

LOGISTIC REGRESSION

Concha Bielza, Pedro Larrañaga

Computational Intelligence Group
Departamento de Inteligencia Artificial
Universidad Politécnica de Madrid



Machine Learning

Outline

- 1 Logistic regression model
- 2 Maximum likelihood estimation of parameters
- 3 Conclusions

Outline

- 1 **Logistic regression model**
- 2 Maximum likelihood estimation of parameters
- 3 Conclusions

Motivation

Objectives

- 1 Determine the existence/absence of **relationship** between independent variables and a dependent variable
predictor response
- 2 Use the identified variables to **predict the probability** of the response taking each value, as a function of the predictor values
- 3 Use these probabilities to **classify** future observations

Approach

What

- Since '67, standard for regression with dichotomic data (Health Sciences)

- We have: $Y = C = 0, 1$

$$X_1, \dots, X_n$$

- N observations like

$$\mathcal{D} = \{(c^j, x_1^j, \dots, x_n^j) = (c^j, \mathbf{x}^j), j = 1, \dots, N\} \text{ with}$$

$c^j = 1$: observation j has the characteristic;

$c^j = 0$: it hasn't

- Dependent variable is

$$\pi^j = p(C = 1 | \mathbf{x}^j) = p(C = 1 | X_1 = x_1^j, \dots, X_n = x_n^j)$$

and since C is Bernoulli, its mean is $E(C | \mathbf{x}^j) = \pi^j$ $\pi^j * 1 + (1 - \pi^j) * 0$

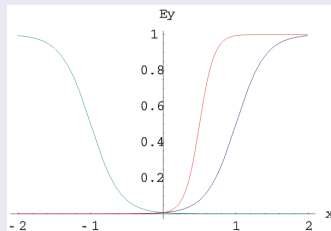
⇒ We look for a relationship between the response mean and the predictors

⇒ Scatterplots are not useful: no relation between y-axis and data Now, we only see two horizontal lines ($C=0, C=1$)

Intuitions

If $n = 1$, $C = 1$ = heart attack, X = cholesterol level, what relationship we expect between π and x ?

- $\pi \approx 1$ for large x values; $\pi \approx 0$ for small x values
- Non-linear for many values of X : for medium x 's, almost linear; asymptotic in extremes



- $\beta_1 = 5$
- $\beta_1 = 10$
- $\beta_1 = -5$

Beta_1 is more important than beta_0
Beta_1's sign determines whether the graph increases or decreases

Logistic function

$$\pi = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

Always between 0 and 1

⇒ satisfies $\pi \in [0, 1]$

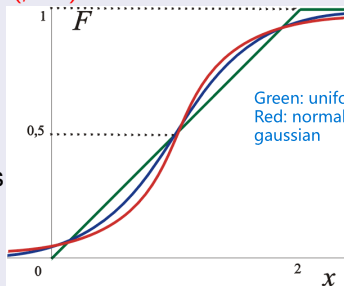
Intuitions

logic model -> logistic
probit model -> gaussi

In general

- In general, to guarantee $\pi \in [0, 1]$, we apply a nonlinear transformation: $\pi = F(\beta^t \mathbf{x})$

- F any distribution function
- $\beta = (\beta_1, \dots, \beta_n)$
vector of coefficients
- $\mathbf{x}^j = (x_1^j, \dots, x_n^j)$
data



Green: uniform distribution
Red: normal distribution / gaussian

Expressions: π and $1 - \pi$

Logistic model

$\forall j = 1, \dots, N$:

$$\pi^j = p(C = 1 | \mathbf{x}^j) = \frac{e^{\beta^t \mathbf{x}^j}}{1 + e^{\beta^t \mathbf{x}^j}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1^j + \dots + \beta_n x_n^j)}}$$

$$\Rightarrow 1 - \pi^j = p(C = 0 | \mathbf{x}^j) = \frac{1}{1 + e^{(\beta_0 + \beta_1 x_1^j + \dots + \beta_n x_n^j)}}$$

- $\beta_0, \beta_1, \dots, \beta_n$ are the parameters, to be estimated from data
- Decision boundary is **linear**:

$$p(C = 1 | \mathbf{x}^j) = p(C = 0 | \mathbf{x}^j) \iff p(C = 1 | \mathbf{x}^j) = 0.5$$

$$\iff \beta_0 + \beta_1 x_1^j + \dots + \beta_n x_n^j = 0$$

This expression defines a hyperplane (the decision boundary)

Expressions: Risk Ratio $RR(\mathbf{x}, \mathbf{x}')$

We can use any kind of variable
(continuous, discrete)

