Regressão Logística

Especialização em Ciência de Dados Aplicada

Paulo Henrique Ribeiro Gabriel

Faculdade de Computação
Universidade Federal de Uberlândia

2024

CLASSIFICAÇÃO

- A classificação é uma das áreas mais importantes do aprendizado de máquina

CLASSIFICAÇÃO

- A classificação é uma das áreas mais importantes do aprendizado de máquina
- Diversos problemas práticos recaem nessa área

CLASSIFICAÇÃO

- A classificação é uma das áreas mais importantes do aprendizado de máquina
- Diversos problemas práticos recaem nessa área
- A regressão logística é uma das técnicas mais básicas de classificação

CLASSIFICAÇÃO

O QUE É CLASSIFICAÇÃO?

- Algoritmos de aprendizado de máquina supervisionado definem modelos de relacionamentos entre dados

Classificação

O que é classificação?

- Algoritmos de aprendizado de máquina supervisionado definem modelos de relacionamentos entre dados
- A classificação visa prever a qual classe ou categoria pertence uma entidade

- Algoritmos de aprendizado de máquina supervisionado definem modelos de relacionamentos entre dados
- A classificação visa prever a qual classe ou categoria pertence uma entidade
  - Com base nas características dessa entidade

Os atributos ou variáveis podem assumir uma das duas formas:

Classificação

O que é classificação?

Os atributos ou variáveis podem assumir uma das duas formas:

1. Variáveis independentes
   - Também chamadas de entradas ou preditores
   - Não dependem de outras características de interesse

Classificação

O que é classificação?

Os atributos ou variáveis podem assumir uma das duas formas:

1. Variáveis independentes
   - Também chamadas de entradas ou preditores
   - Não dependem de outras características de interesse

2. Variáveis dependentes
   - Também chamadas de saídas ou respostas
   - Dependem das variáveis independentes

Classificação
O que é classificação?

Exemplo

- Analisar os funcionários de alguma empresa

Classificação

O que é classificação?

Exemplo

- Analisar os funcionários de alguma empresa
- Tentar estabelecer uma dependência de variáveis:
    - Nível de escolaridade
    - Número de anos no cargo atual
    - Idade
    - Salário
    - Chances de ser promovido

Classificação

O que é classificação?

Exemplo

- Analisar os funcionários de alguma empresa
- Tentar estabelecer uma dependência de variáveis:
  - Nível de escolaridade
  - Número de anos no cargo atual
  - Idade
  - Salário
  - Chances de ser promovido
- O conjunto de dados relativos a um único funcionário é uma observação

Classificação

O que é classificação?

Exemplo

- Analisar os funcionários de alguma empresa
- Tentar estabelecer uma dependência de variáveis:
  - Nível de escolaridade
  - Número de anos no cargo atual
  - Idade
  - Salário
  - Chances de ser promovido
- O conjunto de dados relativos a um único funcionário é uma observação
- Podemos presumir que nível de escolaridade, tempo no cargo atual e idade são mutuamente independentes (entradas)

Classificação

O que é classificação?

Exemplo

- Analisar os funcionários de alguma empresa
- Tentar estabelecer uma dependência de variáveis:
  - Nível de escolaridade
  - Número de anos no cargo atual
  - Idade
  - Salário
  - Chances de ser promovido
- O conjunto de dados relativos a um único funcionário é uma observação
- Podemos presumir que nível de escolaridade, tempo no cargo atual e idade são mutuamente independentes (entradas)
- Salário e chances de promoção podem ser os resultados que dependem das entradas

Observação

- Algoritmos de aprendizado de máquina supervisionado analisam uma série de observações e tentam expressar matematicamente a dependência entre as entradas e as saídas

Observação

- Algoritmos de aprendizado de máquina supervisionado analisam uma série de observações e tentam expressar matematicamente a dependência entre as entradas e as saídas
- Essas representações matemáticas de dependências são os modelos

- A natureza das variáveis dependentes diferencia problemas de regressão e classificação

Classificação

O que é classificação?

