

Linear Models for Classification – Part II

Liang Sun
Arizona State University

July 24, 2007

- 1 Probabilistic Discriminative Models
- 2 The Laplace Approximation
- 3 Bayesian Logistic Regression

Generative Model Vs. Discriminative Model

- Probabilistic Generative Model

Model the class-conditional densities $p(\mathbf{x}|\mathcal{C}_k)$ and prior probabilities $p(\mathcal{C}_k)$, and compute $p(\mathcal{C}_k|\mathbf{x})$ using Bayes theorem

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})} \quad (1)$$

- Probabilistic Discriminative Model

Use the functional form of the generalized linear model explicitly and to determine its parameters directly by using maximum likelihood method

- Advantages of Discriminative Models

- Fewer adaptive parameters need to be determined
- Performance will be improved, especially when the class-conditional density assumption give a poor approximation to the true distributions

Fixed Basis Function

- The original data \mathbf{x} are mapped into $\phi(\mathbf{x})$ using a vector of basis functions $\phi(\mathbf{x})$
- The resulting model is linear in the feature space ϕ , but may not be linear in the original \mathbf{x} space
- The next discussions are based on the ϕ space

Logistic Regression – I

- In two-class problem, the posterior probability of class \mathcal{C}_1 can be written as a logistic sigmoid acting on a linear function of the feature vector ϕ so that

$$p(\mathcal{C}_1|\phi) = y(\phi) = \sigma(\mathbf{w}^T \phi) \quad (2)$$

- And $p(\mathcal{C}_2|\phi) = 1 - p(\mathcal{C}_1|\phi)$, σ is the logistic sigmoid function
- In statistics, the model defined by (2) is known as **logistic regression**
- The model defined by (2) is for **classification**, not for **regression**
- In logistic regression, we estimate the parameter \mathbf{w} directly

Logistic Regression – II

- Comparison of logistic regression and generative model in M -dimensional space
 - In logistic regression, only M parameters (components of \mathbf{w})
 - In generative model, suppose Gaussian class-conditional densities and maximum likelihood method are used, the number of parameters is $M(M + 5)/2 + 1$
 - Means: $2M$ parameters
 - Shared covariance: $(M + 1)M/2$ parameters
 - Prior $p(C_1)$: 1 parameter
- **Maximum Likelihood** method is used to determine the parameters of logistic regression model

Logistic Regression – III – Review of Logistic Sigmoid function

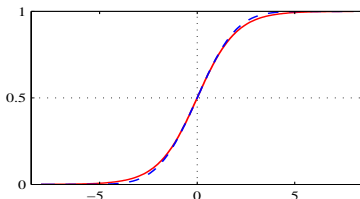
- Definition of **Logistic Sigmoid** function $\sigma(a)$

$$\sigma(a) = \frac{1}{1 + \exp(-a)} \quad (3)$$

- Properties of Logistic Sigmoid function $\sigma(a)$

$$\sigma(-a) = 1 - \sigma(a) \quad (4)$$

$$\frac{d\sigma}{da} = \sigma(1 - \sigma) \quad (5)$$



Logistic Regression – IV – Use Maximum Likelihood to estimate \mathbf{w}

- For a training data set $\{\phi_n, t_n\}$ where $t_n \in \{0, 1\}$ and $n = 1, 2, \dots, N$, the likelihood function is

$$p(\mathbf{t}|\mathbf{w}) = \prod_{n=1}^N y_n^{t_n} (1 - y_n)^{1-t_n} \quad (6)$$

- Definition of t_n , \mathbf{t} and y_n

$$t_n = \begin{cases} 1 & \text{if } n \in \mathcal{C}_1 \\ 0 & \text{if } n \in \mathcal{C}_2 \end{cases} \quad (7)$$

$$\mathbf{t} = (t_1, t_2, \dots, t_N)^T \quad (8)$$

$$y_n = p(\mathcal{C}_1|\phi_n) = \sigma(\mathbf{w}^T \phi_n) \quad (9)$$

- The error function is the negative logarithm of the likelihood, namely, **Cross-entropy** error function

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}) = \sum_{n=1}^N \{t_n \ln y_n + (1 - t_n) \ln(1 - t_n)\} \quad (10)$$