Example

- Variables: C Coronary Disease (1 yes, 0 no); X_1 Cholesterol (1 high, 0 low), X_2 Age, and X_3 Electrocardiogram res. (1 abnormal, 0 normal)
- Parameters ($N = 609$ obs): $\widehat{\beta}_0 = -3.911$ $\widehat{\beta}_1 = 0.652$ $\widehat{\beta}_2 = 0.029$
 $\widehat{\beta}_3 = 0.342$
- Compare the risk for two patterns: $\mathbf{x} = (1, 40, 0)$ and $\mathbf{x}' = (0, 40, 0)$:
 - $p(C = 1|\mathbf{x}) = p(C = 1|X_1 = 1, X_2 = 40, X_3 = 0) =$

$$= \frac{1}{1 + e^{-(-3.911 + 0.652(1) + 0.029(40) + 0.342(0))}} = 0.109$$
 - $p(C = 1|\mathbf{x}') = p(C = 1|X_1 = 0, X_2 = 40, X_3 = 0) =$

$$= \frac{1}{1 + e^{-(-3.911 + 0.652(0) + 0.029(40) + 0.342(0))}} = 0.060$$
- $RR(\mathbf{x}, \mathbf{x}') = \frac{p(C=1|\mathbf{x})}{p(C=1|\mathbf{x}')} = \frac{p(C=1|X_1=1, X_2=40, X_3=0)}{p(C=1|X_1=0, X_2=40, X_3=0)} = \frac{0.109}{0.060} = 1.82$
- For a person who is 40 years old and with normal electrocardiogram, the risk is multiplied by almost 2 when going from low Cholesterol level (0) to high (1)

Expressions: Odds and logit

Logistic model in *logit* form

- $$\text{Odds}(\mathbf{x}) = \frac{p(C = 1 | \mathbf{x})}{1 - p(C = 1 | \mathbf{x})} = e^{(\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n)}$$

Increasing 1 unit x_1 , $\frac{p(C=1|\mathbf{x})}{1-p(C=1|\mathbf{x})}$ multiplies by the factor e^{β_1} .

Not very interpretable

- logit* form:

$$\text{logit}(p(C = 1 | \mathbf{x})) = \ln \text{Odds}(\mathbf{x}) = \ln \left[\frac{p(C = 1 | \mathbf{x})}{1 - p(C = 1 | \mathbf{x})} \right] = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n$$

A **linear** model with this transformation, that represents in a logarithmic scale the difference between the probabilities of belonging to both classes. Interpretable in this scale

- Example: $\text{logit}(p(C = 1 | \mathbf{0})) = \ln \text{Odds}(\mathbf{0}) = \beta_0$

Interpreting parameters β_i

Proposition: *In a logistic regression model, coefficient β_i represents the logit change when the i -th variable X_i ($i = 1, \dots, n$) increases 1 unit*

Proof: Let \mathbf{x} and \mathbf{x}' be vectors such that $x_l = x'_l$ for all $l \neq i$ and $x'_i = x_i + 1$, then

$$\begin{aligned} \text{logit}(p(C = 1|\mathbf{x}')) - \text{logit}(p(C = 1|\mathbf{x})) = \\ \beta_0 + \sum_{l=1}^n \beta_l x'_l - \left(\beta_0 + \sum_{l=1}^n \beta_l x_l \right) = \beta_i x'_i - \beta_i x_i = \beta_i (x_i + 1 - x_i) = \beta_i \end{aligned}$$

- In the example: $\mathbf{x} = (1, 40, 0)$, $\mathbf{x}' = (0, 40, 0)$

$$\text{logit}(p(C = 1|\mathbf{x})) = \beta_0 + 1 \cdot \beta_1 + 40 \cdot \beta_2 + 0 \cdot \beta_3$$

$$\text{logit}(p(C = 1|\mathbf{x}')) = \beta_0 + 0 \cdot \beta_1 + 40 \cdot \beta_2 + 0 \cdot \beta_3$$

$$\Rightarrow \text{logit}(p(C = 1|\mathbf{x})) - \text{logit}(p(C = 1|\mathbf{x}')) = \beta_1$$

