Logistic Regression for Classification

Statistical Data Mining I

Rachael Hageman Blair

Outline

- Recap: Regression of an Indicator Matrix
- Motivation through Example
- The Logistic Function and Model
- Model Fitting
- Binary Classification
- Newton-Raphson
- Connection with LDA
- Example: Prostate Cancer
- Extensions: Penalized Logistic Regression

Recap

- Categorical data is not meant to carry quantitative values.
- The quantitative scheme is arbitrary, but implies a distance between “classes”.

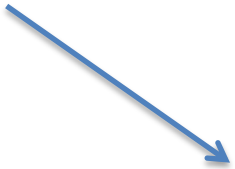
$$Y = \begin{cases} 1 & heroin \\ 2 & alcohol \\ 3 & cigarettes \\ 4 & cocaine \end{cases}$$

- Less of an issue for a 2-class problem.

Recap: Linear Regression of an Indicator Matrix

- If G has K classes, there will be K class indicators $y_K, k = 1 \ldots K$.

g	y1	y2	y3	y4
3	0	0	1	0
1	1	0	0	0
2	0	1	0	0
4	0	0	0	1
1	1	0	0	0



Indicator matrix

- The idea is to fit a regression model for each $y_K, k = 1 \ldots K$,

$$\hat{Y} = X(X^T X)^{-1} X^T Y.$$

- We need to estimate a coefficient vector for each response column (class) $Y(:,k)$ this yields a $(p+1) \times K$ coefficient matrix.

Recap: Linear Regression of an Indicator Matrix

- Define: $\hat{B} = (X^T X)^{-1} X^T Y \in \Re^{(p+1)\times K}$.
- For a new observation with input x, compute the fitted output:

$$\begin{aligned}\hat{f}(x) &= \left[(1, x)\hat{B}\right]^T \\ &= \left[(1, x_1, x_2, \ldots, x_p)\hat{B}\right]^T \\ &= \begin{pmatrix} \hat{f}_1(x) \\ \hat{f}_2(x) \\ \vdots \\ \hat{f}_K(x) \end{pmatrix}.\end{aligned}$$

- Find the largest component of $\hat{f}(x)$ and classify:

$$\hat{G}(x) = \arg\max_{k \in G} \hat{f}_k(x).$$

Recap: Linear Regression of an Indicator Matrix

- The linear regression Y_k on X is a linear approximation to $E(Y_k \mid X = x)$. Note that: $E(Y_k \mid X = x) = \Pr(G = k \mid X = x)$

- According to Bayes Rule, the optimal classifier is given as:

$$G^*(x) = \arg\max_{k \in G} \Pr(G = k \mid X = x).$$

- Linear regression of an indicator matrix:
  - Approximate $\Pr(G = k \mid X = x)$ by a linear function of x using linear regression.

The question is, how well do this do?

Recap: Linear Regression of an Indicator Matrix

- The question: Are the $\hat{f}_k(x)$ functions good approximations of the posterior probabilities?

- Note: $\sum_{k \in G} \hat{f}_k(x) = 1$ for any x when there is an intercept in the model.

- However, $\hat{f}_k(x)$ can be negative or greater than one. This is especially true if we make predictions outside of the training set.

- These observations do not make the model invalid, or suggest it is not working.

Logistic Regression

Motivation: desire to model the posterior probabilities of the K classes via linear functions in x, and ensure that they sum to one and remain in [0,1].

Logistic Regression

Data: Customer Default records from a credit card company. For simplicity, assume a one dimensional predictor balance, for the default classification.

Question: How should we model the default probability:

$$p(X) = \Pr(G = 1 \mid X).$$

Logistic Regression

Data: Customer Default records from a credit card company. For simplicity, assume a one dimensional predictor balance, for the default classification.

Question: How should we model the default probability:

$$p(X) = \Pr(G = 1 \mid X).$$

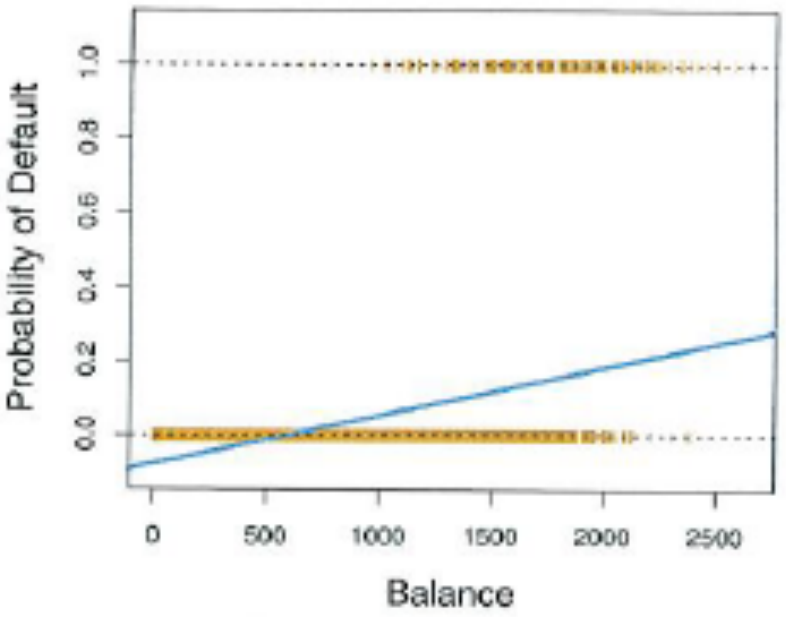
We could use a linear model (regression on an indicator):

