Bayesian Logistic Regression

Sargur N. Srihari
University at Buffalo, State University of New York
USA

Topics in Linear Models for Classification

- Overview
- 1. Discriminant Functions
- 2. Probabilistic Generative Models
- 3. Probabilistic Discriminative Models
- 4. The Laplace Approximation
- 5. Bayesian Logistic Regression

Topics in Bayesian Logistic Regression

- Recap of Logistic Regression
- Roadmap of Bayesian Logistic Regression
- Laplace Approximation
- Evaluation of posterior distribution
 - Gaussian approximation
- Predictive Distribution
 - Convolution of Sigmoid and Gaussian
 - Approximate sigmoid with probit
- Variational Bayesian Logistic Regression

Recap of Logistic Regression

- Feature vector ϕ , two-classes C_1 and C_2
- A posteriori probability $p(C_1|\phi)$ can be written as $p(C_1|\phi) = y(\phi) = \sigma(\mathbf{w}^{\mathrm{T}}\phi)$ where ϕ is a M-dimensional feature vector σ (.) is the logistic sigmoid function
- Goal is to determine the M parameters
- Known as logistic regression in statistics
 - Although a model for classification rather than for regression

Determining Logistic Regression parameters

Maximum Likelihood Approach to determine w

Data set $\{\phi_n, t_n\}$ where t_n ϵ $\{0,1\}$ and $\phi_n = \phi(\boldsymbol{x}_n), \ n=1,...,N$ Since t_n is binary we can use Bernoulli Let y_n be the probability that $t_n = 1$, i.e., $y_n = p(C_1|\phi_n)$ Denote $\mathbf{t} = (t_1, ..., t_N)^T$

Likelihood function associated with N observations

$$p(\mathbf{t} \mid \boldsymbol{w}) = \prod_{n=1}^{N} y_n^{t_n} \left\{ 1 - y_n \right\}^{1 - t_n}$$

Simple sequential solution

Error function is the negative of the log-likelihood

$$E(\boldsymbol{w}) = -\ln p(t \mid \boldsymbol{w}) = -\sum_{n=1}^{N} \left\{ t_n \ln y_n + (1 - t_n) \ln(1 - y_n) \right\}$$
Cross-entropy error function

- ullet No closed-form maximum likelihood solution for determining $oldsymbol{w}$
- Given Gradient of error function

$$\nabla E_n = (y_n - t_n) \phi_n$$

· Solve using an iterative approach

• where
$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N} (y_n - t_n) \phi_n$$

$$oldsymbol{w}^{ au+1} = oldsymbol{w}^{ au} - oldsymbol{\eta}
abla E_n$$
 Error x Feature Vector $oldsymbol{w}^{ au+1}$

Solution has severe over-fitting problems for linearly separable data So use IRLS algorithm 6

IRLS for Logistic Regression

• Posterior probability of class C_i is

$$p(C_1|\phi) = y(\phi) = \sigma(\mathbf{w}^T\phi)$$

Likelihood Function

for data set $\{\phi_n, t_n\}$, $t_n \in \{0,1\}$, $\phi_n = \phi(\boldsymbol{x}_n)$

$$p(\mathbf{t} \mid \boldsymbol{w}) = \prod_{n=1}^{N} y_n^{t_n} \left\{ 1 - y_n \right\}^{1 - t_n}$$

1. Error Function

Log-likelihood yields Cross-entropy

$$E(\mathbf{w}) = -\sum_{n=1}^{N} \left\{ t_n \ln y_n + (1 - t_n) \ln(1 - y_n) \right\}$$

IRLS for Logistic Regression

Gradient of Error Function:

$$\nabla E(\boldsymbol{w}) = \sum_{n=1}^{N} (y_n - t_n) \phi_n = \boldsymbol{\Phi}^T (\boldsymbol{y} - \boldsymbol{t})$$

