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, \textit{Pattern Recognition and Machine Learning}
Intuition

**GP for Regression:** Let $X, Z, t, \epsilon$ be generic variables

\[ t = Z_X + \epsilon \]

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

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

\[ t \sim \mathcal{N}(0, K) \]

where $K$ 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.

\[ a_X = Z_X + \epsilon \]

\[ p(t|X) = \sigma(a_X) \]
Figure 6.11  The left plot shows a sample from a Gaussian process prior over functions $a(x)$, 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 \( p(t_* | X_*, t, X) \), where \( X \) and \( t \) refer to training data with labels.

\[ p(t_* | t) = \int p(t_* | a_*) p(a_* | t) \, da_* \]

with all data samples \( X \) and \( X_* \) omitted on the right-hand side of conditional bars.

Plan:

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

\[ p(a_* | t) \simeq \mathcal{N}(\cdot) \]

\[ \Phi(\cdot) \ast \mathcal{N}(\cdot) = \Phi(\cdot) \simeq \sigma(\cdot) \]
Solving the Predictive Density (2)

\[ p(a_*|t_N) = \int p(a_*|a_N)p(a_N|t_N) \, da_N \]

where

\[ p(a_*|a_N) \sim \mathcal{N}(a_* | k^T C_N^{-1} a_N, c - k^T C_N^{-1} k) \]

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

\[ p(a_N|t_N) = p(a_N)p(t_N|a_N) \sim \mathcal{N}(\cdot) \]

\[ p(a_N) \sim \mathcal{N} \text{ by GP assumption} \]

\[ \ln p(t_N|a_N) = \prod_{i=1}^{n} \sigma(a^{(i)})^{t^{(i)}} (1 - \sigma(a^{(i)}))^{1-t^{(i)}} \]

\[ = \prod_{i=1}^{n} e^{a^{(i)} t^{(i)}} \sigma(-a^{(i)}) \]
Solving the Predictive Density (3)

Laplace approximation for $p(a_N|t_N)$

✓ Mode matches
✓ $\nabla^2 \ln \tilde{p}(\cdot)$ matches

\[
\Psi(a_N) \triangleq \ln p(a_N|t_N) = \ln p(a_N) + \ln p(t_N|a_N)
\]

\[
= -\frac{1}{2} a_N^T C_N^{-1} a_N + t_N^T a_N + \sum_{i=1}^{n} \ln(1 + e^{a(i)}) + \text{const}
\]

The second equation holds by noticing that

\[
\ln p(t_N|a_N) = \prod_{i=1}^{n} e^{a(i)t(i)} \sigma\left(-a(i)\right)
\]
Solving the Predictive Density (4)

\[ \Psi(a_N) = -\frac{1}{2} a_N^T C_N^{-1} a_N + t_N^T a_N + \sum_{i=1}^{n} \ln(1 + e^{a(i)}) + \text{const} \]

\[ \nabla \Psi(a_N) = t_N - \sigma_N - C_N^{-1} a_N \]

\[ \nabla \nabla \Psi(a_N) = -W_N - C_N^{-1} \]

where

\[ \sigma_N = [\sigma(a^{(1)}), \ldots, \sigma(a^{(n)})]^T \]

\[ W_N \] is a diagonal matrix with elements \( \sigma(a^{(i)}) (1 - \sigma(a^{(i)})) \)

Necessary condition of a mode

\[ \nabla \Psi(a_N) = 0 \]

\[ a_N^* = C_N (t_N - \sigma_N) \]

Hence,

\[ q(a_N|t_N) = \mathcal{N} \left( a_N \left| a_N^*, (W_N + C_N^{-1})^{-1} \right. \right) \]
Solving the Predictive Density (5)

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