$$p(X) = \beta_0 + \beta_1 X.$$

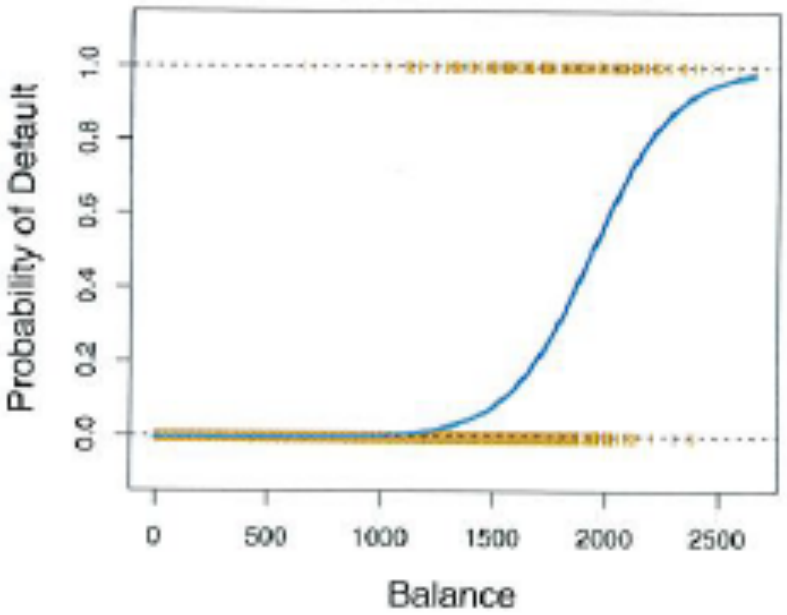
… tempting but…

Logistic Regression

Linear Regression

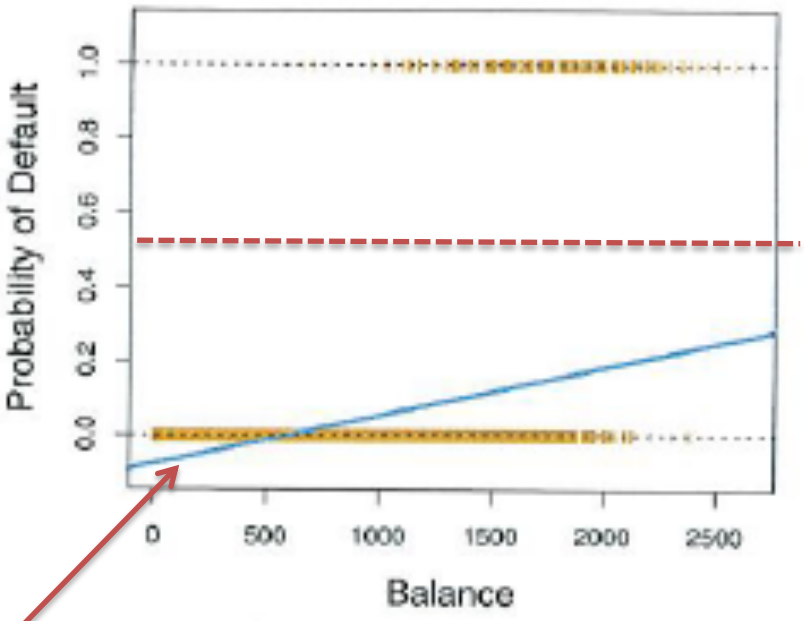


Logistic Regression

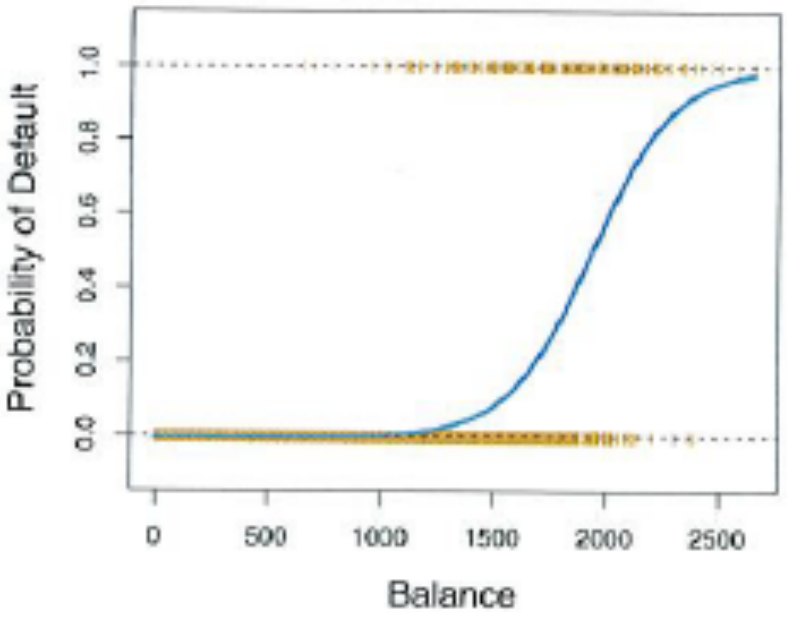


Logistic Regression

Linear Regression

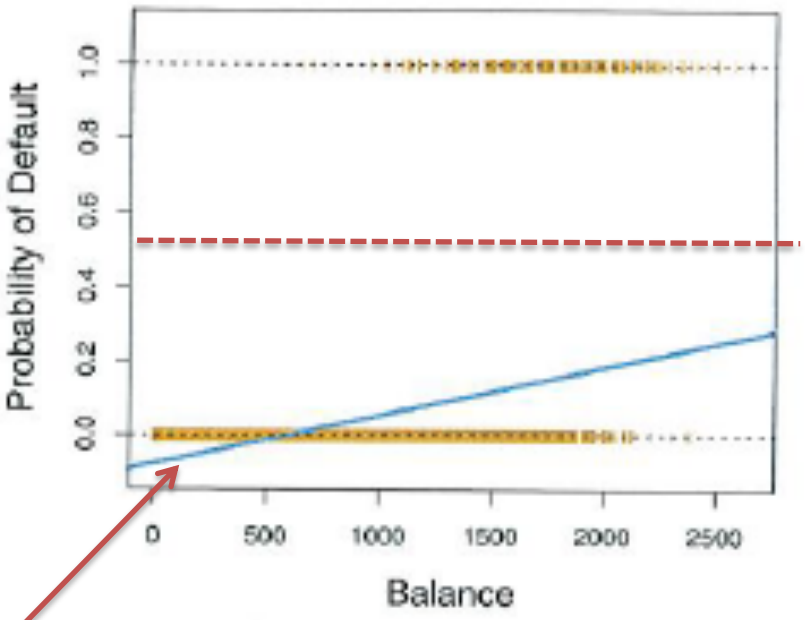


Logistic Regression

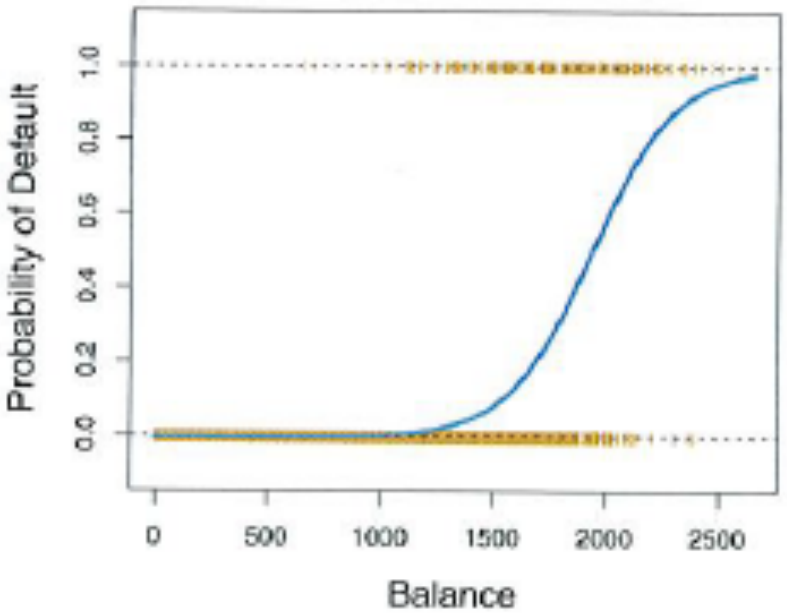


Logistic Regression

Linear Regression



Logistic Regression



The Logistic Function

$$p(X) = \frac{\exp(\beta_0 + \beta_1 X).}{1 + \exp(\beta_0 + \beta_1 X)}$$

Logistic Regression

The Logistic Function:

$$p(X) = \frac{\exp(\beta_0 + \beta_1 X).}{1 + \exp(\beta_0 + \beta_1 X)}$$

With some manipulation:

$$\frac{p(X)}{1 - p(x)} = \exp(\beta_0 + \beta_1 X)$$

Odds ratio

$$\log\left(\frac{p(X)}{1 - p(x)}\right) = \beta_0 + \beta_1 X.$$

Log-odds ratio aka “logit” function

Logistic Regression

Odds:

$$\frac{p(x)}{1-p(x)}$$

Can take values $[0, \infty)$

Example: $p(X) = 0.20$ is the probability of default, then $\frac{0.20}{(1-0.20)} = \frac{1}{4}$ is the odds of default.

Example: If 9 out of 10 people default, then $p(X) = 0.90$ implies an odds of $\frac{0.90}{(1-0.90)} = 9$.

Logistic Regression

- By the Bayes rule:

$$\hat{G}(x) = \arg\max_k \Pr(G = k \mid X = x).$$

- Decision boundary between class k and l is determined by:

$$\Pr(G = k \mid X = x) = \Pr(G = l \mid X = x).$$

- Divide both sides by $\Pr(G = l \mid X = x)$ and take the logarithm.

$$\log \frac{\Pr(G = k \mid X = x)}{\Pr(G = l \mid X = x)} = 0.$$

Logistic Regression

- We want to model the posterior of the K classes in terms of linear functions in X and preserve certain properties.
- Assuming a linear boundary, we can write:

$$\log \frac{\Pr(G=k \mid X=x)}{\Pr(G=l \mid X=x)} = \beta_0^{(k,l)} + \sum_{j=1}^{p} \beta_j^{(k,l)} x_j .$$

There are restrictive relations between these coefficients for different (k,l) combinations.

