Modelos de Classificação
Logistic Regression

Prof. Gustavo Willam Pereira

INSTITUTO FEDERAL
Sudeste de Minas Gerais

Classificação Supervisionada

- Conforme já vimos em aulas anteriores, a aprendizagem supervisionada é realizada quando temos disponíveis dados de entrada e conhecemos, para cada entrada, o resultado da saída.
- A regressão, que é um tipo de aprendizagem supervisionada, utilizada para dados contínuos.
- Agora iremos descrever alguns algoritmos classificação supervisionada.
- Na classificação supervisionada os dados de entrada são utilizados para prever classes, que são variáveis categóricas.
- A principal diferença é que na classificação supervisionada temos que decidir sobre determinada classe, como por exemplo, se um determinado indivíduo apresenta ou não uma determinada doença.

Classificação Supervisionada

- Nesse caso, temos que nos preocupar com os erros e suas consequências.
- Se o algoritmo der como resposta que uma planta tem uma doença, então iremos iniciar um tratamento, no entanto, se o algoritmo estiver errado? Qual risco corremos?
- Esse é um tipo de erro conhecido como falso positivo.
- Em caso contrário, se nosso algoritmo der como resposta que não existe doença na planta, qual o risco corremos se houver um erro? Esse é um tipo de erro conhecido como falso negativo.
- Dessa forma, em classificação, a análise da acurácia do modelo é diferente da regressão.

Classificação Supervisionada

- Para classificação vamos tratar dos algoritmos:
  - Logistic Regression
  - K-Nearest Neighbors(K-NN)
  - Naive Bayes,
  - Support Vector Machine (SVM),
  - Decision Tree
  - Random Forest.

Logistic Regression

- Para exemplificar de forma intuitiva como funciona o algoritmo Logistic Regression, vamos imaginar que temos um problema de classificação de doença em pessoas.
- Então, temos os seguintes rótulos no banco de dados. Se a pessoa está doente, o rótulo será 1 (sim), caso contrário, 0 (não).
- Para cada indivíduo, medimos um determinado índice (um índice de doença), que é nossa variável (x1).
- Veja na Figura 1 como seria essa regressão. Então se o valor estimado de y fosse próximo ou maior que 1 então o resultado da classificação seria 1 (sim). Caso o resultado de y fosse menor ou próximo de 0, então poderíamos decidir que não tem a doença 0 (não).