- A natureza das variáveis dependentes diferencia problemas de regressão e classificação
- Problemas de regressão têm resultados contínuos e geralmente ilimitados

Classificação

O que é classificação?

- A natureza das variáveis dependentes diferencia problemas de regressão e classificação
- Problemas de regressão têm resultados contínuos e geralmente ilimitados
  - Um exemplo é quando estimamos o salário em função da experiência e do nível de escolaridade

Classificação

O que é classificação?

- A natureza das variáveis dependentes diferencia problemas de regressão e classificação
- Problemas de regressão têm resultados contínuos e geralmente ilimitados
  - Um exemplo é quando estimamos o salário em função da experiência e do nível de escolaridade
- Problemas de classificação têm saídas discretas e finitas chamadas classes ou categorias

Classificação

O que é classificação?

- A natureza das variáveis dependentes diferencia problemas de regressão e classificação
- Problemas de regressão têm resultados contínuos e geralmente ilimitados
  - Um exemplo é quando estimamos o salário em função da experiência e do nível de escolaridade
- Problemas de classificação têm saídas discretas e finitas chamadas classes ou categorias
  - Por exemplo, prever se um funcionário será promovido ou não (verdadeiro ou falso) é um problema de classificação

Classificação

O que é classificação?

Existem dois tipos principais de problemas de classificação:

Existem dois tipos principais de problemas de classificação:

1. Classificação binária ou binomial:
   exatamente duas classes para escolher (geralmente verdadeiro e falso, 0 e 1, ou positivo e negativo)

Existem dois tipos principais de problemas de classificação:

1. Classificação binária ou binomial:
   exatamente duas classes para escolher (geralmente verdadeiro e falso, 0 e 1, ou positivo e negativo)

2. Classificação multi-classe ou multinomial:
   três ou mais classes de resultados para escolher

Regressão Logística

- A regressão logística é uma técnica de classificação fundamental

Regressão Logística

- A regressão logística é uma técnica de classificação fundamental
- Pertence ao grupo dos classificadores lineares e semelhante à regressão polinomial e linear

Regressão Logística

- A regressão logística é uma técnica de classificação fundamental
- Pertence ao grupo dos classificadores lineares e semelhante à regressão polinomial e linear
- A regressão logística é rápida e é conveniente para interpretarmos os resultados

Regressão Logística

- A regressão logística é uma técnica de classificação fundamental
- Pertence ao grupo dos classificadores lineares e semelhante à regressão polinomial e linear
- A regressão logística é rápida e é conveniente para interpretarmos os resultados
- Essencialmente, é um método para classificação binária

Regressão Logística

- A regressão logística é uma técnica de classificação fundamental
- Pertence ao grupo dos classificadores lineares e semelhante à regressão polinomial e linear
- A regressão logística é rápida e é conveniente para interpretarmos os resultados
- Essencialmente, é um método para classificação binária
  - Mas pode ser aplicada a problemas multi-classes

Vamos, inicialmente, compreender dois conceitos fundamentais:

Regressão Logística

Fundamentação matemática

Vamos, inicialmente, compreender dois conceitos fundamentais:

1. Função sigmoide
2. Função logaritmo natural

- A função sigmoide ($\sigma(x)$) tem valores de saída próximos de 0 ou 1 na maior parte de seu domínio