3. Hessian:

$$\mathbf{H} = \nabla \nabla E(\mathbf{w}) = \sum_{n=1}^{N} y_n (1 - y_n) \phi_n \phi_n^T = \mathbf{\Phi}^T \mathbf{R} \mathbf{\Phi}$$

 Φ is $N \times M$ design matrix whose n^{th} row is ϕ_n^T

R is $N \times N$ diagonal matrix with elements $R_{nn} = y_n (1 - y_n)$

Hessian is not constant and depends on \boldsymbol{w} through R Since H is positive-definite (i.e., for arbitrary $\boldsymbol{u}, \ \boldsymbol{u}^T H \boldsymbol{u} > 0$) error function is a concave function of \boldsymbol{w} and so has a unique minimum

IRLS for Logistic Regression

4. Newton-Raphson update:

$$\boldsymbol{w}^{(\text{new})} = \boldsymbol{w}^{(\text{old})} - H^{-1} \nabla E(\boldsymbol{w})$$

Substituting $H = \Phi^T R \Phi$ and $\nabla E(w) = \Phi^T (y - t)$

$$w^{(\text{new})} = w^{(\text{old})} - (\Phi^{T}R\Phi)^{-1}\Phi^{T}(y-t)$$

$$= (\Phi^{T}R\Phi)^{-1}\{\Phi\Phi w^{(\text{old})} - \Phi^{T}(y-t)\}$$

$$= (\Phi^{T}R\Phi)^{-1}\Phi^{T}Rz$$

where z is a N-dimensional vector with elements

$$z = \Phi w$$
 (old)-R⁻¹(y - t)

Update formula is a set of normal equations

Since Hessian depends on w

Apply them iteratively each time using the new weight vector

Roadmap of Bayesian Logistic Regression

- Logistic regression is a discriminative probabilistic linear classifier: $p(C_1|\phi) = \sigma(w^T\phi)$
- Exact Bayesian inference for Logistic Regression $p(C_1|\phi) = \int \sigma(w^T\phi)p(w)dw$ is intractable, because:
- 1.Evaluation of posterior distribution $p(\boldsymbol{w}|t)$
 - Needs normalization of prior $p(\mathbf{w}) = N(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0)$ times likelihood (a product of sigmoids) $p(\mathbf{t}|\mathbf{w}) = \prod_{n=1}^{N} y_n^{t_n} \{1 y_n\}^{1-t_n}$
 - Solution: use Laplace approximation to get Gaussian q(w)
- 2. Evaluation of predictive distribution

$$p(C_1 | \boldsymbol{\phi}) \simeq \int \boldsymbol{\sigma}(\boldsymbol{w}^T \boldsymbol{\phi}) q(\boldsymbol{w}) d\boldsymbol{w}$$

- Convolution of sigmoid and Gaussian
 - Solution: Approximate Sigmoid by Probit

Laplace Approximation (summary)

- Need mode w_0 of posterior distribution p(w|t)
 - Done by a numerical optimization algorithm
- Fit a Gaussian centered at the mode

$$q(\boldsymbol{w}) = \frac{1}{W} f(\boldsymbol{w}) = \frac{|\mathbf{A}|^{1/2}}{(2\pi)^{M/2}} \exp\left\{-\frac{1}{2} (\boldsymbol{w} - \boldsymbol{w}_0)^{\mathrm{T}} \mathbf{A} (\boldsymbol{w} - \boldsymbol{w}_0)\right\}$$
$$= N(\boldsymbol{w} | \boldsymbol{w}_0, \mathbf{A}^{-1})$$

- Needs second derivatives of log posterior $A = -\nabla \nabla \ln f(\mathbf{w}) I_{\mathbf{w} = \mathbf{w}_0}$
 - Equivalent to finding Hessian matrix

$$S_N = -\nabla \nabla \ln p(\mathbf{w} \mid t) = S_0^{-1} + \sum_{i=1}^n y_i (1 - y_i) \phi_i \phi_i^T$$

