MATH 829: Introduction to Data Mining and Analysis
Logistic regression

Dominique Guillot

Departments of Mathematical Sciences
University of Delaware

March 7, 2016

Logistic regression

Suppose we work with binary outputs, i.e., $y_i \in \{0,1\}$.

Linear regression may not be the best model.

- $x^T\beta \in \mathbb{R}$ not in $\{0,1\}$.
- Linearity may not be appropriate. Does doubling the predictor doubles the probability of $Y=1$? (e.g. probability of going to the beach vs outdoors temperature).

Logistic regression: Different perspective. Instead of modelling the $\{0,1\}$ output, we model the probability that $Y=0,1$.

Idea: We model $P(Y=1|X=x)$.

- Now: $P(Y=1|X=x) \in [0,1]$ instead of $\{0,1\}$.
- We want to relate that probability to $x^T\beta$.

We assume

$$\begin{aligned}\operatorname{logit}(P(Y=1|X=x)) &= \log \frac{P(Y=1|X=x)}{1-P(Y=1|X=x)} \\ &= \log \frac{P(Y=1|X=x)}{P(Y=0|X=x)} = x^T\beta.\end{aligned}$$

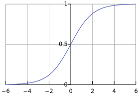
Logistic regression (cont.)

Equivalently,

$$P(Y=1|X=x) = \frac{e^{x^T\beta}}{1+e^{x^T\beta}}$$

$$P(Y=0|X=x) = 1 - P(Y=1|X=x) = \frac{1}{1+e^{x^T\beta}}$$

The function $f(x) = e^x/(1+e^x) = 1/(1+e^{-x})$ is called the *logistic function*.



$\log \frac{P(Y=1|X=x)}{P(Y=0|X=x)}$ is the *log-odds* ratio.

- Larger positive values of $x^T\beta \Rightarrow p \approx 1$.
- Larger negative values of $x^T\beta \Rightarrow p \approx 0$.

Logistic regression (cont.)

In summary, we are assuming:

- $Y|X=x \sim \text{Bernoulli}(p)$.
- $\operatorname{logit}(p) = \operatorname{logit}(E(Y|X=x)) = x^T\beta$.

More generally, one can use a *generalized linear model* (GLM). A GLM consists of:

- A probability distribution for $Y|X=x$ from the exponential family.
- A linear predictor $\eta = x^T\beta$.
- A *link function* g such that $g(E(Y|X=x)) = \eta$.

Logistic regression: estimating the parameters

In logistic regression, we are assuming a model for Y. We typically estimate the parameter β using maximum likelihood.

Recall: If $Y \sim \mathrm{Bernoulli}(p)$, then

$$P(Y = y) = p^y(1-p)^{1-y}, \qquad y \in \{0, 1\}.$$

Thus, $L(p) = \prod_{i=1}^n p^{y_i}(1-p)^{1-y_i}$.

Here $p = p(x_i, \beta) = \frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}}$. Therefore,

$$L(\beta) = \prod_{i=1}^n p(x_i, \beta)^{y_i}(1 - p(x_i, \beta))^{1-y_i}.$$

Taking the logarithm, we obtain

$$\begin{aligned} l(\beta) &= \sum_{i=1}^n y_i \log p(x_i, \beta) + (1 - y_i)\log(1 - p(x_i, \beta)) \\ &= \sum_{i=1}^n y_i(x_i^T\beta - \log(1 + x_i^T\beta)) - (1 - y_i)\log(1 + e^{x_i^T\beta}) \\ &= \sum_{i=1}^n [y_i x_i^T\beta - \log(1 + e^{x_i^T\beta})]. \end{aligned}$$

Logistic regression: estimating the parameters

Taking the derivative:

$$\frac{\partial}{\partial\beta_j} l(\beta) = \sum_{i=1}^n \left[y_i x_{ij} - x_{ij}\frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}} \right].$$

Needs to be solved using numerical methods (e.g. Newton-Raphson).

Logistic regression often performs well in applications.

As before, penalties can be added to regularize the problem or induce sparsity. For example,

$$\min_\beta -l(\beta) + \alpha\|\beta\|_1$$
$$\min_\beta -l(\beta) + \alpha\|\beta\|_2.$$

Example

South African Heart Disease (ESL):

- Subset of the Coronary Risk-Factor Study (CORIS) baseline survey.
- Carried out in three rural areas of the Western Cape, South Africa (Rousseauw et al., 1983).
- Aim of the study was to establish the intensity of ischemic heart disease risk factors in that high-incidence region
- Data represent white males between 15 and 64, and the response variable is the presence or absence of myocardial infarction (MI) at the time of the survey.
- 160 cases in dataset, and a sample of 302 controls.

Dataset variables

sbp	systolic blood pressure
tobacco	cumulative tobacco (kg)
ldl	low densiity lipoprotein cholesterol
adiposity	
famhist	family history of heart disease (Present, Absent)
typea	type-A behavior
obesity	
alcohol	current alcohol consumption
age	age at onset
chd	response, coronary heart disease

Example (cont.)

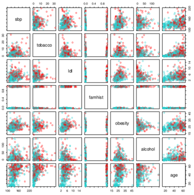


FIGURE 4.12. *A scatterplot matrix of the South African heart disease data. Each plot shows a pair of risk factors, and the cases and controls are color coded (red is a case). The variable family history of heart disease (famhist) is binary (yes or no).*

ESL

Example (cont.)

```
import pandas as pd
import numpy as np
from sklearn.linear_model import LogisticRegression
from sklearn.cross_validation import train_test_split

data = pd.read_csv('../../../data/SouthAfrica_Heart/SAheart.csv')

y = np.array(data['chd'])
X = np.array(data.drop('chd',axis=1))

# Separate data into train/test
N = 100  # Number of repetitions

log_model = LogisticRegression(fit_intercept=True)
score = np.zeros((N,1))
for i in range(N):
    X_train, X_test, y_train, y_test =
     train_test_split(X, y, test_size=0.25)
    log_model.fit(X_train,y_train)
    score[i] = log_model.score(X_test, y_test)

print score.mean()
print score.std()
```

We obtain about 72% accuracy with a standard deviation of $\approx 4\%$.

Logistic regression with more than 2 classes

- Suppose now the response can take any of $\{1, \ldots, K\}$ values.
- Can still use logistic regression.
- We use the categorical distribution instead of the Bernoulli distribution.
- $P(Y = i|X = x) = p_i$, $0 \leq p_i \leq 1$, $\sum_{i=1}^K p_i = 1$.
- Each category has its own set of coefficients:

$$P(Y = i|X = x) = \frac{e^{x^T \beta^{(i)}}}{\sum_{i=1}^K e^{x^T \beta^{(i)}}}.$$

- Estimation can be done using maximum likelihood as for the binary case.

Example: handwritten digits

- Normalized handwritten digits, automatically scanned from envelopes by the U.S. Postal Service.
- Images here have been deslanted and size normalized, resulting in 16 x 16 grayscale images (Le Cun et al., 1990).
- Each line consists of the digit id (0-9) followed by the 256 grayscale values.
- There are 7291 training observations and 2007 test observations.
- The test set is notoriously “difficult”, and a 2.5% error rate is excellent.
- These data were kindly made available by the neural network group at AT&T research labs (thanks to Yann Le Cunn).

Exercise: Use logistic regression to predict the handwritten digits. Compute the prediction error of your model on the given test set.