Multi-class logistic regression: $\Omega_C = \{1, \dots, R\}, R > 2$

- $C|\mathbf{x} \sim$ **categorical** distribution (rather than Bernoulli)
- Equation of the logit is now **a set of $R - 1$** logit transformations:

$$\ln \frac{p(C = 1|\mathbf{x})}{p(C = R|\mathbf{x})} = \beta_{10} + \beta_{11}x_1 + \dots + \beta_{1n}x_n \quad \text{Logistic between 1 and R}$$

$$\vdots$$

$$\ln \frac{p(C = R - 1|\mathbf{x})}{p(C = R|\mathbf{x})} = \beta_{(R-1)0} + \beta_{(R-1)1}x_1 + \dots + \beta_{(R-1)n}x_n \quad \text{Logistic between R-1 and R}$$

- Convention: using the last category R as the denominator (estimates do not vary under other choice). We get:

$$p(C = r|\mathbf{x}) = \frac{e^{\beta_{r0} + \beta_{r1}x_1 + \dots + \beta_{rn}x_n}}{1 + \sum_{l=1}^{R-1} e^{\beta_{l0} + \beta_{l1}x_1 + \dots + \beta_{ln}x_n}}, \quad r = 1, \dots, R - 1$$

$$p(C = R|\mathbf{x}) = \frac{1}{1 + \sum_{l=1}^{R-1} e^{\beta_{l0} + \beta_{l1}x_1 + \dots + \beta_{ln}x_n}}$$

which add up to 1. There are $(n + 1)(R - 1)$ parameters: $\{\beta_{10}, \dots, \beta_{(R-1)n}\}$

Feature subset selection

Multicollinearity among predictors

- Important to **remove** it, as done in linear regression
⇒ Unstable $\hat{\beta}_i$ (correlated, high std error)
- Detect it as usually Use the correlation matrix
- Remove correlated predictors

Outline

- 1 Logistic regression model
- 2 Maximum likelihood estimation of parameters**
- 3 Conclusions

Maximum likelihood estimates

(Conditional) likelihood function \mathcal{L}

- Probability function: $p(C = c^j | \mathbf{x}^j) = (\pi^j)^{c^j} (1 - \pi^j)^{1 - c^j}$, $c^j = 0, 1$
(each obs is a Bernoulli trial)

- $\mathcal{L}(\beta | \mathcal{D}) = \prod_{j=1}^N p(C = c^j | \mathbf{x}^j) = \prod_{j=1}^N (\pi^j)^{c^j} (1 - \pi^j)^{1 - c^j}$

- Conditional log-likelihood: $\ln \mathcal{L}(\beta | \mathcal{D}) = \sum_{j=1}^N \ln p(C = c^j | \mathbf{x}^j)$

$$= \sum_{j=1}^N \left[c^j \ln \pi^j + (1 - c^j) \ln(1 - \pi^j) \right]$$

$$= \sum_{j=1}^N c^j \ln \frac{\pi^j}{1 - \pi^j} + \sum_{j=1}^N \ln(1 - \pi^j)$$

$$= \sum_{j=1}^N c^j \left(\beta_0 + \beta_1 x_1^j + \cdots + \beta_n x_n^j \right) - \sum_{j=1}^N \ln \left(1 + e^{(\beta_0 + \beta_1 x_1^j + \cdots + \beta_n x_n^j)} \right)$$

Maximum likelihood estimates

MLE $\hat{\beta}_i$ for β_i

- If the derivative is equal to zero: *–likelihood equations–*

$$\frac{\partial \ln \mathcal{L}(\beta)}{\partial \beta_0} = \sum_{j=1}^N c^j - \sum_{j=1}^N \frac{e^{(\beta_0 + \beta_1 x_1^j + \dots + \beta_n x_n^j)}}{1 + e^{(\beta_0 + \beta_1 x_1^j + \dots + \beta_n x_n^j)}} = 0$$

$$\frac{\partial \ln \mathcal{L}(\beta)}{\partial \beta_1} = \sum_{j=1}^N c^j x_1^j - \sum_{j=1}^N x_1^j \frac{e^{(\beta_0 + \beta_1 x_1^j + \dots + \beta_n x_n^j)}}{1 + e^{(\beta_0 + \beta_1 x_1^j + \dots + \beta_n x_n^j)}} = 0$$

$$\vdots$$

$$\frac{\partial \ln \mathcal{L}(\beta)}{\partial \beta_n} = \sum_{j=1}^N c^j x_n^j - \sum_{j=1}^N x_n^j \frac{e^{(\beta_0 + \beta_1 x_1^j + \dots + \beta_n x_n^j)}}{1 + e^{(\beta_0 + \beta_1 x_1^j + \dots + \beta_n x_n^j)}} = 0$$