Srihari Machine Learning

Evaluation of Posterior Distribution

Gaussian prior

$$p(\boldsymbol{w}) = N(\boldsymbol{w}|\boldsymbol{m}_0, S_0)$$

- Where m_0 and S_0 are hyper-parameters
- Posterior distribution

$$p(\boldsymbol{w}|\mathbf{t})$$
 α $p(\boldsymbol{w})p(\mathbf{t}|\boldsymbol{w})$ where $\mathbf{t} = (t_1,...,t_N)^{\mathrm{T}}$

- Substituting
$$p(\mathbf{t} \mid \boldsymbol{w}) = \prod_{n=1}^{N} y_n^{t_n} \left\{ 1 - y_n \right\}^{1 - t_n}$$

$$\ln p(\mathbf{w}|\mathbf{t}) = -\frac{1}{2}(\mathbf{w} - m_0)^T S_0^{-1}(\mathbf{w} - m_0) + \sum_{i=1}^n (t_n \ln y_n + (1 - t_n) \ln(1 - y_n) + const$$

• where
$$y_n = \sigma(\mathbf{w}^T \phi_n)$$

Gaussian Approximation of Posterior

- Maximize posterior p(w|t) to give
 - MAP solution $oldsymbol{w}_{ ext{map}}$
 - Done by numerical optimization
 - Defines mean of the Gaussian
- Covariance given by
 - Inverse of matrix of 2^{nd} derivatives of negative log-likelihood $S_N = -\nabla \nabla \ln p(\mathbf{w}|\mathbf{t}) = S_0^{-1} + \sum_{i=1}^n y_i(1-y_i)\phi_i\phi_i^T$
- Gaussian approximation to posterior

$$q(\boldsymbol{w}) = N(\boldsymbol{w} \mid \boldsymbol{w}_{\text{map}}, S_N)$$

 Need to marginalize wrt this distribution to make predictions

Predictive Distribution

- Predictive distribution for class C_1 , given new feature vector $\phi(x)$
 - Obtained by marginalizing wrt posterior $p(\boldsymbol{w}|t)$

$$p(C_1 | \phi, \mathbf{t}) = \int p(C_1, \mathbf{w} | \phi, \mathbf{t}) d\mathbf{w} \quad \text{Sum rule}$$

$$= \int p(C_1 | \phi, \mathbf{t}, \mathbf{w}) p(\mathbf{w} | \mathbf{t}) d\mathbf{w} \quad \text{Product rule}$$

$$= \int p(C_1 | \phi, \mathbf{w}) p(\mathbf{w} | \mathbf{t}) d\mathbf{w} \quad \text{Given } \phi \text{ and } \mathbf{w}, C_1 \text{ is indep of } \mathbf{t}$$

$$= \int \sigma(\mathbf{w}^T \phi) q(\mathbf{w}) d\mathbf{w} \quad \text{Approximate } p(\mathbf{w} | \mathbf{t}) \text{ by Gaussian } q(\mathbf{w})$$
corresponding probability for class C_2

$$p(C_2 | \phi, t) = 1 - p(C_1 | \phi, t)$$

Predictive distrib. is a Convolution

$$p(C_1 | \phi, \mathbf{t}) \simeq \int \sigma(\mathbf{w}^T \phi) q(\mathbf{w}) d\mathbf{w}$$