- The gradient of cross-entropy function with respect to \mathbf{w} is

$$\nabla E(\mathbf{w}) = \sum_{n=1}^N (y_n - t_n) \phi_n \quad (11)$$

Implementations of Logistic Regression – I

- Two different algorithms
 - Batch version
 - Online version
- The cross-entropy error function is concave, and a unique minimum is ensured
- Newton-Raphson Algorithm
It uses a local quadratic approximation to the cross-entropy error function to update w iteratively

$$\mathbf{w}^{(\text{new})} = \mathbf{w}^{(\text{old})} - H^{-1} \nabla E(\mathbf{w}) \quad (12)$$

Implementations of Logistic Regression – II

- For function $y = \mathbf{w}^T \phi$, the gradient and Hessian matrix are given

$$\nabla E(\mathbf{w}) = \sum_{n=1}^N (\mathbf{w}^T \phi_n - t_n) \phi_n = \mathbf{\Phi}^T \mathbf{\Phi} \mathbf{w} - \mathbf{\Phi}^T \mathbf{t} \quad (13)$$

$$H = \nabla \nabla E(\mathbf{w}) = \sum_{n=1}^N \phi_n \phi_n^T = \mathbf{\Phi}^T \mathbf{\Phi} \quad (14)$$

where n-th row of $\mathbf{\Phi}$ is ϕ_n^T

- The Newton-Raphson update for this function is

$$\mathbf{w}^{(\text{new})} = \mathbf{w}^{(\text{old})} - (\mathbf{\Phi}^T \mathbf{\Phi})^{-1} \left\{ \mathbf{\Phi}^T \mathbf{\Phi} \mathbf{w}^{(\text{old})} - \mathbf{\Phi}^T \mathbf{t} \right\} \quad (15)$$

$$= (\mathbf{\Phi}^T \mathbf{\Phi})^{-1} \mathbf{\Phi}^T \mathbf{t} \quad (16)$$

- It shows that if the error function is quadratic, the Newton-Raphson method gives the exact solution in one step

Implementations of Logistic Regression – III

- For cross-entropy error function, we can compute gradient and Hessian matrix

$$\nabla E(\mathbf{w}) = \sum_{n=1}^N (y_n - t_n) \phi_n = \Phi^T (\mathbf{y} - \mathbf{t}) \quad (17)$$

$$H = \nabla \nabla E(\mathbf{w}) = \sum_{n=1}^N y_n (1 - y_n) \phi_n \phi_n^T = \Phi^T \mathbf{R} \Phi \quad (18)$$

where \mathbf{R} is an $N \times N$ diagonal matrix, and

$$R_{nn} = y_n (1 - y_n) \quad (19)$$

- Note that $0 < y_n < 1$, then \mathbf{R} is positive definite, and H is also positive definite
- Since H is positive definite, the cross-entropy error function is concave, and it has a unique minimum

Implementations of Logistic Regression – IV

- The Newton-Raphson update for cross-entropy error function is

$$\begin{aligned}\mathbf{w}^{(\text{new})} &= \mathbf{w}^{(\text{old})} - (\Phi^T \mathbf{R} \Phi)^{-1} \Phi^T (\mathbf{y} - \mathbf{t}) \\ &= (\Phi^T \mathbf{R} \Phi)^{-1} \Phi^T \mathbf{R} \mathbf{z} \\ \text{where } \mathbf{z} &= \Phi \mathbf{w}^{(\text{old})} - \mathbf{R}^{-1} (\mathbf{y} - \mathbf{t})\end{aligned}\tag{20}$$

- (20) takes the form of a set of normal equations for a weighted least-squares problem, and diagonal matrix \mathbf{R} is the weight in each iteration
- Newton-Raphson algorithm is also known as **iterative reweighted least squares**

Multiclass Logistic Regression – I

- For the case of $K > 2$ classes, the posterior probability can be represented as

$$\begin{aligned} p(C_k|\phi) &= \frac{p(\phi|C_k)p(C_k)}{\sum_j p(\phi|C_j)p(C_j)} \\ &= \frac{\exp(a_k)}{\sum_j \exp(a_j)} \end{aligned} \quad (21)$$

$$= y_k(\phi) \quad (22)$$

where

$$a_k = \mathbf{w}_k^T \phi \quad (23)$$

- In discriminative model, we consider the use of maximum likelihood to determine the parameters $\{\mathbf{w}_k\}$ directly