Logistic Regression

- The model has the form:

$$\log\frac{\Pr(G=1\mid X=x)}{\Pr(G=K\mid X=x)}=\beta_{10}+\beta_1^T x$$

$$\log\frac{\Pr(G=2\mid X=x)}{\Pr(G=K\mid X=x)}=\beta_{20}+\beta_2^T x$$

$$\vdots$$

$$\log\frac{\Pr(G=K-1\mid X=x)}{\Pr(G=K\mid X=x)}=\beta_{(K-1)0}+\beta_{K-1}^T x$$

Number of parameters: $(K-1)(p+1)$.

Denote the entire parameter set: $\theta=\{\beta_{10},\beta_1,\beta_{20},\beta_2,\ldots,\beta_{(K-1)0},\beta_{(K-1)}\}$.

Logistic Regression

$$\theta = \begin{pmatrix} \beta_{10} \\ \beta_{1} \\ \beta_{20} \\ \beta_{2} \\ \vdots \\ \beta_{(K-1)0} \\ \beta_{K-1} \end{pmatrix} = \begin{pmatrix} \beta_{10} \\ \beta_{11} \\ \vdots \\ \beta_{1p} \\ \beta_{20} \\ \vdots \\ \beta_{2p} \\ \vdots \\ \beta_{(K-1)0} \\ \vdots \\ \beta_{(K-1)_p} \end{pmatrix}$$

Logistic Regression

- The posterior probabilities are given by:

$$\Pr(G=k \mid X=x)=\frac{\exp\left(\beta_{k0}+\beta_k^T x\right)}{1+\sum_{l=1}^{K-1}\exp\left(\beta_{l0}+\beta_l^T x\right)} \quad \text{for} \quad k=1,\ldots,K-1.$$

$$\Pr(G=K \mid X=x)=\frac{1}{1+\sum_{l=1}^{K-1}\exp\left(\beta_{l0}+\beta_l^T x\right)}$$

- It can be shown that (homework):

$$\sum_{k=1}^{K}\Pr(G=k \mid X=x)=1$$

Comparison with LR on Indicators

- Similarities:
  - Both attempt to estimate $\Pr(G = k \mid X = x)$.
  - Both have linear classification boundaries.

- Differences:
  - **Linear regression** on an indicator matrix: we approximate $\Pr(G = k \mid X = x)$ by a *linear* function of x. The property of falling between 0 and 1 are NOT guaranteed.

  - **Logistic regression**: $\Pr(G = k \mid X = x)$ is approximated by a *nonlinear* function of x. It is guaranteed to range from 0 to 1, and sum to 1.

A note on popularity..

Logistic regression very popular in Applied Statistics for discrete data, why?

- Tradition
- Relationship to the log-odds ratio, and statistical properties of probability.
- Performs well in practice compared to other classifiers.

*Mostly used with a binary response (K=2).

Fitting Logistic Regression Models

- Criteria: find parameters that maximize the conditional likelihood of *G* given *X*.
- Denote $p_k(x_i;\theta) = \Pr(G = k \mid X = x_i;\theta)$.
- Given the first input x_1, the posterior probability of its class being g_1 is $\Pr(G = g_1 \mid X = x_1)$.
- Since samples in the training data set are independent, the posterior probability for the N samples each having class g_i, $i = 1,2,\ldots,N$, given their inputs $x_1, x_2, \ldots, x_N$ is:

$$\prod_{i=1}^{N} \Pr(G = g_i \mid X = x_i).$$

Fitting Logistic Regression Models

- The conditional log-likelihood of the class labels in the training data set is:

$$L(\theta) = \sum_{i=1}^{N} \log \Pr(G = g_i \mid X = x_i)$$

$$= \sum_{i=1}^{N} \log p_{g_i}(x_i; \theta).$$

Binary Classification K=2

- For binary classification, if $g_i = 1$, denote $y_i = 1$; if $g_i = 2$, denote $y_i = 0$ (WLOG).

- Let $p_1(x;\theta) = p(x;\theta)$, the

$$p_2(x;\theta) = 1 - p_1(x;\theta) = 1 - p(x;\theta).$$

- Since $K = 2$, the parameters are $\theta = \{\beta_{10}, \beta_1\}$. We write them in vector form: $\beta = (\beta_{10}, \beta_1)^T$.