Regressão Logística

Fundamentação matemática

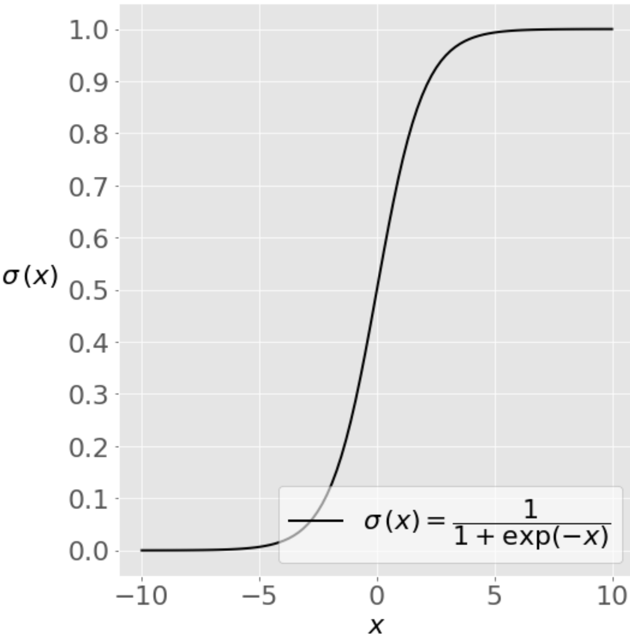
- A função sigmoide ($\sigma(x)$) tem valores de saída próximos de 0 ou 1 na maior parte de seu domínio
- Este fato a torna adequada para aplicação em métodos de classificação

Regressão Logística

Fundamentação matemática

- A função sigmoide ($\sigma(x)$) tem valores de saída próximos de 0 ou 1 na maior parte de seu domínio

- Este fato a torna adequada para aplicação em métodos de classificação



- Na função logaritmo , à medida que x tende para zero, o valor de $\ln(x)$ tende a menos infinito $(-\infty)$

- Na função logaritmo , à medida que x tende para zero, o valor de $\ln(x)$ tende a menos infinito $(-\infty)$

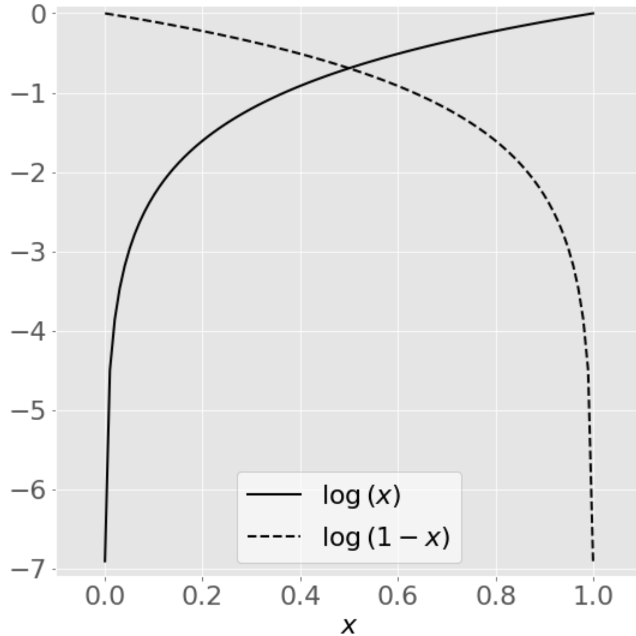
- Quando $x = 1$, $\ln(x) = 0$

- Na função logaritmo , à medida que x tende para zero, o valor de $\ln(x)$ tende a menos infinito $(-\infty)$
- Quando $x = 1$, $\ln(x) = 0$
- O oposto ocorre para $\ln(1 - x)$

Regressão Logística

Fundamentação matemática

- Na função logaritmo, à medida que x tende para zero, o valor de $\ln(x)$ tende a menos infinito $(-\infty)$
- Quando $x = 1$, $\ln(x) = 0$
- O oposto ocorre para $\ln(1 - x)$



Regressão Logística

Formulação do problema

- Vamos focar, aqui, no caso mais comum de regressão logística: a classificação binária

- Vamos focar, aqui, no caso mais comum de regressão logística: a classificação binária
- Se houver apenas uma variável de entrada, ela geralmente será denotada por x