Multiclass Logistic Regression – II

- According to (22), we have

$$\frac{\partial y_k}{\partial a_j} = y_k(I_{kj} - y_j) \quad (24)$$

where

$$I_{kj} = \begin{cases} 1 & \text{if } k = j \\ 0 & \text{otherwise} \end{cases} \quad (25)$$

- The likelihood function for a training set of size N is given as follows:

$$p(\mathbf{T}|\mathbf{w}_1, \dots, \mathbf{w}_K) = \prod_{n=1}^N \prod_{k=1}^K p(C_k|\phi_n)^{t_{nk}} = \prod_{n=1}^N \prod_{k=1}^K y_{nk}^{t_{nk}} \quad (26)$$

where $y_{nk} = y_k(\phi_n)$, T is an $N \times K$ matrix of target variables

Multiclass Logistic Regression – III

- For target variable \mathbf{t}_n ($K \times 1$ vector) for data ϕ_n is defined as follows:

$$\mathbf{t}_n(k) = \begin{cases} 1 & \text{if } \phi_n \in \mathcal{C}_k \\ 0 & \text{otherwise} \end{cases} \quad (27)$$

And matrix $\mathbf{T}^T = [\mathbf{t}_1, \dots, \mathbf{t}_N]$

- Based on the likelihood function, we can define the error function

$$E(\mathbf{w}_1, \dots, \mathbf{w}_K) = -\ln p(\mathbf{T} | \mathbf{w}_1, \dots, \mathbf{w}_K) = -\sum_{n=1}^N \sum_{k=1}^K t_{nk} \ln y_{nk} \quad (28)$$

- (28) is the cross-entropy function defined for multi-class problem

Multiclass Logistic Regression – IV

- Like two-class problem, we can still use Newton-Raphson algorithm to estimate parameters $\{\mathbf{w}_k\}$
- To use Newton-Raphson algorithm, we need the closed form of gradient and Hessian matrix of the error function
- The gradient of the error function is

$$\nabla_{\mathbf{w}_j} E(\mathbf{w}_1, \dots, \mathbf{w}_K) = \sum_{n=1}^N (y_{nj} - t_{nj}) \phi_n \quad (29)$$

- The Hessian is

$$\nabla_{\mathbf{w}_k} \nabla_{\mathbf{w}_j} E(\mathbf{w}_1, \dots, \mathbf{w}_K) = - \sum_{n=1}^N y_{nk} (I_{kj} - y_{nj}) \phi_n \phi_n^T \quad (30)$$

Probit Regression – I

- In generalized linear model, we have

$$p(t = 1|a) = f(a) \quad (31)$$

where $a = \mathbf{w}^T \phi$, and f is the activation function

- In probit regression, a noisy threshold model is considered
For each input ϕ_n , we evaluate $a_n = \mathbf{w}^T \phi_n$, and the target variable is defined as:

$$t_n = \begin{cases} 1 & \text{if } a_n \geq \theta \\ 0 & \text{otherwise} \end{cases} \quad (32)$$

where θ is drawn from a probability density $p(\theta)$

- Thus the corresponding activation function is given as follows:

$$f(a) = \int_{-\infty}^a p(\theta) d\theta \quad (33)$$

Probit Regression – II

- A special example $p(\theta) \sim \mathcal{N}(\theta|0, 1)$
- Probit function

$$\Phi(a) = \int_{-\infty}^a \mathcal{N}(\theta|0, 1) d\theta \quad (34)$$

- Probit function is the cumulative distribution function of the standard normal distribution
- $\Phi(a)$ is a monotonic increasing function, and $0 < \Phi(a) < 1$ for $\forall a \in \mathbb{R}$

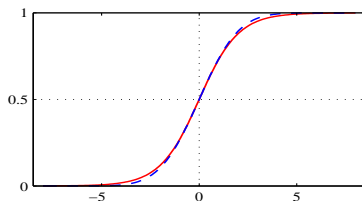


Figure: Plot of the probit function $\Phi(\lambda a)$, and $\lambda^2 = \pi/8$ to make sure the derivatives of both lines are equal for $a = 0$ (Blue Line)

Probit Regression – III

- The use of a more general Gaussian distribution does not change the model because it is equivalent to a re-scaling of the linear coefficients \mathbf{w}
- Many numerical packages provide for the evaluation of a function called “erf function” or “error function”

$$\text{erf}(a) = \frac{2}{\sqrt{\pi}} \int_0^a \exp\left(-\frac{\theta^2}{2}\right) d\theta \quad (35)$$