Binary Classification

- If $y_i = 1$, i.e., $g_i = 1$,

$$\begin{aligned}\log p_{g_i}(x;\beta) &= \log p_1(x;\beta) \\ &= 1 \cdot \log p(x;\beta) \\ &= y_i \log p(x;\beta).\end{aligned}$$

- If $y_i = 0$, i.e., $g_i = 2$,

$$\begin{aligned}\log p_{g_i}(x;\beta) &= \log p_2(x;\beta) \\ &= 1 \cdot \log\left(1 - p(x;\beta)\right) \\ &= (1 - y_i)\log\left(1 - p(x;\beta)\right).\end{aligned}$$

- Since either $y_i = 0$, or $1 - y_i = 0$, we have the log-likelihood:

$$\log p_{g_i}(x;\beta) = y_i \log p(x;\beta) + (1 - y_i)\log\left(1 - p(x;\beta)\right).$$

Binary Classification

- The log-likelihood (in general):

$$L(\beta) = \sum_{i=1}^{N} \log p_{g_i}(x_i;\beta)$$

$$= \sum_{i=1}^{N} \left[y_i \log p(x_i;\beta) + (1-y_i) \log\left(1 - p(x_i;\beta)\right) \right]$$

Binary Classification

- Recall: under the assumptions of logistic regression model:

$$\Pr(G = k \mid X = x) = \frac{\exp\left(\beta_{k0} + \beta_k^T x\right)}{1 + \sum_{l=1}^{K-1} \exp\left(\beta_{l0} + \beta_l^T x\right)}$$

$$\Pr(G = K \mid X = x) = \frac{1}{1 + \sum_{l=1}^{K-1} \exp\left(\beta_{l0} + \beta_l^T x\right)}$$

- Substitute the above in $L(\beta)$:

$$L(\beta) = \sum_{i=1}^{N} \left[y_i \beta^T x_i - \log\left(1 + \exp(\beta^T x_i)\right) \right].$$

Binary Classification

- In matrix form, we write:

$$\frac{\partial L(\beta)}{\partial \beta_{1j}} = \sum_{i=1}^{N} x_i \left(y_i - p(x_i;\beta) \right).$$

- To solve the set of $p+1$ nonlinear equations $\frac{\partial L(\beta)}{\partial \beta_{1j}} = 0$

  $j = 0,1,\ldots,p,$ use the **Newton-Raphson algorithm**.

- Requires the second derivatives or Hessian Matrix:

$$\frac{\partial^2 L(\beta)}{\partial \beta^2} = -\sum_{i=1}^{N} x_i x_i^T p(x_i;\beta)\left(1 - p(x_i;\beta)\right).$$

Newton-Raphson

- Newton-Raphson algorithm is an iterative method, providing updates to the parameter vector at every step:

$$\beta^{new} = \beta^{old} - \left(\frac{\partial^2 L(\beta)}{\partial \beta^2} \right)^{-1} \frac{\partial L(\beta)}{\partial \beta}.$$

Note: the derivatives are evaluated in the iteration before β^{old}.

Newton-Raphson (in computation)

- The iteration can be expressed in matrix form:

$$\frac{\partial L(\beta)}{\partial \beta} = X^T (y - p)$$

$$\frac{\partial^2 L(\beta)}{\partial \beta^2} = -XWX,$$

y is the column vector of y_i.

X is the $N \times (p+1)$ input matrix

p is the N vector of fitted probabilities with the ith element $p(x_i; \beta^{old})$.

W is an $N \times N$ diagonal matrix of weights with the ith element equal to $p(x_i; \beta^{old})(1 - p(x_i; \beta^{old}))$.

Newton-Raphson

- The Newton-Raphson step is:

$$\beta^{new} = \beta^{old} + \left(X^T W X\right)^{-1} X^T \left(y - p\right)$$

$$= \left(X^T W X\right)^{-1} X^T W z$$

where $z = X\beta^{old} + W^{-1}\left(y - p.\right)$

- If **z** is viewed as a response and **X** is the input matrix, β^{new} is the solution to a weighted LS problem:

$$\beta^{new} \leftarrow \arg\min_{\beta} \left(z - X\beta\right)^T W \left(z - X\beta\right).$$

- The algorithm is known as **Iteratively Reweighted Least Squares** (IRLS).

Psuedocode

1. $0 \rightarrow \beta$
2. Compute y by setting its elements:

$$y_i = \begin{cases} 1 \ if \ g_i = 1 \\ 0 \ if \ g_i = 2 \end{cases}, \quad i = 1 \ldots N.$$

3. Compute p by setting its elements equal to:

$$p(x_i;\beta) = \frac{e^{\beta^T x_i}}{1 + e^{\beta^T x_i}}, \quad i = 1, 2, \ldots, N.$$

4. Compute W, with elements: $w(i,i) = p(x_i;\beta)(1 - p(x_i;\beta))$.
5. $z \leftarrow X\beta + W^{-1}(y - p)$
6. $\beta \leftarrow (X^T W X)^{-1} X^T W z$
7. Stop if stopping criteria met, else, go to step 3.

Psuedocode

Note: Alternative Approaches Available for Computational Efficiency.

1. $0 \rightarrow \beta$
2. Compute y by setting its elements:

$$y_i = \begin{cases} 1 \ if \ g_i = 1 \\ 0 \ if \ g_i = 2 \end{cases}, \quad i = 1 \ldots N.$$

3. Compute p by setting its elements equal to:

$$p(x_i;\beta) = \frac{e^{\beta^T x_i}}{1 + e^{\beta^T x_i}}, \quad i = 1, 2, \ldots, N.$$

4. Compute W, with elements: $w(i,i) = p(x_i;\beta)(1 - p(x_i;\beta))$.
5. $z \leftarrow X\beta + W^{-1}(y - p)$
6. $\beta \leftarrow (X^T W X)^{-1} X^T W z$
7. Stop if stopping criteria met, else, go to step 3.

