Non-parametric Bayesian Methods in Machine Learning

Dr. Simon Rogers
School of Computing Science
University of Glasgow
simon.rogers@glasgow.ac.uk
@sdrogers

April 29, 2014

Outline

- **FIX ME AT THE END**
- (My) Bayesian philosophy
- Gaussian Processes for Regression and Classification
  - GP preliminaries
  - Classification (including semi-supervised)
  - Regression application 1: clinical (dis)-agreement
  - Regressopn application 2: typing on touch-screens
- Dirichlet Process flavoured Cluster Models
  - DP preliminaries
  - Idenfitying metabolites
  - (if time) Cluster models for multiple data views

About me

- I'm not a statistican by training (don't ask me to prove anything!).
- Education:
  - Undergraduate Degree: Electrical and Electronic Engineering (Bristol)
  - PhD: Machine Learning Techniques for Microarray Analysis (Bristol)
- Currently:
  - Lecturer: Computing Science
  - Research Interests: Machine Learning and Applied Statistics in Computational Biology and Human-Computer Interaction (HCI)

Lecture 4: GPs for classification and ordinal regression

Dr. Simon Rogers
School of Computing Science
University of Glasgow
simon.rogers@glasgow.ac.uk
@sdrogers

April 29, 2014

GPs for Classification and ordinal regression

- What if our observation model is non-Gaussian?
  - Classification:

$$P(y_n = 1|f_n) = \int_{-\infty}^{f_n} \mathcal{N}(z|0,1)\; dz = \phi(f_n)$$

  - Logistic Regression:

$$P(y_n = k|f_n) = \phi(b_{k+1}) - \phi(b_k)$$

  - etc
- Analytical inference is no longer possible
- I'll cover how to do inference in these models with the *auxiliary variable trick*

Binary classification

- Problem setup: we observe N data / target pairs $(\mathbf{x}_n, y_n)$ where $y_n \in \{0, 1\}$
- Place a GP prior on a set of latent variables f_n

$$\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$$

- Use the probit likelihood:

$$P(y_n = 1 | f_n) = \phi(f_n) = \int_{-\infty}^{f_n} \mathcal{N}(z|0, 1)\ dz$$

- Inference in this form is hard

Auxiliary Variable Trick

- Re-write the probit function:

$$
\begin{aligned}
P(y_n = 1|f_n) &= \int_{-\infty}^{f_n} N(z|0,1)\ dz \\
&= \int_{-\infty}^{0} N(z|-f_n,1)\ dz \\
&= \int_{0}^{\infty} N(z|f_n,1)\ dz \\
&= \int_{-\infty}^{\infty} \delta(z>0)\mathcal{N}(z|f_n,1)\ dz
\end{aligned}
$$

where $\delta(expr)$ is 1 if *expr* is true, and 0 otherwise.

Auxiliary Variable Trick

- If we define $P(y_n = 1|z_n) = \delta(z_n > 0)$ then we have:

$$P(y_n = 1|f_n) = \int_{-\infty}^{\infty} P(y_n = 1|z_n)p(z_n|f_n)\ dz_n$$

- and could therefore remove the integral to obtain a model including z_n:

$$p(y_n = 1, z_n|f_n) = P(y_n = 1|z_n)p(z_n|f_n)$$

- Doing inference in this model (i.e. with additional variables z_n) is much easier (but still not analytically tractable)
- Note: $P(y_n = 0|z_n) = \delta(z_n < 0)$

Example - Gibbs sampling for binary classification

- An easy way to perform inference in the augmented model is via Gibbs sampling
- Sample $z_n | f_n, y_n$:

$$\begin{aligned} p(z_n | f_n, y_n = 0) &\propto \delta(z_n < 0)\mathcal{N}(z_n | f_n, 1) \\ p(z_n | f_n, y_n = 1) &\propto \delta(z_n < 1)\mathcal{N}(z_n | f_n, 1) \end{aligned}$$

Example - Gibbs sampling for binary classification

- An easy way to perform inference in the augmented model is via Gibbs sampling
- Sample $z_n | f_n, y_n$:

$$
\begin{aligned}
p(z_n | f_n, y_n = 0) &\propto \delta(z_n < 0) \mathcal{N}(z_n | f_n, 1) \\
p(z_n | f_n, y_n = 1) &\propto \delta(z_n < 1) \mathcal{N}(z_n | f_n, 1)
\end{aligned}
$$

- Sample $\mathbf{f} | \mathbf{z}, \mathbf{C}$

$$p(\mathbf{f} | \mathbf{z}, \mathbf{C}) = \mathcal{N}(\boldsymbol{\mu}_f, \boldsymbol{\Sigma}_f)$$

where

$$\boldsymbol{\Sigma}_f = \left(\mathbf{I} + \mathbf{C}^{-1}\right)^{-1}, \quad \boldsymbol{\mu}_f = \boldsymbol{\Sigma}_f^{-1} \mathbf{z}$$

- Repeat ad infinitum