Logistic Regression

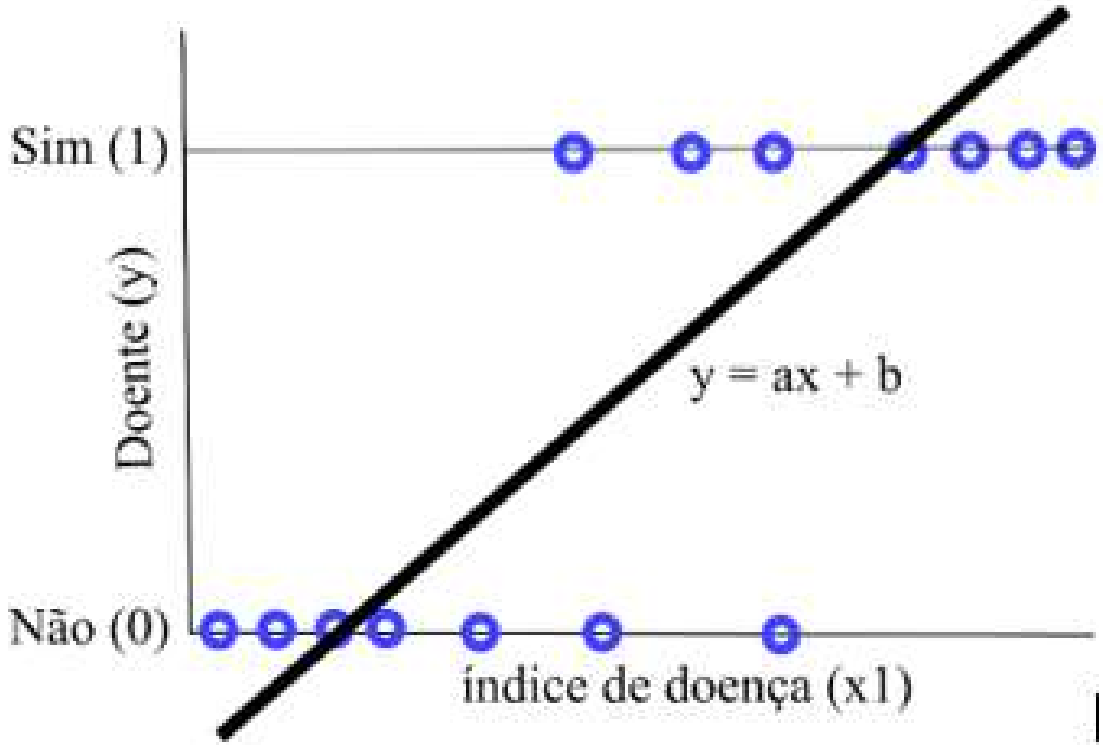


Figura 1 – Problema de classificação de doença (1 e 0) utilizando regressão linear simples.

Logistic Regression

- A regressão da Figura 1 é descrita pela seguinte equação.

$$y = b_0 + b_1 X$$

- A seguir uma equação sigmoide ou logística que assume valores entre 0 e 1.

$$p = \frac{1}{1 + e^{-y}}$$

- Nessa equação sigmoide poderíamos substituir o y da função de regressão. Então a solução seria a seguinte equação.

$$ln\left(\frac{p}{1-p}\right) = b_0 + b_1 X$$

Logistic Regression

- Nessa equação p é a estimativa da probabilidade. Se ajustássemos a equação acima aos dados do problema anterior, obteríamos a função da Figura 2.

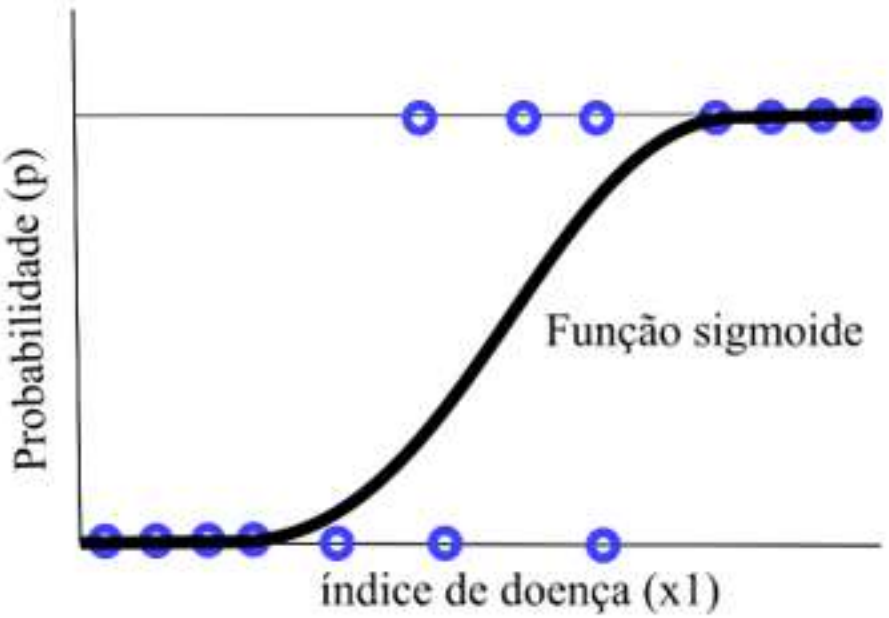


Figura 2 – Transformação da função de regressão em uma função sigmoide de probabilidade

Logistic Regression

- Então, para cada valor da variável x1 teríamos uma probabilidade associada.
- Na Figura 3 é apresentado um exemplo que para x1 = 0,40 a probabilidade seria de 5% de ter a doença,
- Se x1 = 0,65 a probabilidade seria de 40% de ter a doença,
- Se x1 = 0,85 a probabilidade seria de 80% de ter a doença.