Example

Diabetes data set

- Input X is two dimensional.

  X_1 and X_2 are the first two principal components calculated from the original 8 variables.

- **Class1**: with diabetes and **Class2**: without diabetes
- Applying logistic regression, we obtain:

$$\beta = \left(0.7679, -0.6816, -0.38664\right)^T .$$

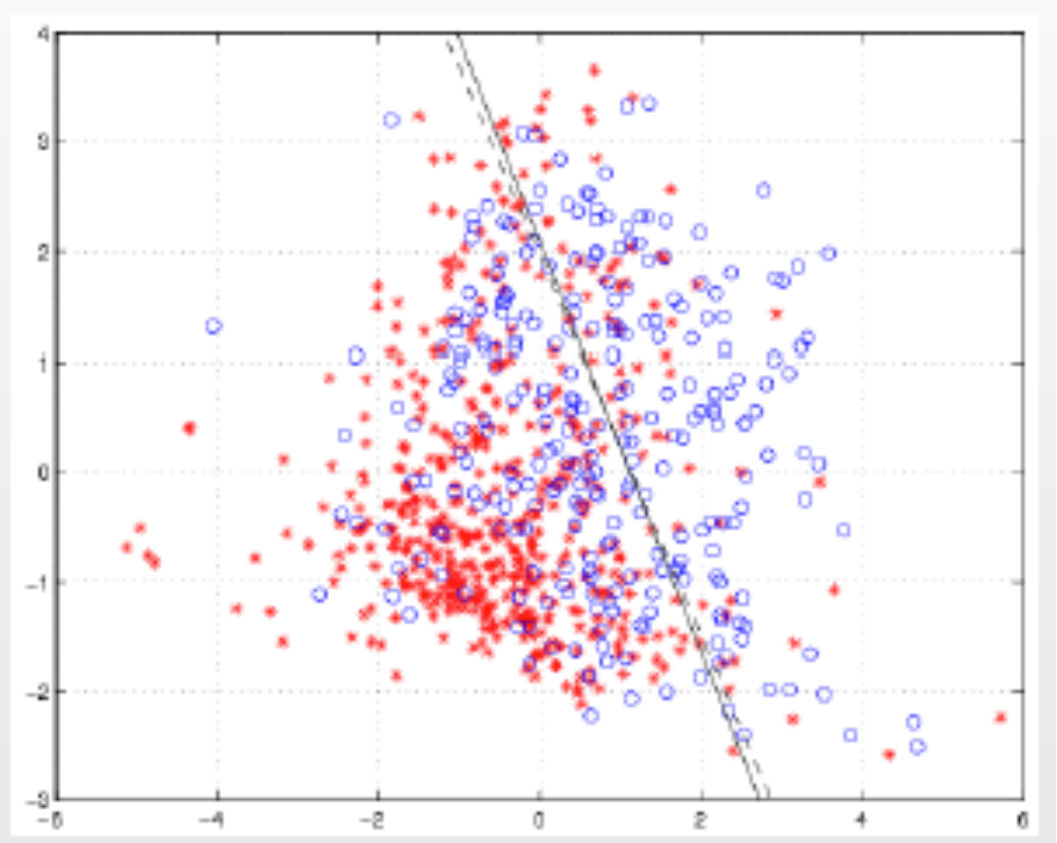
$$\Pr(G = 1 \mid X = x) = \frac{e^{0.7679 - 0.6816 X_1 - 0.3664 X_2}}{1 + e^{0.7679 - 0.6816 X_1 - 0.3664 X_2}}$$

$$\Pr(G = 2 \mid X = x) = \frac{1}{1 + e^{0.7679 - 0.6816 X_1 - 0.3664 X_2}}$$

Example

Dashed line – decision boundary
obtained by logistic regression.

Solid line – decision boundary
obtained by LDA.



Connection with LDA

- If the additional assumption made by LDA is appropriate, LDA tends to estimate the parameters more efficiently by using more information about the data.
- When classes are well-separated, parameter estimates for the logistic regression model are surprisingly unstable. LDA does not suffer from this problem.
- If N is small, the LDA model is again more stable.
- As logistic regression relies on fewer assumptions, and is more generalizable.
- For K>2 is not used much in practice.
- In practice, logistic regression and LDA often give similar results.

Logistic Regression: Implementation

glmnet: Lasso and elastic-net regularized generalized linear models

Extremely efficient procedures for fitting the entire lasso or elastic-net regularization path for linear regression, logistic and multinomial regression models, poisson regression and the Cox model. Two recent additions are the multiresponse gaussian, and the grouped multinomial. The algorithm uses cyclical coordinate descent in a pathwise fashion, as described in the paper listed below.