- We can use erf function to compute the probit function

$$\Phi(a) = \frac{1}{2} \left\{ 1 + \frac{1}{2} \text{erf}(a) \right\} \quad (36)$$

- Formally, **Probit Regression** is given below:

$$p(t = 1|a) = \Phi(a) = \Phi(\mathbf{w}^T \phi) \quad (37)$$

Probit Regression – IV

- In practice, we can also determine the parameter \mathbf{w} by using maximum likelihood method
- The results found using probit regression tend to be similar to those of logistic regression
- Compared with logistic regression, probit regression is more sensitive to outliers

Canonical Link Functions – I

- **Canonical Link Function** is a general framework of assuming a conditional distribution for the target variable from the exponential family, along with a corresponding choice for the activation function known as the canonical link function
- Cross-entropy error function can be considered as a special case of canonical link function
- Assumption of canonical link function:
The target variable t satisfy some exponential family distribution.
Formally,

$$p(t|\eta, s) = \frac{1}{s} h\left(\frac{t}{s}\right) g(\eta) \exp\left\{\frac{\eta t}{s}\right\} \quad (38)$$

where s is a scale parameter

Canonical Link Functions – II

- y is defined as the conditional mean of t

$$y \equiv \mathbb{E}[t|\eta] = -s \frac{d}{d\eta} g(\eta) \quad (39)$$

- It means that y and η must be related, and we denote this relation through

$$\eta = \psi(y) \quad (40)$$

- The generalized linear model

$$y = f(\mathbf{w}^T \phi) \quad (41)$$

where f is the **activation function** in machine learning literature, and f^{-1} is known as the **link function** in statistics

Canonical Link Functions – III

- Similarly, for a training set, we can compute the log likelihood function

$$\ln p(\mathbf{t}|\eta, s) = \sum_{n=1}^N \ln p(t_n|\eta, s) = \sum_{n=1}^N \left\{ \ln g(\eta_n) + \frac{\eta_n t_n}{s} \right\} + \text{const} \quad (42)$$

where we assume s is a scale parameter shared by all observations, just like the covariance matrix Σ in Gaussian distribution, thus s is independent of n

Canonical Link Functions – IV

- The derivative of the log likelihood with respect to \mathbf{w} is

$$\nabla_{\mathbf{w}} \ln p(\mathbf{t}|\eta, s) = \sum_{n=1}^N \left\{ \frac{d}{d\eta_n} \ln g(\eta_n) + \frac{t_n}{s} \right\} \frac{d\eta_n}{dy_n} \frac{dy_n}{da_n} \nabla a_n \quad (43)$$

$$= \sum_{n=1}^N \frac{1}{s} (t_n - y_n) \psi'(y_n) f'(a_n) \phi_n \quad (44)$$

where $a_n = \mathbf{w}^T \phi_n$, $y_n = f(a_n)$, $\eta = \psi(y)$, and $y = -s \frac{d}{d\eta} g(\eta)$

- If we choose a particular form for the link function $f^{-1}(y)$, (44) will be simplified

Canonical Link Functions – V

- A choice of link function $f^{-1}(y)$

$$f^{-1}(y) = \psi(y) \quad (45)$$

- Some properties of this link function

$$f(\psi(y)) = y \Rightarrow f'(\psi)\psi'(y) = 1 \quad (46)$$

$$a = f^{-1}(y) \Rightarrow a = \psi \Rightarrow f'(a)\psi'(y) = 1 \quad (47)$$

- The gradient of the error function can be simplified:

$$\nabla \ln E(\mathbf{w}) = \frac{1}{s} \sum_{n=1}^N (y_n - t_n) \phi_n \quad (48)$$

- For Gaussian $s = \beta^{-1}$
- For logistic model, $s = 1$

- 1 Probabilistic Discriminative Models
- 2 The Laplace Approximation
- 3 Bayesian Logistic Regression

Basic Idea of Laplace Approximation

The basic idea of Laplacian Approximation is to use some Gaussian distribution to approximate a probability density defined over a set of **continuous** variables

