Classificação

Machine Learning

Prof. Neylson Crepalde

Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani

Utilizamos as análises de regressão quando nossa variável resposta é quantitativa. Entretanto, em alguns casos, podemos ter variáveis resposta que são **qualitativas**, ou **categóricas**. Nesses casos, utilizaremos os métodos e algoritmos de **classificação**.

Alguns problemas comuns de classificação:

- Uma pessoa chega na sala de emergência de um hospital com uma série de sintomas que podem se encaixar em 3 diagnósticos. Qual dos 3 o indivíduo tem?
- Um serviço de banco online deve ser capaz de determinar se uma transação efetuada em seu site é ou não fraudulenta com base no endereço de IP do usuário, histórico de transações, etc.
- Com base em dados de sequência de DNA para um número de pacientes com e sem uma determinada doença, um bioinformata desejaria conhecer quais mutações de DNA são deletérias (causam a doença) e quais não são.

De maneira mais substantiva, podemos interpretar coeficientes positivos como um aumento nas probabilidades de sucesso de Y quando X também aumenta (**embora essa relação não seja linear!!!**). Para termos um coeficiente mais interpretável, duas transformações são possível:

- e^{β_1} nos dá a chance relativa (*odds*);
- (e^{β_1} -1) X 100 nos dá o percentual de mudança na probabilidade (aumento ou diminuição).

A regressão logística

No caso do banco de dados *Default* que utilizaremos para prever o não pagamento do cartão de crédito. Neste caso, nossa variável resposta é uma variável binária que assume apenas dois valores, SIM (1) e NÃO (0). O modelo logístico, portanto, difere essencialmente do modelo linear no sentido de que aqui prevemos a probabilidade de sucesso da variável resposta dadas as covariáveis.

O modelo pode ser definido da seguinte maneira:

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

Transformando, temos:

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0+\beta_1 X}$$

A equação

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}$$

representa as *chances (odds)* de sucesso da variável em questão. Por exemplo, podemos dizer que 1 em 5 pessoas não vão pagar o cartão de crédito se obtivermos odds = 1/4. Se a probabilidade p(X) = 0.2, então as chances serão 0.2/(1-0.2) = 1/4.

A equação pode ser transformada para obtermos o logaritmo natural das chances, ou *logit*:

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

A parte esquerda é chamada de *logit* ou *log das chances (log odds)*. Podemos perceber que o *logit* é linear em relação a X. Entretanto, a interpretação do coeficiente estimado não corresponde, como na regressão linear, a um aumento linear na probabilidade.

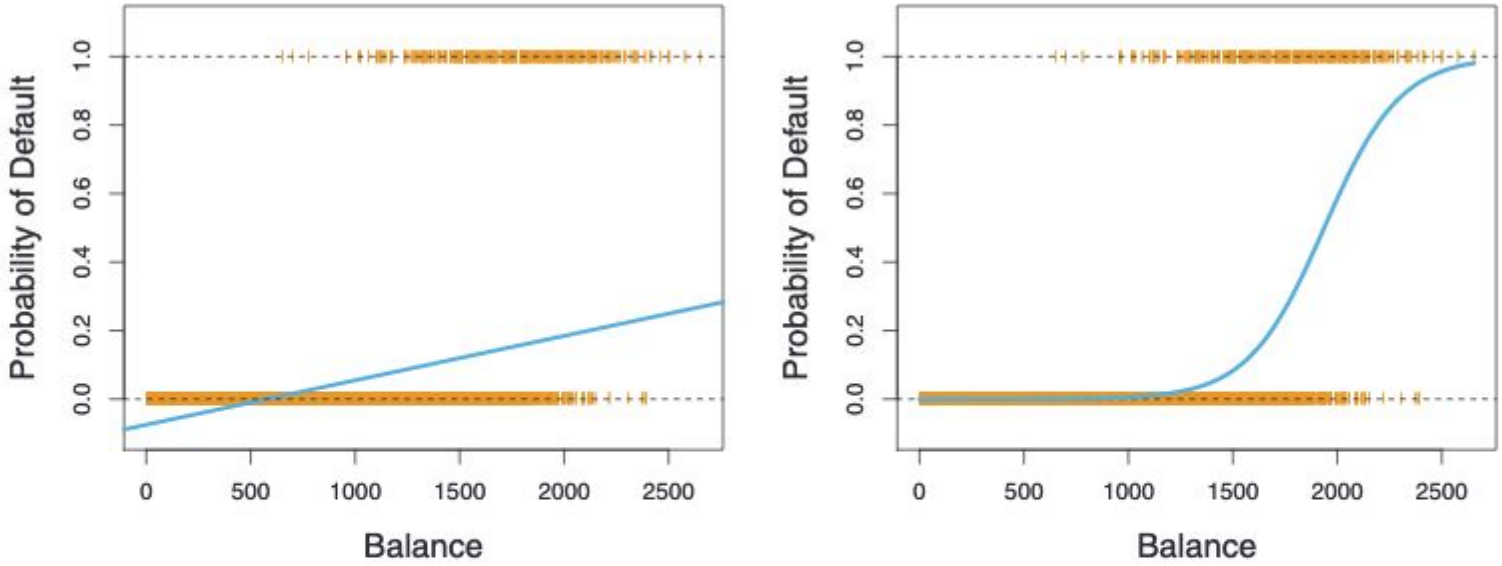

FIGURE 4.2. *Classification using the* `Default` *data.* Left: *Estimated probability of* `default` *using linear regression. Some estimated probabilities are negative! The orange ticks indicate the 0/1 values coded for* `default` *(*`No` *or* `Yes`*).* Right: *Predicted probabilities of* `default` *using logistic regression. All probabilities lie between* 0 *and* 1.

A estimação dos coeficientes geralmente é feita por **máxima verossimilhança** (*maximum likelihood*). Intuição básica desse método de estimação pode ser descrita da seguinte maneira:

Procuramos estimar beta0 e beta1 de modo que as probabilidades estimadas de Y dado X correspondam tão próximas quanto possível das classificações observadas. Dito de outra forma, a estimação é realizada tentando obter probabilidades mais próximas de 1 para indivíduos que tenham sucesso para Y e probabilidades mais próximas de 0 para indivíduos que tenham fracasso para Y. Isto pode ser formalizado na seguinte equação:

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'}))$$

Exercício!