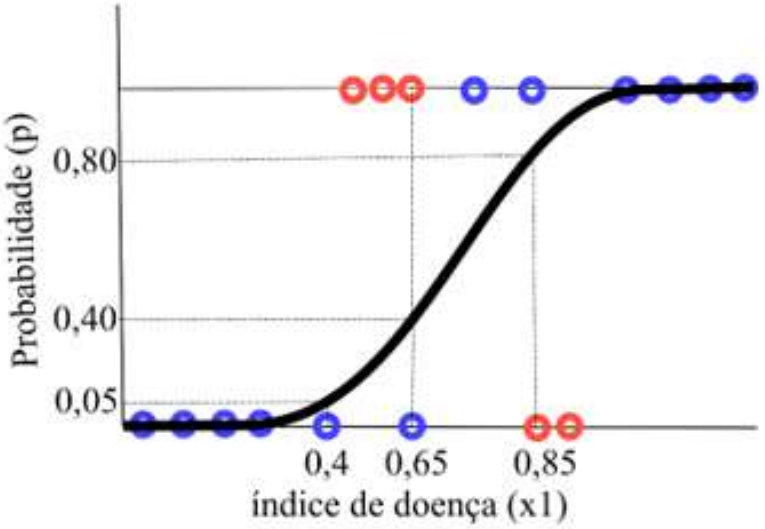


Figura 3 – Valores da variável X1 e suas probabilidades.

INSTITUTO FEDERAL
Sudeste de Minas Gerais

Logistic Regression

- Com base na probabilidade de ocorrência, se decidíssemos que se a probabilidade for menor que 50%, então a pessoa não tem a doença, e se a probabilidade for maior que 50%, então a pessoa tem a doença.
- Veja na Figura 3, que com base nesse critério de 50%, 3 amostras doentes seriam erroneamente classificadas como não doentes (Falso negativo).
- Isso ocorre pois se visualmente, usando a Figura 3, projetássemos os pontos na função logística os pontos em vermelho teríamos valores de probabilidade menores que 50%.
- Por outro lado, 2 amostras (em vermelho) não doentes seriam classificadas como doentes (Falso positivo). Isso ocorre pois os valores de probabilidade seriam maiores que 50%.

Logistic Regression

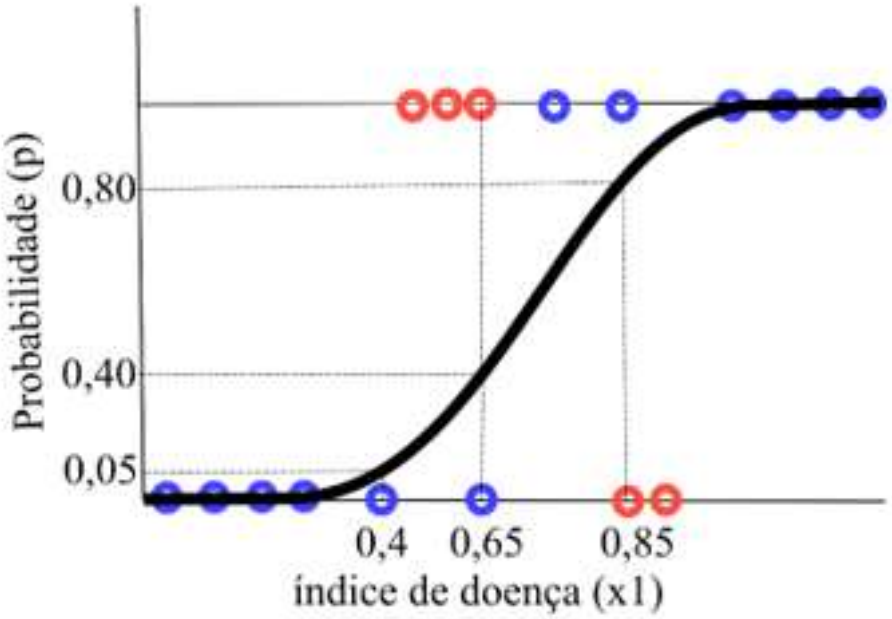


Figura 3 – Valores da variável X1 e suas probabilidades.

- Como exercício, imagine o que ocorreria se modificássemos o limite de 50% para 20%. Veja que nesse caso todas as amostras que apresentasse probabilidade acima de 20% seriam classificadas como doentes, caso contrário, não doentes. O que ocorreria com os erros do tipo falso positivo ou falso negativo?