Regressão Logística

Formulação do problema

- Vamos focar, aqui, no caso mais comum de regressão logística: a classificação binária
- Se houver apenas uma variável de entrada, ela geralmente será denotada por x
- Para mais de uma entrada, normalmente utilizamos a notação vetorial

$$\mathbf{x} = \{x_1, \ldots, x_r\}$$

onde r é o número de preditores (ou atributos) independentes

Regressão Logística

Formulação do problema

- Vamos focar, aqui, no caso mais comum de regressão logística: a classificação binária
- Se houver apenas uma variável de entrada, ela geralmente será denotada por x
- Para mais de uma entrada, normalmente utilizamos a notação vetorial

$$\mathbf{x} = \{x_1, \ldots, x_r\}$$

  onde r é o número de preditores (ou atributos) independentes
- A variável de saída é denotada por y e assume os valores 0 ou 1

Regressão Logística

Formulação do problema

- Começamos com os valores conhecidos dos preditores $\mathbf{x}_i$ e a resposta real correspondente (ou saída) y_i para cada observação $i = 1, \ldots, n$

Regressão Logística

Formulação do problema

- Começamos com os valores conhecidos dos preditores $\mathbf{x}_i$ e a resposta real correspondente (ou saída) y_i para cada observação $i = 1, \ldots, n$

- Nosso objetivo é encontrar a função de regressão logística $p(\mathbf{x})$ tal que as respostas previstas $p(\mathbf{x}_i)$ sejam o mais próximas possível da resposta real y_i para cada observação $i = 1, \ldots, n$

Regressão Logística

Formulação do problema

- Começamos com os valores conhecidos dos preditores $\mathbf{x}_i$ e a resposta real correspondente (ou saída) y_i para cada observação $i = 1, \ldots, n$
- Nosso objetivo é encontrar a função de regressão logística $p(\mathbf{x})$ tal que as respostas previstas $p(\mathbf{x}_i)$ sejam o mais próximas possível da resposta real y_i para cada observação $i = 1, \ldots, n$
- Lembrando que a resposta real pode ser apenas 0 ou 1 em problemas de classificação binária!

- Isto significa que cada $p(\mathbf{x}_i)$ deve estar próximo de 0 ou 1

Regressão Logística

Formulação do problema

- Isto significa que cada $p(\mathbf{x}_i)$ deve estar próximo de 0 ou 1
  - Por isso que é conveniente utilizar a função sigmoide!

Regressão Logística

Formulação do problema

- Isto significa que cada $p(\mathbf{x}_i)$ deve estar próximo de 0 ou 1
  - Por isso que é conveniente utilizar a função sigmoide!
- Depois de ter a função de regressão logística $p(\mathbf{x})$, podemos usá-la para prever os resultados para entradas novas e não vistas

- Isto significa que cada $p(\mathbf{x}_i)$ deve estar próximo de 0 ou 1
  - Por isso que é conveniente utilizar a função sigmoide!
- Depois de ter a função de regressão logística $p(\mathbf{x})$, podemos usá-la para prever os resultados para entradas novas e não vistas
  - Supondo, obviamente, que a dependência matemática subjacente permanece inalterada

Regressão Logística

Metodologia

- A regressão logística é um classificador linear

Regressão Logística

Metodologia

- A regressão logística é um classificador linear
- Portanto, utilizaremos uma função linear

$$f(\mathbf{x}) = b_0 + b_1 x_1 + \ldots + b_r x_r$$

também chamada de logit

Regressão Logística
Metodologia

- A regressão logística é um classificador linear
- Portanto, utilizaremos uma função linear

  $$f(\mathbf{x}) = b_0 + b_1 x_1 + \ldots + b_r x_r$$

  também chamada de logit
- As variáveis $b_0, b_1, \ldots, b_r$ são os estimadores dos coeficientes de regressão

