boldsymbol{u})h(\boldsymbol{x}';\boldsymbol{u})]$

Calculate $V(\boldsymbol{x}, \boldsymbol{x}') \stackrel{\text{def}}{=} E\boldsymbol{u}[h(\boldsymbol{x};\boldsymbol{u})h(\boldsymbol{x}';\boldsymbol{u})]$

by using 2 specific transfer functions (with Gaussian weight priors)

- 1) Sigmoidal function
- 2) Gaussian

3.1 Sigmoidal transfer function

- very common choice in NN
- $h(\boldsymbol{x};\boldsymbol{u})=\Phi\left(u_0+\sum_{i=1}^{d} u_j x_i\right)$ (where $\boldsymbol{u}\sim N(0,\Sigma)$)
- $\Phi(z)=2/\sqrt{\pi}\int_0^z e^{-t^2}\,dt$ (erf function, CDF of Gaussian)

$$V_{\text{erf}}\left(\boldsymbol{x},\boldsymbol{x}'\right)=\frac{1}{(2\pi)^{\frac{d+1}{2}}|\Sigma|^{1/2}}\int\Phi\left(\boldsymbol{u}^T\tilde{\boldsymbol{x}}\right)\Phi\left(\boldsymbol{u}^T\tilde{\boldsymbol{x}}'\right)\exp\left(-\tfrac{1}{2}\boldsymbol{u}^T\Sigma^{-1}\boldsymbol{u}\right)d\boldsymbol{u}$$

$$V_{\text{erf}}\left(x,x'\right)=\frac{2}{\pi}\sin^{-1}\frac{2\tilde{x}^T\Sigma\tilde{x}'}{\sqrt{\left(1+2\tilde{x}^T\Sigma\tilde{x}\right)\left(1+2\tilde{x}'^T\Sigma\tilde{x}'\right)}}$$ (this is not stationary!)

But, if

- set $\Sigma=\operatorname{diag}\left(\sigma_0^2,\sigma_I^2,\ldots,\sigma_I^2\right)$
- $|x|^2,|x'|^2\gg\left(1+2\sigma_0^2\right)/2\sigma_I^2$

Then, $V_{\text{erf}}\left(\boldsymbol{x},\boldsymbol{x}'\right)\simeq 1-2\theta/\pi,$ (where θ is the angle between $\boldsymbol{x}$ and $\boldsymbol{x}'$)

3.2 Gaussian transfer function

- very common choice in NN

  (Gaussian basis function are often used in RBF networks)
- $h(\boldsymbol{x};\boldsymbol{u})=\exp\left[-(\boldsymbol{x}-\boldsymbol{u})^T(\boldsymbol{x}-\boldsymbol{u})/2\sigma_g^2\right]$ (where $u\sim N\left(0,\sigma_u^2 I\right)$)

$$V_G\left(\boldsymbol{x},\boldsymbol{x}'\right)=\frac{1}{\left(2\pi\sigma_u^2\right)^{d/2}}\int\exp-\frac{(\boldsymbol{x}-\boldsymbol{u})^T(\boldsymbol{x}-\boldsymbol{u})}{2\sigma_g^2}\exp-\frac{(\boldsymbol{x}'-\boldsymbol{u})^T(\boldsymbol{x}'-\boldsymbol{u})}{2\sigma_g^2}\exp-\frac{\boldsymbol{u}^T\boldsymbol{u}}{2\sigma_u^2}c$$

(by completing the square & integrating out u)

$$V_G\left(x,x'\right)=\left(\frac{\sigma_e}{\sigma_u}\right)^d\exp\left\{-\frac{x^Tx}{2\sigma_m^2}\right\}\exp\left\{-\frac{(x-x')^T(x-x')}{2\sigma_s^2}\right\}\exp\left\{-\frac{x'^Tx'}{2\sigma_m^2}\right\}$$ (t his is not stationary!)

where $1/\sigma_e^2=2/\sigma_g^2+1/\sigma_u^2,\sigma_s^2=2\sigma_g^2+\sigma_g^4/\sigma_u^2$ and $\sigma_m^2=2\sigma_u^2+\sigma_g^2$

But, If $\sigma_u^2\rightarrow\infty$

$V_G\left(x,x'\right)\propto\exp\left\{-\left(x-x'\right)^T\left(x-x'\right)/4\sigma_g^2\right\}^4.$

For a finite value of σ_u^2,

$V_G\left(x,x'\right)$ is a stationary covariance function "modulated" by the Gaussian decay function $\exp\left(-\boldsymbol{x}^T\boldsymbol{x}/2\sigma_m^2\right)\exp\left(-\boldsymbol{x}'^T\boldsymbol{x}'/2\sigma_m^2\right).$

Clearly if σ_m^2 is much larger than the largest distance in x -space then the predictions made with V_G and a Gaussian process with only the stationary part of V_G will be very similar.

3.3 Comparing covariance functions

시공간자료분석 수강 후에...