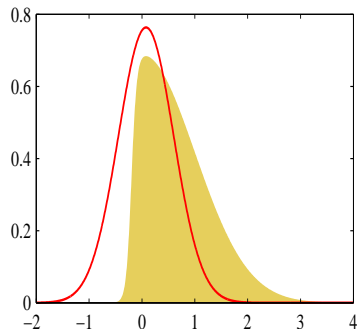


Figure: The illustration of Laplace Approximation

Laplace Approximation for 1-dimensional Space

- The original distribution

$$p(z) = \frac{1}{Z}f(z), \quad \text{where } Z = \int f(z)dz \quad (49)$$

- Detail procedure of Laplace approximation

Step 1 Find a point z_0 such that $p'(z_0) = 0$ or $f'(z_0) = 0$

Step 2 Use Taylor expansion of $\ln f(z)$ centered on z_0 to find the coefficient of quadratic term in Gaussian distribution

$$\ln f(z) \approx \ln f(z_0) - \frac{1}{2}A(z - z_0)^2 \quad \text{where } A = -\frac{d^2}{dz^2} \ln f(z)|_{z=z_0}$$

$$\therefore f(z) \approx f(z_0) \exp\left(-\frac{1}{2}A(z - z_0)^2\right)$$

Laplace Approximation for 1-dimensional Space

- The normalized distribution $q(z)$ is given as follows:

$$q(z) = \left(\frac{A}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2}A(z - z_0)^2\right) \quad (50)$$

- Requirement of Laplace Approximation
 - According to the definition of Gaussian distribution, $A > 0$
 - In other words, z_0 must be a local maximum

Laplacian Approximation for M -dimensional Space

- In M -dimensional space, also suppose the original distribution is

$$p(\mathbf{z}) = \frac{1}{Z} f(\mathbf{z}) \quad (51)$$

- Detail procedure of Laplace approximation

Step 1 Find a stationary point \mathbf{z}_0 the gradient $\nabla f(\mathbf{z}) = 0$

Step 2 Use Taylor expansion of $\ln f(\mathbf{z})$ centered on \mathbf{z}_0 to find the coefficient of quadratic term in Gaussian distribution

$$\ln f(\mathbf{z}) \approx \ln f(\mathbf{z}_0) - \frac{1}{2}(\mathbf{z} - \mathbf{z}_0)^T A(\mathbf{z} - \mathbf{z}_0)$$

$$\text{where } A = \nabla \nabla f(\mathbf{z})|_{\mathbf{z}=\mathbf{z}_0}$$

$$\therefore f(\mathbf{z}) \approx f(\mathbf{z}_0) \exp\left(-\frac{1}{2}(\mathbf{z} - \mathbf{z}_0)^T A(\mathbf{z} - \mathbf{z}_0)\right)$$

Laplace Approximation for M -dimensional Space

- The normalized distribution $q(\mathbf{z})$ is given as follows:

$$q(\mathbf{z}) = \frac{|A|^{1/2}}{(2\pi)^{M/2}} \exp\left(-\frac{1}{2}(\mathbf{z} - \mathbf{z}_0)^T A (\mathbf{z} - \mathbf{z}_0)\right) \quad (52)$$

- Requirement of Laplace Approximation
 - According to the definition of Gaussian distribution, A is positive definite
 - In other words, \mathbf{z}_0 must be a local maximum

- 1 Probabilistic Discriminative Models
- 2 The Laplace Approximation
- 3 Bayesian Logistic Regression

Introduction to Bayesian Logistic Regression

- Exact Bayesian inference for logistic regression is intractable
Evaluation of the posterior distribution would require normalization of the product of a prior distribution and a likelihood function that itself comprises a product of logistic sigmoid functions, one for every data point
- Basic idea of Bayesian Logistic Regression
Use Laplace approximation to the problem of Bayesian logistic regression

Laplace Approximation of the Posterior Distribution $p(\mathbf{w}|\mathbf{t})$

- The prior of \mathbf{w} is Gaussian

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0) \quad (53)$$

where \mathbf{m}_0 and \mathbf{S}_0 are both fixed hyperparameters

- Using Bayes Theorem, the posterior distribution over \mathbf{w} is

$$p(\mathbf{w}|\mathbf{t}) \propto p(\mathbf{w})p(\mathbf{w}|\mathbf{t}) \quad (54)$$

where $\mathbf{t} = (t_1, \dots, t_N)^T$

- Or equivalently,

$$\ln p(\mathbf{w}|\mathbf{t}) = -\frac{1}{2}(\mathbf{w} - \mathbf{m}_0)^T \mathbf{S}_0^{-1}(\mathbf{w} - \mathbf{m}_0) \quad (55)$$

$$+ \sum_{n=1}^N \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\} \quad (56)$$