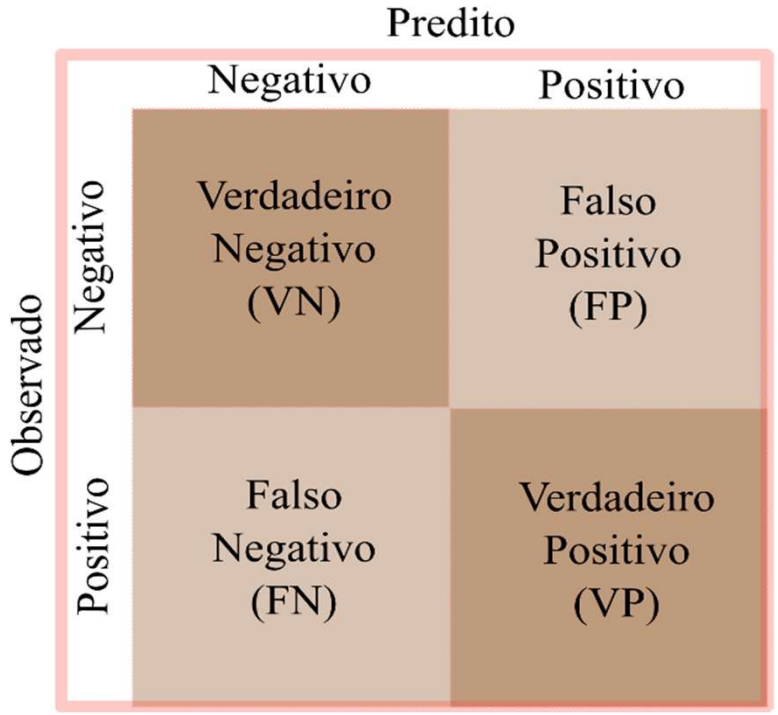
Logistic Regression

- Agora vamos fazer um exemplo de classificação supervisionada utilizando Logistic Regression no Python.
- Para teste vamos utilizar o banco de dados Framingham, disponível em: https://www.kaggle.com/navink25/framingham.
    - O conjunto de dados está disponível publicamente no site Kaggle e é de um estudo cardiovascular em andamento em moradores da cidade de Framingham, Massachusetts. O objetivo da classificação é prever se o paciente tem risco de contrair doença coronariana futura (DCF). O conjunto de dados fornece as informações dos pacientes. Inclui mais de 4.000 registros e 15 atributos.

Métricas para Classificação

Matriz de confusão

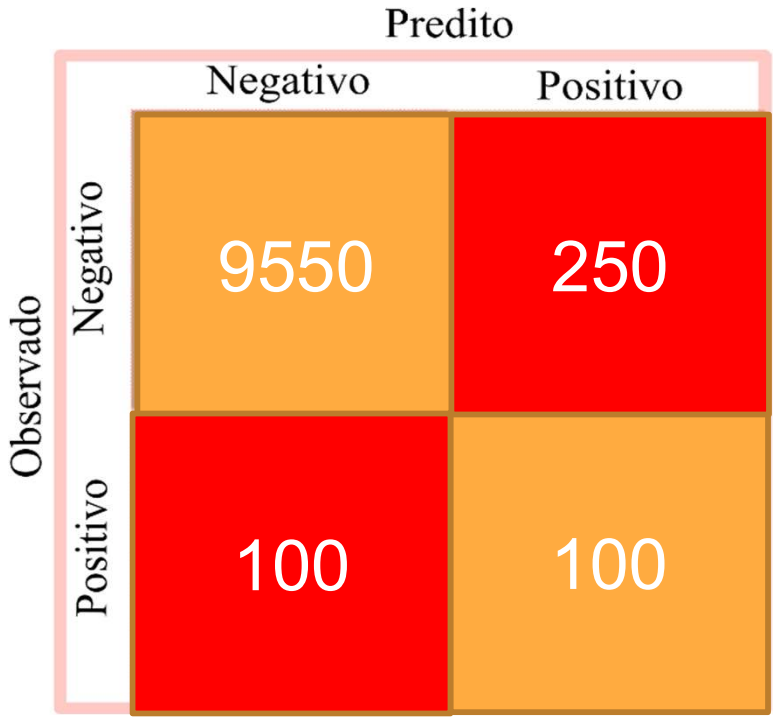
$$acurácia\ global = \frac{VN + VP}{VN + FP + FN + VP}$$



INSTITUTO FEDERAL
Sudeste de Minas Gerais

Métricas para Classificação

Matriz de confusão



$$acurácia\ global = \frac{VN + VP}{VN + FP + FN + VP}$$

$$acurácia\ global = \frac{9550 + 100}{10000} = 0{,}965$$

INSTITUTO FEDERAL
Sudeste de Minas Gerais

INSTITUTO FEDERAL
Sudeste de Minas Gerais