Lecture Notes: Bayesian Logistic Regression

Marc Toussaint

September 18, 2012

actual title:
Bayesian [Kernel|RBF|polynomial] [Ridge|Lasso] [Logistic] Regression

This is an attempt to cleanly document the basic family of methods generalizing from linear and logistic regression. Note:

- Bayesian Kernel Ridge Regression = Gaussian Process (Welling: Kernel Ridge Regression Lecture Notes; Rasmussen & Williams sections 2.1 & 6.2; Bishop sections 3.3.3 & 6)
- Bayesian Kernel Ridge Logistic Regression = Gaussian Process classification
- Kernel Ridge [Logistic] Regression $\sim$ SVM (when replacing the hinge loss by a squared loss)

1 Ridge Logistic Regression

2-class case only [lecture slides have multi-class case]

We have data $D = \{(x_i, y_i)\}_{i=1}^n$ with $x_i \in \mathbb{R}^d$ and $y_i \in \{0, 1\}$. For both classes we have discriminative functions $f(0, x)$ and $f(1, x)$; w.l.o.g. we assume $f(0, x) = 0$ and write $f(x) \equiv f(1, x)$. The discriminative function is linear in the parameters β,

$$f(x) = \phi(x)^\top \beta \tag{1}$$

$$\hat{y}(x) = \operatorname*{argmax}_y f(y, x) = \begin{cases} 0 & \text{else} \\ 1 & \text{if } \phi(x)^\top \beta > 0 \end{cases} \tag{2}$$

The conditional class probabilities are

$$p(1 \,|\, x) = \frac{e^{f(1,x)}}{e^{f(0,x)} + e^{f(1,x)}} = \sigma(f(x)) \tag{3}$$

with the *logistic sigmoid function* $\sigma(z) = \frac{e^z}{1+e^z} = \frac{1}{e^{-z}+1}$. The cost function function is the data neglog-likelihood plus regularization

$$L^{\text{logistic}}(\beta) = -\sum_{i=1}^n \log p(y_i \,|\, x_i) + \lambda \|\beta\|^2 \tag{4}$$

$$= \sum_{i=1}^n \Big[y_i \log p(1 \,|\, x_i) + (1 - y_i) \log[1 - p(1 \,|\, x_i)] \Big] - \lambda \|\beta\|^2 \tag{5}$$

The optimal parameters can be found by a Newton method, with Gradient and Hessian:

$$\nabla = \frac{\partial L^{\text{logistic}}(\beta)}{\partial \beta} \tag{6}$$

$$= \sum_{i=1}^n (y_i - p_i)\phi(x_i) - 2\lambda I \beta \;, \quad p_i := p(y{=}1 \,|\, x_i) \tag{7}$$

$$= \boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{p}) - 2\lambda I \beta \;, \quad \boldsymbol{X} = \begin{pmatrix} \phi(x_1)^\top \\ \vdots \\ \phi(x_n)^\top \end{pmatrix} \in \mathbb{R}^{n \times k} \tag{8}$$

$$H = \frac{\partial^2 L^{\text{logistic}}(\beta)}{\partial \beta^2} \tag{9}$$
$$= -\boldsymbol{X}^\top W \boldsymbol{X} - 2\lambda I \tag{10}$$
$$\beta \leftarrow \beta - H^{-1} \nabla \tag{11}$$
$$= \beta + (\boldsymbol{X}^\top W \boldsymbol{X} + 2\lambda I)^{-1} (\boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{p}) - 2\lambda I \beta) \tag{12}$$

Here, W diagonal with $W_{ii} = p_i(1-p_i)$. The latter is an interated weighted least square estimate (compare to $\hat{\beta}^{\text{ridge}} = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda I)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$)

2 Bayesian Ridge Logistic Regression

The regularizsation translates to a prior and we have

$$P(X) = \text{arbitrary (we're talking about a } \textit{conditional} \text{ model } P(Y|X)) \tag{13}$$
$$P(\beta) = \mathcal{N}(\beta|0, \frac{2}{\lambda}) \propto \exp\{-\lambda \|\beta\|^2\} \tag{14}$$
$$P(Y=1 \,|\, X, \beta) = \sigma(\beta^\top \phi(x)) \tag{15}$$

The parameter posterior is

$$P(\beta|D) \propto P(D \,|\, \beta)\; P(\beta) \propto \exp\{-L^{\text{logistic}}(\beta)\} \tag{16}$$

using a local Gaussian approximation (2nd order Taylor for L) around $\beta^* = \text{argmin}_\beta L(\beta)$:

$$L^{\text{logistic}}(\beta) \approx L(\beta^*) + \bar{\beta}^\top \nabla + \frac{1}{2} \bar{\beta}^\top H \bar{\beta}\,, \quad \bar{\beta} = \beta - \beta^* \tag{17}$$
$$P(\beta|D) \propto \exp\{-\bar{\beta}^\top \nabla - \frac{1}{2} \bar{\beta}^\top H \bar{\beta}\} \tag{18}$$
$$= \mathcal{N}[\bar{\beta}| - \nabla, H] = \mathcal{N}(\bar{\beta}| - H^{-1}\nabla, H^{-1}) \tag{19}$$
$$= \mathcal{N}(\beta|\beta^*, H^{-1}) \tag{20}$$

Given the Gaussian approximation of the parameter posterior, the predictive distribution of the *discriminative function* is also Gaussian, as for Bayesian Regression

$$P(f(x) \,|\, D) = \textstyle\int_\beta P(f(x) \,|\, \beta)\; P(\beta \,|\, D)\; d\beta \tag{21}$$
$$= \textstyle\int_\beta \mathcal{N}(f(x) \,|\, \phi(x)^\top \beta, \sigma^2)\; \mathcal{N}(\beta \,|\, \beta^*, H^{-1})\; d\beta \tag{22}$$
$$= \mathcal{N}(f(x) \,|\, \phi(x)^\top \beta^*, \sigma^2 + \phi(x)^\top H^{-1} \phi(x)) \tag{23}$$
$$=: \mathcal{N}(f(x) \,|\, f^*, s^2) \tag{24}$$

The predictive distribution over the label $y \in \{0, 1\}$ is the convolution of this Gaussian with the logistic function. This is more easily approximated by the convolution of the Gaussian with the probit function $\varphi(x) = \int_{-\infty}^{x} \mathcal{N}(0,1)\; dx$ (Gaussian cumulative function), which becomes again a probit function:

$$P(y(x) \,|\, D) = \textstyle\int_{f(x)} \sigma(f(x))\; P(f(x)|D)\; df \tag{25}$$
$$\approx \sigma(\sqrt{1 + s^2 \pi/8} f^*) \tag{26}$$