Regressão Logística
Metodologia

- A regressão logística é um classificador linear
- Portanto, utilizaremos uma função linear

$$f(\mathbf{x}) = b_0 + b_1 x_1 + \ldots + b_r x_r$$

  também chamada de logit
- As variáveis $b_0, b_1, \ldots, b_r$ são os estimadores dos coeficientes de regressão
  - Também chamados de pesos previstos ou apenas coeficientes

Regressão Logística

Metodologia

- A função de regressão logística $p(\mathbf{x})$ é a função sigmoide de $f(\mathbf{x})$:

$$p(\mathbf{x}) = \frac{1}{1 + \exp(-f(\mathbf{x}))}$$

Regressão Logística

Metodologia

- A função de regressão logística $p(\mathbf{x})$ é a função sigmoide de $f(\mathbf{x})$:

$$p(\mathbf{x}) = \frac{1}{1 + \exp(-f(\mathbf{x}))}$$

- A função $p(\mathbf{x})$ é frequentemente interpretada como a probabilidade prevista de que a saída para um determinado $\mathbf{x}$ ser igual a 1

- A função de regressão logística $p(\mathbf{x})$ é a função sigmoide de $f(\mathbf{x})$:

$$p(\mathbf{x}) = \frac{1}{1 + \exp(-f(\mathbf{x}))}$$

- A função $p(\mathbf{x})$ é frequentemente interpretada como a probabilidade prevista de que a saída para um determinado $\mathbf{x}$ ser igual a 1

- Portanto, $1 - p(\mathbf{x})$ é a probabilidade de que a saída seja 0

A regressão logística determina os melhores pesos previstos

$$b_0, b_1, \ldots, b_r$$

de modo que a função $p(\mathbf{x})$ seja o mais próximo possível de todas as respostas reais

$$y_i,\ i = 1, \ldots, n,$$

onde n é o número de observações

- O processo de cálculo dos melhores pesos usando as observações disponíveis é chamado de treinamento ou ajuste de modelo

Regressão Logística

Metodologia

- O processo de cálculo dos melhores pesos usando as observações disponíveis é chamado de treinamento ou ajuste de modelo
- Para obter os melhores pesos, geralmente buscamos maximizar a função *log-verossimilhança* (*log-likelihood function*, LLF) para todas as observações $i = 1, \ldots, n$

Regressão Logística

Metodologia

- O processo de cálculo dos melhores pesos usando as observações disponíveis é chamado de treinamento ou ajuste de modelo
- Para obter os melhores pesos, geralmente buscamos maximizar a função *log-verossimilhança* (*log-likelihood function*, LLF) para todas as observações $i = 1, \ldots, n$
- Este método é chamado de estimativa de máxima verossimilhança e é representado pela equação
$$LLF = \sum_i y_i \ln(p(\mathbf{x})) + (1 - y_i) \ln(1 - p(\mathbf{x}_i))$$

Regressão Logística

Metodologia

$$LLF = \sum_i y_i \ln(p(\mathbf{x})) + (1 - y_i) \ln(1 - p(\mathbf{x}_i))$$

Regressão Logística

Metodologia

$$LLF = \sum_i y_i \ln(p(\mathbf{x})) + (1 - y_i) \ln(1 - p(\mathbf{x}_i))$$

- Quando $y_i = 0$, o LLF da observação correspondente é igual a $\ln(1 - p(\mathbf{x}))$

Regressão Logística

Metodologia

$$LLF = \sum_i y_i \ln(p(\mathbf{x})) + (1 - y_i) \ln(1 - p(\mathbf{x}_i))$$

- Quando $y_i = 0$, o LLF da observação correspondente é igual a $\ln(1 - p(\mathbf{x}))$
- Se $p(\mathbf{x})$ está próximo de $y_i = 0$, então $\ln(1 - p(\mathbf{x}_i))$ está próximo de 0

Regressão Logística

Metodologia

$$LLF = \sum_i y_i \ln(p(\mathbf{x})) + (1 - y_i) \ln(1 - p(\mathbf{x}_i))$$

- Quando $y_i = 0$, o LLF da observação correspondente é igual a $\ln(1 - p(\mathbf{x}))$
- Se $p(\mathbf{x})$ está próximo de $y_i = 0$, então $\ln(1 - p(\mathbf{x}_i))$ está próximo de 0
  - Este é o resultado que desejamos!