- Function $\sigma(\mathbf{w}^{\mathrm{T}}\phi)$ depends on \mathbf{w} only through its projection onto ϕ
- Denoting $a = \mathbf{w}^T \phi$ we have $\sigma(\mathbf{w}^T \phi) \simeq \int \delta(a \mathbf{w}^T \phi) \sigma(a) da$
 - where δ is the Dirac delta function
- Thus $\int \sigma(w^T \phi) q(w) dw = \int \sigma(a) p(a) da$ where $p(a) = \int \delta(a w^T \phi) q(w) dw$
 - Can evaluate p(a) because
 - the delta function imposes a linear constraint on $oldsymbol{w}$
 - Since q(w) is Gaussian, its marginal is also Gaussian
 - Evaluate its mean and covariance

$$\mu_{a} = \operatorname{E}[a] = \int p(a)da = \int q(\mathbf{w})\mathbf{w}^{T}\phi d\mathbf{w} = \mathbf{w}_{map}^{T}\phi$$

$$\sigma_{a}^{2} = \operatorname{var}[a] = \int p(a)\left\{a^{2} - \operatorname{E}[a]^{2}\right\}da$$

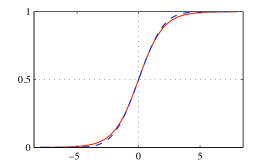
$$= \int q(\mathbf{w})\left\{(\mathbf{w}^{T}\phi)^{2} - (m_{N}^{T}\phi)^{2}\right\}d\mathbf{w} = \phi^{T}S_{N}\phi$$
15

Variational Approximation to Predictive Distribution

Predictive distribution is

$$p(C_1 \mid \mathbf{t}) = \int \sigma(a) p(a) da$$
$$= \int \sigma(a) N(a \mid \mu_a, \sigma_a^2) da$$

- Convolution of Sigmoid-Gaussian is intractable
- Use probit instead of logistic sigmoid



Approximation using Probit

$$p(C_1 \mid \mathbf{t}) = \int \sigma(a) N(a \mid \mu_a, \sigma_a^2) da$$

- Use probit which is similar to Logistic sigmoid
 - Defined as $\Phi(a) = \int_{0}^{a} N(\theta \mid 0,1) d\theta$



• Find λ such that two functions have same slope at origin Approximate $\sigma(a)$ by $\Phi(\lambda a)$

Find suitable value of λ by requiring that two have same slope at origin, which yields $\lambda^2 = \pi/8$

Convolution of probit with Gaussian is a probit

$$\int \Phi(\lambda a) N(a \mid \mu, \sigma^2) da = \Phi\left(\frac{\mu}{\left(\lambda^{-2} + \sigma^2\right)^{1/2}}\right)$$

$$p(C_1 | \phi, \mathbf{t}) = \int \sigma(a) N(a | \mu_a, \sigma_a^2) da$$
$$\simeq \sigma(\kappa(\sigma_a^2) \mu_a)$$

where

$$\kappa(\sigma^2) = (1 + \pi \sigma^2 / 8)^{-1/2}$$

Probit Classification

Applying it to

$$p(C_1 \mid \mathbf{t}) = \int \sigma(a) N(a \mid \mu_a, \sigma_a^2) da$$

We have

$$p(C_1 \mid \phi, \mathbf{t}) = \sigma(\kappa(\sigma_a^2)\mu_a)$$

where

$$\mu_a = \mathbf{w}_{map}^T \phi \qquad \qquad \boldsymbol{\sigma}_a^2 = \boldsymbol{\phi}^T S_N \boldsymbol{\phi}$$

Decision boundary corresponding to $p(C_1|\phi,t) = 0.5$ is given by

$$\mu_a = 0$$

This is the same solution as

$$\mathbf{w}_{map}^T \phi = 0$$

Thus marginalization has no effect!

When minimizing misclassification rate with equal prior probabilities

For more complex decision criteria it plays important role₁₈

Summary

- Logistic regression is a linear probabilistic discriminative model $p(C_1|x) = \sigma(w^T\phi)$
- Bayesian Logistic Regression is intractable
- Using Laplacian the posterior parameter distribution $p(\boldsymbol{w}|\mathbf{t})$ can be approximated as a Gaussian
- Predictive distribution is convolution of sigmoids and Gaussian $p(C_1|\phi) \simeq \int \sigma(w^T\phi)q(w)dw$
 - Probit yields convolution as probit