Non-linear in β

Maximum likelihood estimates

MLE $\hat{\beta}_i$ for β_i

- It is impossible to have a closed formula (analytic solution) for MLE
- Newton-Raphson's** numeric algorithm is traditionally used, with an updating formula given by

$$\hat{\beta}^{\text{new}} = \hat{\beta}^{\text{old}} + (\mathbf{Z}^t \mathbf{W}^{\text{old}} \mathbf{Z})^{-1} \mathbf{Z}^t (\mathbf{c} - \hat{\pi}^{\text{old}})$$

\mathbf{c} is N -vector of response values c^j , $j = 1, \dots, N$

\mathbf{X} is $N \times n$ -matrix with rows \mathbf{x}^j

\mathbf{Z} is the matrix $[\mathbf{u} | \mathbf{X}]$, with \mathbf{u} the N -vector of ones \mathbf{u} is for β_0

$\hat{\pi}^{\text{old}}$ is N -vector of estimated values at that iteration, i.e. its j th-component is

$$(\hat{\pi}^j)^{\text{old}} = [1 + e^{-(\hat{\beta}_0^{\text{old}} + \hat{\beta}_1^{\text{old}} x_1^j + \dots + \hat{\beta}_n^{\text{old}} x_n^j)}]^{-1}$$

\mathbf{W}^{old} is a diagonal matrix with elements $(\hat{\pi}^j)^{\text{old}}(1 - (\hat{\pi}^j)^{\text{old}})$

Initialize e.g. with $\hat{\beta} = (0, \dots, 0)$ These values will change, because if they don't, we will always get $\pi = 0.5$ which is not interesting

- ...until convergence

Classifying

Steps

1. Fix a cutoff value $\hat{\pi}^*$ for $\hat{\pi}$
2. Assign $\hat{c}^j = 1$ if $\hat{\pi}^j \geq \hat{\pi}^*$. Otherwise, $\hat{c}^j = 0$ (predicted class)
3. Build the confusion matrix:

	$\hat{c} = 1$	$\hat{c} = 0$
$c = 1$	N_1	N_2
$c = 0$	N_3	N_4

$$N = N_1 + N_2 + N_3 + N_4$$

Assess the model utility:

$$\begin{aligned} \% \text{ correctly classified} &= 100 (N_1 + N_4)/N \\ \text{sensitivity} &= 100 N_1/(N_1 + N_2) \\ \text{specificity} &= 100 N_4/(N_3 + N_4) \end{aligned}$$

Outline

- 1 Logistic regression model
- 2 Maximum likelihood estimation of parameters
- 3 Conclusions**

Software

Logistic regression with WEKA

Classifier \Rightarrow Functions

Logistic

- Binary case $\rightarrow e^{\beta_i}$ is the odds ratio in Weka (> 1 for $\beta_i > 0$, and < 1 for $\beta_i < 0$). Increasing X_i in 1 unit (the remaining variables do not change), the ratio $p(C = 1|\mathbf{x})/p(C = 0|\mathbf{x})$ multiplies by e^{β_i} .
- Multi-class case $\rightarrow e^{\beta_{1i}}$: Increasing X_i in 1 unit (the remaining variables do not change), the ratio $p(C = 1|\mathbf{x})/p(C = R|\mathbf{x})$ multiplies by $e^{\beta_{1i}}$. If $e^{\beta_{1i}} > 1$ ($\beta_{1i} > 0$) then $C = 1$ becomes more likely than $C = R$ for each increment in X_i

Conclusions

Statistical paradigm

- **Discriminative** model: maximize conditional probability
- Assign to each instance the **posterior probability** of belonging to each class
- Interpretation of parameters
- Estimation of parameters by **maximum likelihood**.
Approximate them via iterative numerical methods

$$\operatorname{argmax}_C p(C|X) = \operatorname{argmax}_C p(C,X)/p(X) = \operatorname{argmax}_C p(C,X)$$

The previous expressions are used by generative classifiers. There are two types of generative classifiers:

- Discriminant analysis (like logistic regression)
- Bayesian classifiers

Bibliography

Texts

- Bielza, C., Larrañaga, P. (2021) *Data-Driven Computational Neuroscience. Machine Learning and Statistical Models*, Cambridge University Press [Chap. 8]
- Hosmer, D.W., Lemeshow, S. (2000) *Applied Logistic Regression*, 2nd ed., Wiley Interscience
- Kleinbaum, D.G. (1994) *Logistic Regression*, Springer
- Ryan, T.P. (1997) *Modern Regression Methods*, Wiley [Chap. 9]
- Sharma, S. (1996) *Applied Multivariate Techniques*, Wiley [Chap. 10]