Regressão Logística

Metodologia

$$LLF = \sum_i y_i \ln(p(\mathbf{x})) + (1 - y_i)\ln(1 - p(\mathbf{x}_i))$$

- Quando $y_i = 0$, o LLF da observação correspondente é igual a $\ln(1 - p(\mathbf{x}))$
- Se $p(\mathbf{x})$ está próximo de $y_i = 0$, então $\ln(1 - p(\mathbf{x}_i))$ está próximo de 0
  - Este é o resultado que desejamos!
- Se $p(\mathbf{x})$ estiver longe de 0, então $\ln(1 - p(\mathbf{x}_i))$ cai significativamente

Regressão Logística

Metodologia

$$LLF = \sum_i y_i \ln(p(\mathbf{x})) + (1 - y_i)\ln(1 - p(\mathbf{x}_i))$$

- Quando $y_i = 0$, o LLF da observação correspondente é igual a $\ln(1 - p(\mathbf{x}))$
- Se $p(\mathbf{x})$ está próximo de $y_i = 0$, então $\ln(1 - p(\mathbf{x}_i))$ está próximo de 0
  - Este é o resultado que desejamos!
- Se $p(\mathbf{x})$ estiver longe de 0, então $\ln(1 - p(\mathbf{x}_i))$ cai significativamente
  - Não queremos esse resultado porque nosso objetivo é obter o LLF máximo

Regressão Logística

Metodologia

$$LLF = \sum_i y_i \ln(p(\mathbf{x})) + (1 - y_i) \ln(1 - p(\mathbf{x}_i))$$

- Quando $y_i = 0$, o LLF da observação correspondente é igual a $\ln(1 - p(\mathbf{x}))$
- Se $p(\mathbf{x})$ está próximo de $y_i = 0$, então $\ln(1 - p(\mathbf{x}_i))$ está próximo de 0
  - Este é o resultado que desejamos!
- Se $p(\mathbf{x})$ estiver longe de 0, então $\ln(1 - p(\mathbf{x}_i))$ cai significativamente
  - Não queremos esse resultado porque nosso objetivo é obter o LLF máximo
- Da mesma forma, quando $y_i = 1$ o LLF para essa observação é $y_i \ln(p(\mathbf{x}_i))$

Regressão Logística

Metodologia

$$LLF = \sum_i y_i \ln(p(\mathbf{x})) + (1 - y_i)\ln(1 - p(\mathbf{x}_i))$$

- Quando $y_i = 0$, o LLF da observação correspondente é igual a $\ln(1 - p(\mathbf{x}))$
- Se $p(\mathbf{x})$ está próximo de $y_i = 0$, então $\ln(1 - p(\mathbf{x}_i))$ está próximo de 0
  - Este é o resultado que desejamos!
- Se $p(\mathbf{x})$ estiver longe de 0, então $\ln(1 - p(\mathbf{x}_i))$ cai significativamente
  - Não queremos esse resultado porque nosso objetivo é obter o LLF máximo
- Da mesma forma, quando $y_i = 1$ o LLF para essa observação é $y_i \ln(p(\mathbf{x}_i))$
  - Se $p(\mathbf{x}_i)$ estiver próximo de $y_i = 1$, então $\ln(p(\mathbf{x}_i))$ está próximo de 0

