# **Probit Regression**

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

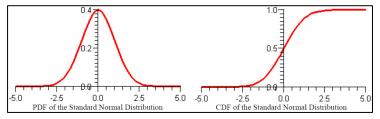
## Topics in Linear Classification using Probabilistic Discriminative Models

- Generative vs Discriminative
- 1. Fixed basis functions in linear classification
- 2. Logistic Regression (two-class)
- 3. Iterative Reweighted Least Squares (IRLS)
- 4. Multiclass Logistic Regression
- 5. Probit Regression
- 6. Canonical Link Functions

# Improving over Logistic Sigmoid

- For many class-conditionals, i.e., exponential family, posterior class probabilities are given by a logistic (or softmax) acting on a linear function of the feature variables
  - Gaussian class-conditional

# $p(\boldsymbol{x} \mid C_k) = \frac{1}{(2\pi)^{D/2}} \frac{1}{\left|\boldsymbol{\Sigma}\right|^{1/2}} \exp\left\{-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k)\right\}$

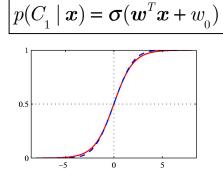


$$p(C_1 \mid \boldsymbol{x}) = \sigma \left( \ln \frac{p(\boldsymbol{x} \mid C_1) p(C_1)}{p(\boldsymbol{x} \mid C_2) p(C_2)} \right)$$

$$\mathbf{w} = \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

$$w_{_{0}} = -\frac{1}{2}\mu_{_{1}}^{^{T}}\boldsymbol{\Sigma}^{^{-1}}\boldsymbol{\mu}_{_{1}} + \frac{1}{2}\mu_{_{2}}^{^{T}}\boldsymbol{\Sigma}^{^{-1}}\boldsymbol{\mu}_{_{2}} + \ln\frac{p(C_{_{1}})}{p(C_{_{2}})}$$

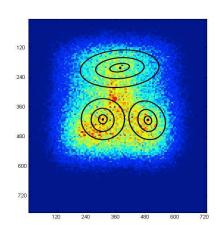
#### Sigmoid posterior

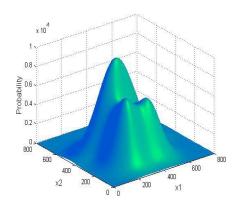


 However not all choices of class-conditional densities give rise to such a simple form for the posterior probabilities

### Gaussian Mixture Model

- Logistic transformation may not be suitable for GMMs
- Alternative discriminative model is based on probit function (which is the cdf of a zero-mean Gaussian)
  - Note that a cdf also goes between 0 and 1





GMM of spatial location of minutiae

## Motivating Alternate Link Function

Two-class case, Generalized Linear Model

$$p(t=1|a) = f(a)$$

where  $a = \mathbf{w}^{\mathrm{T}} \mathbf{\phi}$  and f (.) is the activation function

- Consider stochastic (noisy) threshold model
  - For input  $\varphi_n$ , evaluate  $a_n = \mathbf{w}^T \varphi_n$  and assign target value as

$$\begin{cases} t_n = 1 & \text{if } a_n \ge \theta \\ t_n = 0 & \text{otherwise} \end{cases}$$

– If  $\theta$  is drawn from a PDF  $p(\theta)$  then the corresponding activation function can be seen to be *equivalent* to the CDF

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

As illustrated next

## **Cumulative Distribution Function**

For a PDF  $p(\theta)$  the CDF is

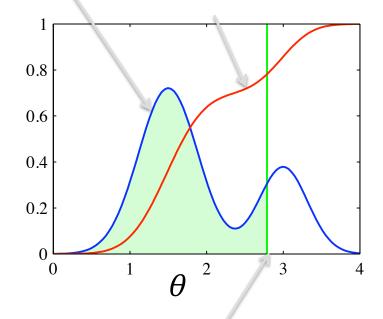
PDF  $p(\theta)$  is a mixture of two Gaussians

CDF *f*(*a*) is near-sigmoidal

$$f(a) = \text{Prob}(\theta \le a) = \int_{-\infty}^{a} p(\theta) d\theta$$

Thus CDF is equivalent to the Noisy threshold activation function

$$\begin{cases} t = 1 & \text{if } a = \mathbf{w}^T \mathbf{\phi} \ge \theta \\ t = 0 & \text{otherwise} \end{cases}$$

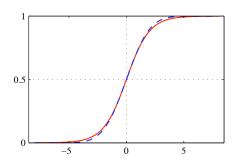


Value of blue curve at point *a* of vertical line is slope of red curve at that point Value of red curve is area under blue curve in shaded region

## Probit Function

#### It is the CDF of a zero-mean unit-variance Gaussian

$$\Phi(a) = \int_{-\infty}^{a} N(\theta \mid 0, 1) d\theta$$



It has a sigmoidal shape and compared to the logistic sigmoid It is closely related to the erf function which is usually tabulated

$$erf(a) = \frac{2}{\sqrt{\pi}} \int_{0}^{a} \exp\left(\frac{-\theta^{2}}{2}\right) d\theta$$
 It represents the probability that the error lies between  $\underline{+}$ 

that the error lies between  $\pm a$ 

with the relationship

erf is also sigmoidal in shape

$$\Phi(a) = \frac{1}{2} \left\{ 1 + \frac{1}{\sqrt{2}} \operatorname{erf}(a) \right\}$$

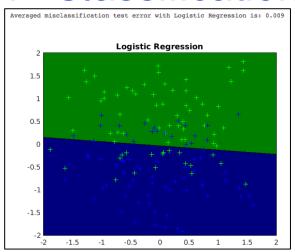
The generalized linear model based on probit activation is known as probit regression 7

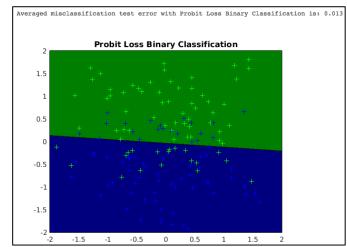
## **Probit Regression**

- Generalized Linear Model based on Probit activation function is known as Probit Regression
- We can determine its parameters using maximum likelihood
- Another application of probit is in Bayesian treatment of logistic regression

## Performance: Logistic vs Probit

#### 1. Classification results are similar





https://www.cs.ubc.ca/~schmidtm/ Software/matLearn/binaryclass/ demos/html/demo\_binaryclass\_ GLM.html

# 1. Probit is significantly more sensitive to outliers than logistic sigmoid

#### 2. Link functions:

map a probability in [0,1] to a value between -∞ and +∞

$$\label{eq:Logistic:f^1(\mu_y) = ln(\frac{p}{1-p})} \\ \text{Probit:} f^{-1}(\mu_y) = \Phi^{-1}(p)$$