Version:	1.8
Depends:	Matrix (≥ 1.0-6), utils
Suggests:	survival
Published:	2012-07-03
Author:	Jerome Friedman, Trevor Hastie, Rob Tibshirani
Maintainer:	Trevor Hastie <hastie at stanford.edu>
License:	GPL-2
URL:	http://www.jstatsoft.org/v33/i01/.
Citation:	glmnet citation info
In views:	MachineLearning
CRAN checks:	glmnet results

Downloads:

Package source:	glmnet_1.8.tar.gz
MacOS X binary:	glmnet_1.8.tgz
Windows binary:	glmnet_1.8.zip
Reference manual:	glmnet.pdf
Vignettes:	Fitting the Penalized Cox Model
News/ChangeLog:	ChangeLog
Old sources:	glmnet archive

*Can handle large models efficiently and regularized models.

Logistic Regression example

- Traditional example: we want to understand the role of the input on a binary output.
- Data: Coronary Risk Factor Study (CORIS) in white males between 15 and 64.
- Goal: Establish the disease risk factors for ischemia in regions with high incidence.
- Response variable: the presence/absence of myocardial infraction.
- Risk Factors: systolic blood pressure (sbp), tobacco, ldl, famhist, obesity, alcohol, and age.

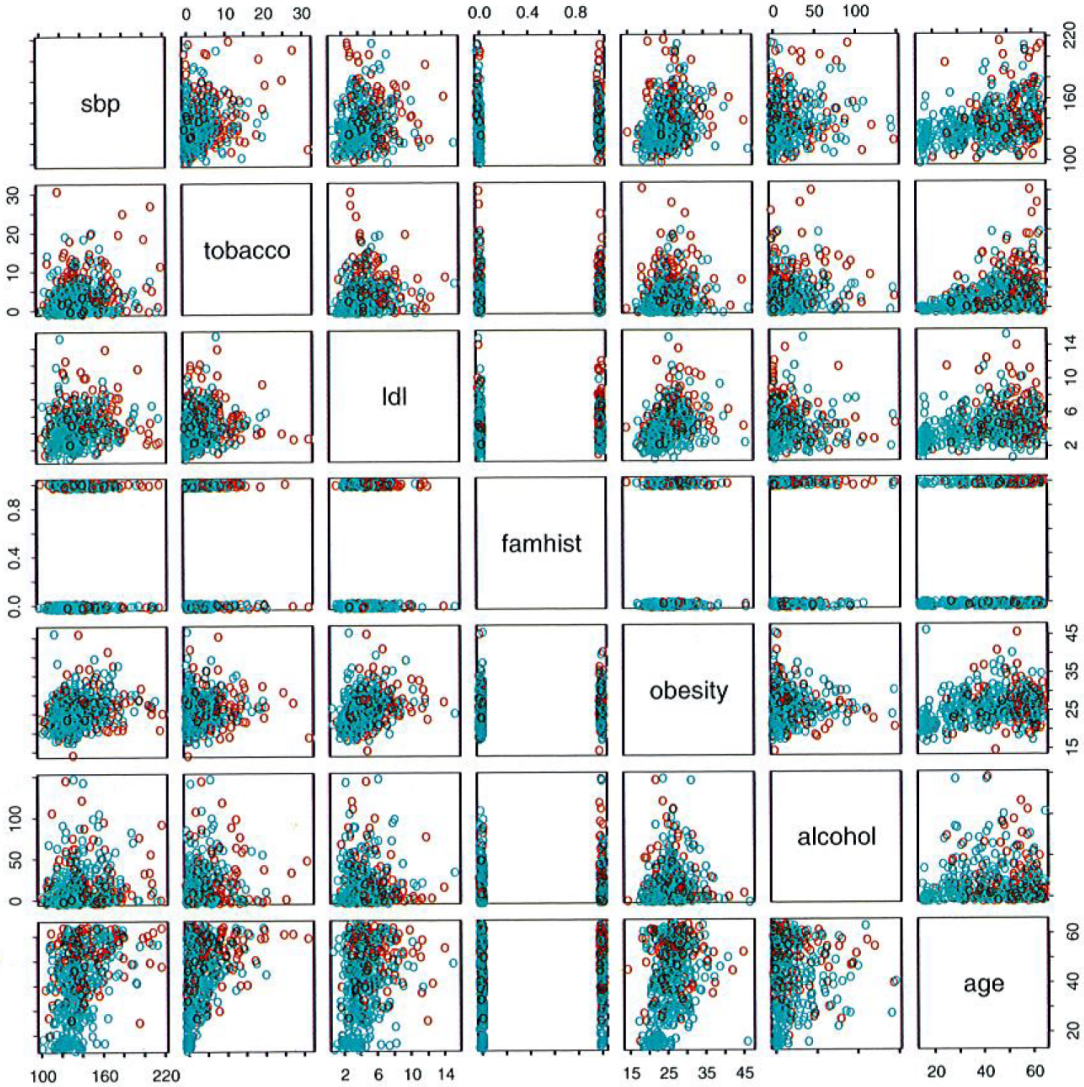


FIGURE 4.12. *A scatterplot matrix of the South African heart disease data. Each plot shows a pair of risk factors, and the cases and controls are color coded (red is a case). The variable* family history of heart disease (`famhist`) *is binary (yes or no).*

Logistic Regression example

Summary of results of the logistic regression:

TABLE 4.2. *Results from a logistic regression fit to the South African heart disease data.*