$$LLF = \sum_i y_i \ln(p(\mathbf{x})) + (1 - y_i)\ln(1 - p(\mathbf{x}_i))$$

- Quando $y_i = 0$, o LLF da observação correspondente é igual a $\ln(1 - p(\mathbf{x}))$
- Se $p(\mathbf{x})$ está próximo de $y_i = 0$, então $\ln(1 - p(\mathbf{x}_i))$ está próximo de 0
  - Este é o resultado que desejamos!
- Se $p(\mathbf{x})$ estiver longe de 0, então $\ln(1 - p(\mathbf{x}_i))$ cai significativamente
  - Não queremos esse resultado porque nosso objetivo é obter o LLF máximo
- Da mesma forma, quando $y_i = 1$ o LLF para essa observação é $y_i \ln(p(\mathbf{x}_i))$
  - Se $p(\mathbf{x}_i)$ estiver próximo de $y_i = 1$, então $\ln(p(\mathbf{x}_i))$ está próximo de 0
  - Se $p(\mathbf{x}_i)$ estiver longe de 1, então $\ln(p(\mathbf{x}_i))$ é um grande número negativo

Regressão Logística

Metodologia

- Existem diversas abordagens matemáticas para calcular os melhores pesos que correspondem ao LLF máximo

- Existem diversas abordagens matemáticas para calcular os melhores pesos que correspondem ao LLF máximo
  - Trata-se de um problema de otimização numérica

Regressão Logística

Metodologia

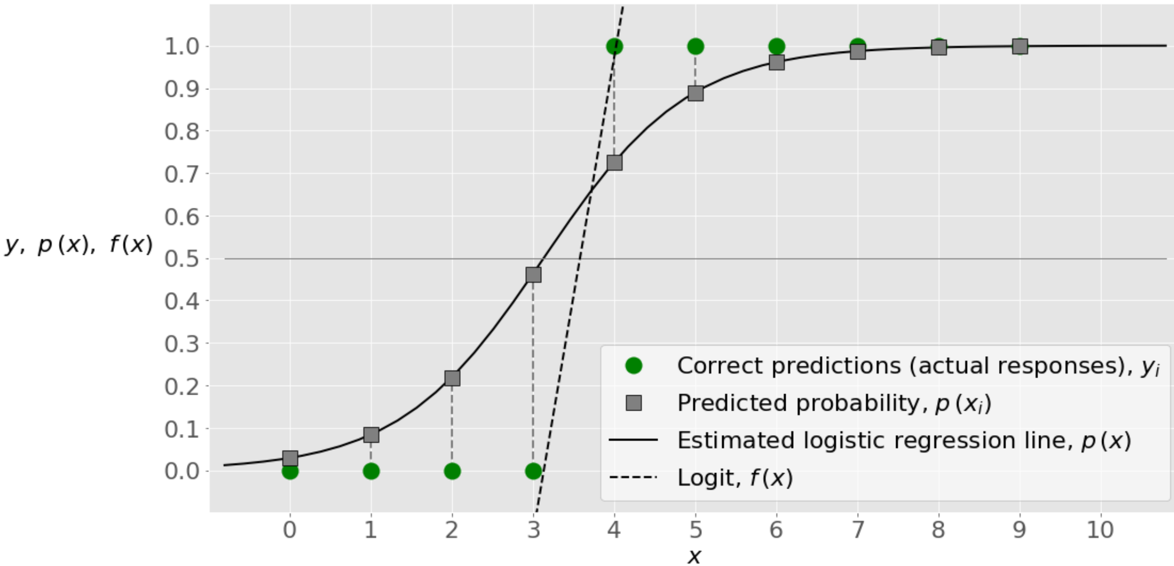
- Existem diversas abordagens matemáticas para calcular os melhores pesos que correspondem ao LLF máximo
  - Trata-se de um problema de otimização numérica
- Depois de determinar os melhores pesos que definem a função $p(\mathbf{x})$, podemos obter as saídas previstas $p(\mathbf{x}_i)$ para qualquer entrada $\mathbf{x}_i$

