CSE 802
Spring 2017
Logistic Regression

Inci M. Baytas
Computer Science
Michigan State University

March 29, 2017

Introduction

- Consider two-class classification problem, the posterior probability of class C_1 can be written as:

$$p\left(C_1|\Phi\right) = y\left(\Phi\right) = \sigma\left(\mathbf{w}^T\Phi\right) \tag{1}$$

- $\sigma\left(\cdot\right)$ is the logistic sigmoid function.
- $p\left(C_2|\Phi\right) = 1 - p\left(C_1|\Phi\right)$
- Φ is a feature vector, a non-linear transformation on original observation space $\mathbf{x}$.
- The model in Eq.1 is called as *Logistic Regression* in the terminology of statistics.

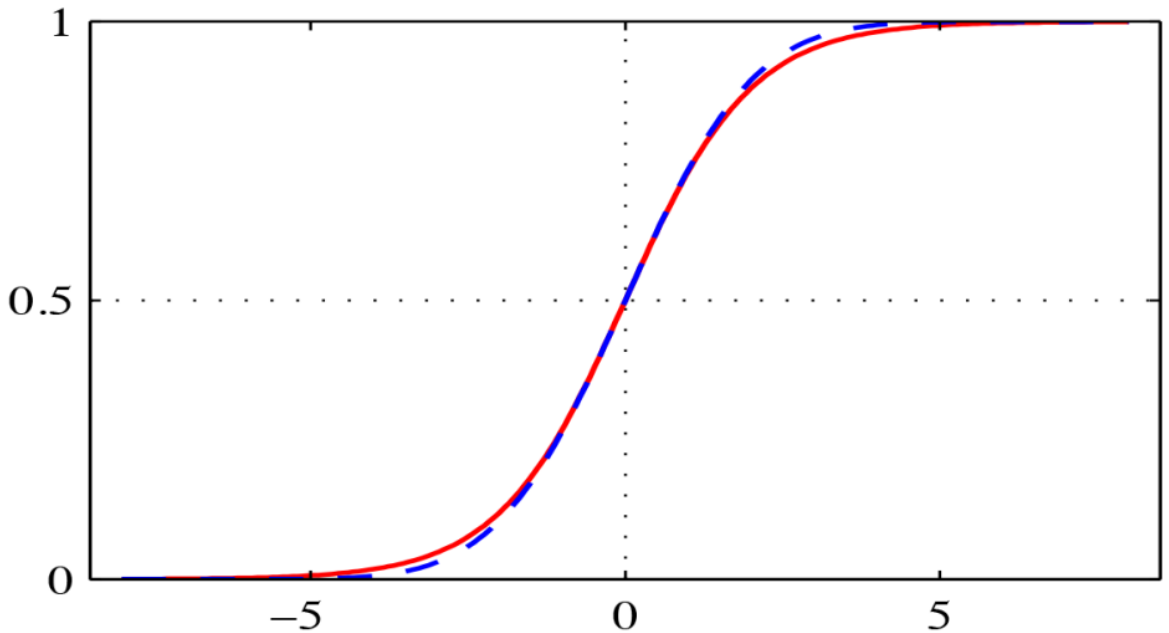
Logistic Regression I

- A classification model rather than regression
- A probabilistic discriminative model
  - We estimate the parameter **w** directly.
- Comparison of logistic regression and generative model in M-dimensional space Φ:
  - Logistic regression: M adjustable parameters.
  - Generative models: Assume we fit Gaussian class conditional densities using maximum likelihood; $M\left(M+5\right)/2+1=$ Means: $2M$ + Shared covariance: $\left(M+1\right)M/2$ + Prior $p\left(C_1\right)$: 1
- **Maximum likelihood** is used to determine the parameters of logistic regression model.

Logistic Regression II

- Definition and properties of logistic sigmoid function:

$$\sigma(a) = \frac{1}{1 + exp(-a)}$$
$$\sigma(-a) = 1 - \sigma(a) \tag{2}$$
$$\frac{d\sigma}{da} = \sigma(1 - \sigma)$$



Logistic Regression III - How to Estimate w

- For a training data set $\{\Phi_n, t_n\}$, where $t_n \in \{0, 1\}$ and $\Phi_n = \Phi(x_n)$, with $n = 1, ..., N$, the log likelihood can be written as:

$$p(\mathbf{t}|\mathbf{w}) = \prod_{n=1}^{N} y_n^{t_n} \{1 - y_n\}^{1-t_n} \tag{3}$$

where $\mathbf{t} = (t_1, ..., t_N)^T$ and $y_n = p(C_1|\Phi_n)$

- The error function is the negative logarithm of the likelihood, known as *Cross-entropy* error function:

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}) = -\sum_{n=1}^{N} \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\} \tag{4}$$

where $y_n = \sigma(a_n)$ and $a_n = \mathbf{w}^T \Phi_n$.

Logistic Regression IV - How to Estimate w

- There is no analytical (closed-form) solution.
- The cross entropy loss is a convex function.
  - There is a global minimum.
  - Can use an iterative approach.
- Calculate the gradient with respect to **w**:

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N} (y_n - t_n) \Phi_n \tag{5}$$

- Use gradient descent (batch or online):

$$\mathbf{w}^{\tau+1} = \mathbf{w}^{\tau} - \eta \nabla E(\mathbf{w}^{\tau}) \tag{6}$$

Logistic Regression V - How to Estimate w

- Newton-Raphson Algorithm

$$\mathbf{w}^{(\text{new})} = \mathbf{w}^{(\text{old})} - \mathbf{H}^{-1}\nabla E(\mathbf{w}) \tag{7}$$

- It uses a local quadratic approximation to the cross-entropy error function to update w iteratively.
- Newton-Raphson algorithm is also known as **iterative reweighted least squares**.
- Convexity: **H** is positive definite (eigenvalues of **H** are non-negative).

Multi-class Logistic Regression

- Cross-entropy for multi-class classification problem:

$$E\left(\mathbf{w}_1, ..., \mathbf{w}_K\right) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \ln y_{nk} \tag{8}$$

  where $y_k\left(\Phi\right) = p\left(C_k|\Phi\right) = \frac{\exp\left(\mathbf{w}_k^T\Phi\right)}{\sum_j \exp\left(\mathbf{w}_j^T\Phi\right)}$ which is called *softmax function*.
- Use maximum likelihood to estimate the parameters.
- Use an iterative approach such as Newton-Rapson.

Over-fitting in Logistic Regression

- Maximum likelihood can suffer from severe over-fitting.
- This can be overcome by finding a MAP solution for $\mathbf{w}$ (Bayesian treatment).
- Another alternative is to use regularization.
- Add regularizers to the loss function, regularized log-likelihood.
  - ℓ_2 norm
  - ℓ_1 norm (Lasso)

References

- Classification lecture of Dr. Jiayu Zhou.
- Christopher Bishop, Pattern Recognition and Machine Learning, *Information Science and Statistics*, Springer-Verlag New York, 2006.