Gaussian Processes for Classification

Lili MOU

moull12@sei.pku.edu.cn
http://sei.pku.edu.cn/~moull12

30 April 2015

Reference

§6.4.5, §6.4.6

C. Bishop, *Pattern Recognition and Machine Learning*

Intuition

GP for Regression: Let X, Z, t, ϵ be generic variables

$$t = Z_X + \epsilon$$

where $Z_X \sim \mathcal{GP}(0, k)$, $\epsilon \sim \mathcal{N}(0, \nu \mathbf{I})$

Let $\mathbf{t} = (t_1, t_2, \cdots, t_n)^T$ be a set of variables of interest, where t_i is a copy of t.

$$\mathbf{t} \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$$

where $\mathbf{K}$ is defined by the covariance function k, evaluated at pairs $(X^{(i)}, X^{(j)})$, $i, j = 1, 2, \cdots, n$.

GP for Classification: t is not Gaussian, but $\sigma^{-1}(t)$ may be.

$$a_X = Z_X + \epsilon$$

$$p(t|X) = \sigma(a_X)$$

Illustration

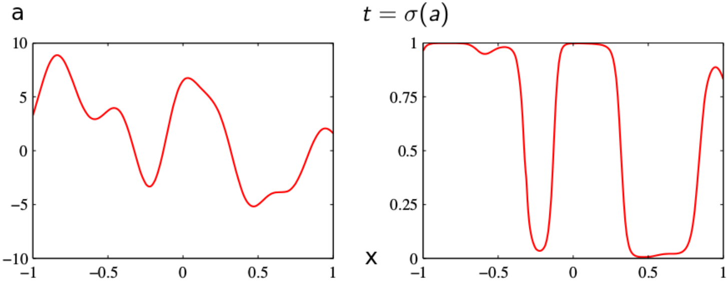

Figure 6.11 The left plot shows a sample from a Gaussian process prior over functions $a(\mathbf{x})$, and the right plot shows the result of transforming this sample using a logistic sigmoid function.

Solving the Predictive Density

"Everything unknown is a random variable."

—Bayesiansim

Model:

$$Z_X \sim \mathcal{GP}(0, k)$$

$$a_X = Z_X + \epsilon$$

$$p(t_X|X) = \sigma(a_X)$$

Goal: to predict $p(t_*|X_*, \mathbf{t}, \mathbf{X})$, where $\mathbf{X}$ and $\mathbf{t}$ refer to training data with labels.

$$p(t_*|\mathbf{t}) = \int p(t_*|a_*)p(a_*|\mathbf{t})\,\mathrm{d}a_*$$

with all data samples $\mathbf{X}$ and X_* omitted on the right-hand side of conditional bars.

Plan:

$$p(t_*|a_*) = \sigma(\cdot) \simeq \Phi(\cdot)$$

$$p(a_*|\mathbf{t}) \simeq \mathcal{N}(\cdot)$$

$$\Phi(\cdot) * \mathcal{N}(\cdot) = \Phi(\cdot) \simeq \sigma(\cdot)$$

Solving the Predictive Density (2)

$$p(a_*|\mathbf{t}_N) = \int p(a_*|\mathbf{a}_N)p(\mathbf{a}_N|\mathbf{t}_N)\,\mathrm{d}\mathbf{a}_N$$

where

- $p(a_*|\mathbf{a}_N) \sim \mathcal{N}\left(a_*|\mathbf{k}^T\mathbf{C}_N^{-1}\mathbf{a}_N, c - \mathbf{k}^T\mathbf{C}_N^{-1}\mathbf{k}\right)$

  Recall the assumption of GP for classification, and also the results of GP regression

- $p(\mathbf{a}_N|\mathbf{t}_N) = p(\mathbf{a}_N)p(\mathbf{t}_N|\mathbf{a}_N) \simeq \mathcal{N}(\cdot)$

  $p(\mathbf{a}_N) \sim \mathcal{N}$ by GP assumption

$$\begin{aligned} \ln p(\mathbf{t}_N|\mathbf{a}_N) &= \prod_{i=1}^{n} \sigma\left(a^{(i)}\right)^{t^{(i)}} \left(1 - \sigma\left(a^{(i)}\right)\right)^{1-t^{(i)}} \\ &= \prod_{i=1}^{n} e^{a^{(i)}t^{(i)}}\sigma\left(-a^{(i)}\right) \end{aligned}$$

Solving the Predictive Density (3)

Laplace approximation for $p(\mathbf{a}_N|\mathbf{t}_N)$

- ✓ Mode matches
- ✓ $\nabla\nabla \ln \tilde{p}(\cdot)$ matches

$$\begin{aligned}\Psi(\mathbf{a}_N) &\stackrel{\Delta}{=} \ln p(\mathbf{a}_N|\mathbf{t}_N)\\ &= \ln p(\mathbf{a}_N) + \ln p(\mathbf{t}_N|\mathbf{a}_N)\\ &= -\frac{1}{2}\mathbf{a}_N^T\mathbf{C}_N^{-1}\mathbf{a}_N + \mathbf{t}_N^T\mathbf{a}_N + \sum_{i=1}^{n}\ln(1+e^{a^{(i)}}) + \text{const}\end{aligned}$$

The second equation holds by noticing that

$$\ln p(\mathbf{t}_N|\mathbf{a}_N) = \prod_{i=1}^{n} e^{a^{(i)}t^{(i)}}\sigma\left(-a^{(i)}\right)$$

Solving the Predictive Density (4)

$$\Psi(\mathbf{a}_N) = -\frac{1}{2}\mathbf{a}_N^T\mathbf{C}_N^{-1}\mathbf{a}_N + \mathbf{t}_N^T\mathbf{a}_N + \sum_{i=1}^{n}\ln(1+e^{a^{(i)}}) + \text{const}$$

$$\nabla\Psi(\mathbf{a}_N) = \mathbf{t}_N - \boldsymbol{\sigma}_N - \mathbf{C}_N^{-1}\mathbf{a}_N$$

$$\nabla\nabla\Psi(\mathbf{a}_N) = -\mathbf{W}_N - \mathbf{C}_N^{-1}$$

where

$$\boldsymbol{\sigma}_N = [\sigma(a^{(1)}), \cdots, \sigma(a^{(n)})]^T$$

$\mathbf{W}_N$ is a diagonal matrix with elements $\sigma\left(a^{(i)}\right)\left(1-\sigma\left(a^{(i)}\right)\right)$

Necessary condition of a mode

$$\nabla\Psi(\mathbf{a}_N) = 0$$

$$\mathbf{a}_N^* = \mathbf{C}_N(\mathbf{t}_N - \boldsymbol{\sigma}_N)$$

Hence,

$$q(\mathbf{a}_N|\mathbf{t}_N) = \mathcal{N}\left(\mathbf{a}_N \middle| \mathbf{a}_N^*, \left(\mathbf{W}_N + \mathbf{C}_N^{-1}\right)^{-1}\right)$$

Solving the Predictive Density (5)

All is done. For detailed equations, please refer to *Pattern Recognition and Machine Learning*.