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

Laplace Approximation of the Posterior Distribution $p(\mathbf{w}|\mathbf{t})$

- Maximize the posterior distribution to give the MAP (maximum posterior) solution \mathbf{w}_{map} , which defines the mean of the Gaussian
- The covariance is then given by the inverse of the matrix of second derivatives of the negative log likelihood, which takes the form

$$\mathbf{S}_N = -\nabla\nabla \ln p(\mathbf{w}|\mathbf{t}) = \mathbf{S}_0^{-1} + \sum_{n=1}^N y_n(1 - y_n)\phi_n\phi_n^T \quad (57)$$

- The Gaussian approximation to the posterior distribution

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{w}_{\text{map}}, \mathbf{S}^{-1}) \quad (58)$$

Predictive Distribution – I

- The predictive distribution for class \mathcal{C}_1 , given a new feature vector $\phi(\mathbf{x})$, is obtained by marginalizing with respect to the posterior distribution $p(\mathbf{w}|\mathbf{t})$

$$p(\mathcal{C}_1|\phi, \mathbf{t}) = \int p(\mathcal{C}_1|\phi, \mathbf{w})p(\mathbf{w}|\mathbf{t})d\mathbf{w} \approx \int \sigma(\mathbf{w}^T\phi)q(\mathbf{w})d\mathbf{w} \quad (59)$$

where $p(\mathcal{C}_1|\phi, \mathbf{w}) = \sigma(\mathbf{w}^T\phi)$ and $p(\mathbf{w}|\mathbf{t}) \approx q(\mathbf{w})$

- Denoting $a = \mathbf{w}^T\phi$, we have

$$\sigma(\mathbf{w}^T\phi) = \int \delta(a - \mathbf{w}^T\phi)\phi(a)da \quad (60)$$

where δ is the Dirac delta function

Predictive Distribution – II

- The predictive distribution can be represented as

$$p(\mathcal{C}_1|\phi, \mathbf{t}) \approx \int \sigma(\mathbf{w}^T \phi) q(\mathbf{w}) d\mathbf{w} = \int \sigma(a) p(a) da \quad (61)$$

where

$$p(a) = \int \delta(a - \mathbf{w}^T \phi) q(\mathbf{w}) d\mathbf{w} \quad (62)$$

- Actually, the distribution $p(a)$ is also Gaussian, and we can evaluate the mean and covariance of this distribution by taking moments, and interchanging the order of integration over a and \mathbf{w}

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

$$\sigma_a^2 = \text{var}[a] = \int p(a) (a^2 - \mathbb{E}[a]^2) da$$

$$= \int q(\mathbf{w}) \left((\mathbf{w}^T \phi)^2 - (\mathbf{m}_N^T \phi)^2 \right) d\mathbf{w} = \phi^T \mathbf{S}_n \phi$$

Predictive Distribution – III

- Therefore,

$$\int \sigma(a)p(a)da = \int \sigma(a)\mathcal{N}(a|\mu_a, \sigma_a^2)da \quad (63)$$

- The integral over a represents the convolution of a Gaussian with a logistic sigmoid, and cannot be evaluated analytically.
- A good approximation can be obtained if we use $\Phi(a)$ to approximate $\sigma(a)$
- If the two functions are required to have the same slope at the origin, then the corresponding probit function $\Phi(a)$ is

$$\Phi(\lambda a), \text{ where } \lambda^2 = \frac{\pi}{8} \quad (64)$$

Predictive Distribution – IV

- If we replace $\sigma(a)$ with $\Phi(\lambda a)$, then we have

$$\int \Phi(\lambda a) \mathcal{N}(a|\mu, \sigma^2) da = \Phi\left(\frac{\mu}{(\lambda^{-2} + \sigma^2)^{1/2}}\right) \quad (65)$$

- Use $\sigma(a)$ to replace $\Phi(\lambda a)$ in both sides of (65), we get

$$\int \sigma(a) \mathcal{N}(a|\mu, \sigma^2) da \approx \sigma(\kappa(\sigma^2)\mu) \quad (66)$$

where

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

- Finally we get

$$p(\mathcal{C}_1|\phi, \mathbf{t}) = \sigma(\kappa(\sigma^2)\mu) \quad (67)$$