	Coefficient	Std. Error	Z Score
(Intercept)	−4.130	0.964	−4.285
sbp	0.006	0.006	1.023
tobacco	0.080	0.026	3.034
ldl	0.185	0.057	3.219
famhist	0.939	0.225	4.178
obesity	-0.035	0.029	−1.187
alcohol	0.001	0.004	0.136
age	0.043	0.010	4.184

Logistic Regression example

Summary of results of the logistic regression:

TABLE 4.2. *Results from a logistic regression fit to the South African heart disease data.*

	Coefficient	Std. Error	Z Score
(Intercept)	−4.130	0.964	−4.285
sbp	0.006	0.006	1.023
tobacco	0.080	0.026	3.034
ldl	0.185	0.057	3.219
famhist	0.939	0.225	4.178
obesity	-0.035	0.029	−1.187
alcohol	0.001	0.004	0.136
age	0.043	0.010	4.184

Summary of stepwise logistic regression:

TABLE 4.3. *Results from stepwise logistic regression fit to South African heart disease data.*

	Coefficient	Std. Error	Z score
(Intercept)	−4.204	0.498	−8.45
tobacco	0.081	0.026	3.16
ldl	0.168	0.054	3.09
famhist	0.924	0.223	4.14
age	0.044	0.010	4.52

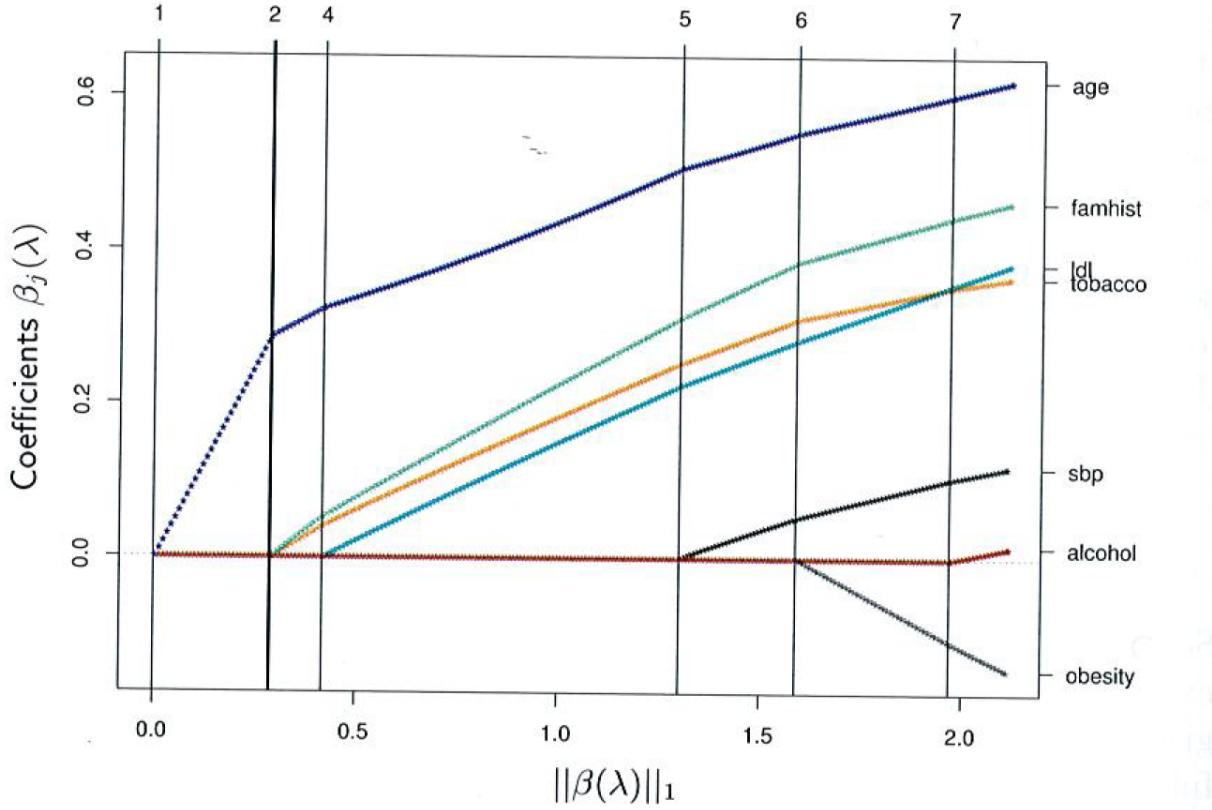
Penalized Logistic Regression

L1-regularized logistic regression –

$$\max\left\{\sum_{i=1}^{N}\left[y_i\left(\beta_0+\beta^T x_i\right)-\log\left(1+e^{\beta_0+\beta^T x_i}\right)\right]-\lambda\sum_{j=1}^{p}\left|\beta_j\right|\right\}.$$

- Path algorithms will not work because the coefficient profiles are piecewise smooth rather than piecewise linear.
- Coordinate descent methods are efficient: R package glmnet.

Penalized Logistic Regression



IGURE 4.13. *L_1 regularized logistic regression coefficients for the South frican heart disease data, plotted as a function of the L_1 norm. The variables ere all standardized to have unit variance. The profiles are computed exactly at ıch of the plotted points.*