- Existem diversas abordagens matemáticas para calcular os melhores pesos que correspondem ao LLF máximo
  - Trata-se de um problema de otimização numérica
- Depois de determinar os melhores pesos que definem a função $p(\mathbf{x})$, podemos obter as saídas previstas $p(\mathbf{x}_i)$ para qualquer entrada $\mathbf{x}_i$
- Para cada observação, a saída prevista será:

$$f(\mathbf{x}_i) = \begin{cases} 1, \text{ se } p(\mathbf{x}_i) > 0.5 \\ 0, \text{ caso contrário} \end{cases}$$

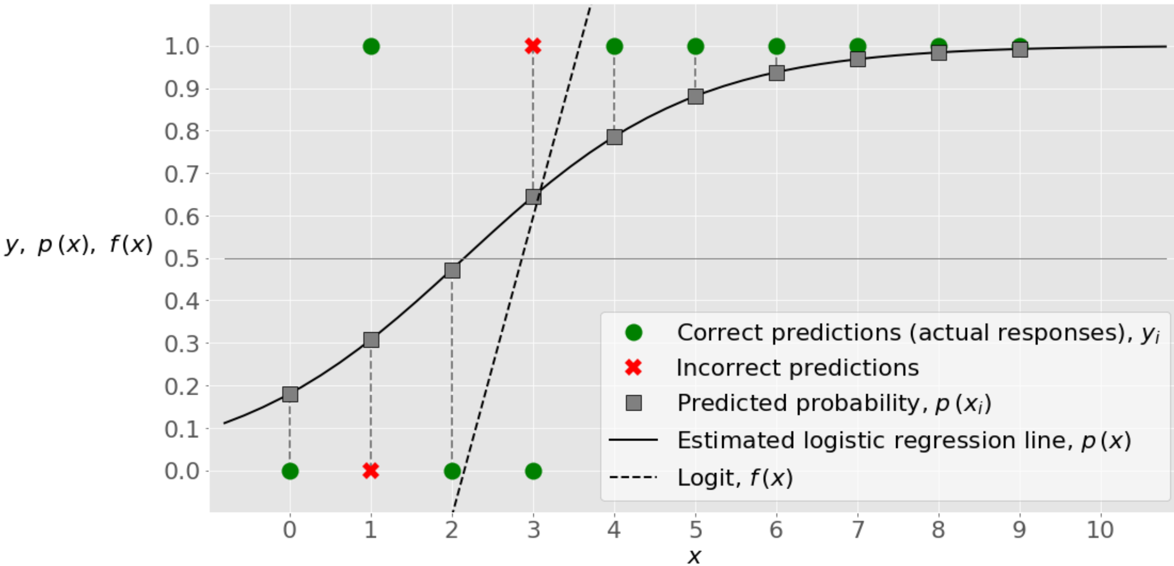
Regressão Logística

Metodologia



Regressão Logística

Metodologia



Regressão Logística
Metodologia

- O limiar (*threshold*) não precisa ser 0.5

Regressão Logística

Metodologia

- O limiar (*threshold*) não precisa ser 0.5
- Podemos definir um valor menor ou maior se for mais conveniente para a situação

Regressão Logística
Metodologia

- O limiar (*threshold*) não precisa ser 0.5
- Podemos definir um valor menor ou maior se for mais conveniente para a situação
- Há mais uma relação importante entre $p(\mathbf{x})$ e $f(\mathbf{x})$:

$$\ln\left(\frac{p(\mathbf{x})}{1-p(\mathbf{x})}\right) = f(\mathbf{x})$$

- O limiar (*threshold*) não precisa ser 0.5
- Podemos definir um valor menor ou maior se for mais conveniente para a situação
- Há mais uma relação importante entre $p(\mathbf{x})$ e $f(\mathbf{x})$:

$$\ln\left(\frac{p(\mathbf{x})}{1 - p(\mathbf{x})}\right) = f(\mathbf{x})$$

- Isso implica que $p(\mathbf{x}) = 0.5$ quando $f(\mathbf{x}) = 0$ e que a saída prevista é 1

- O limiar (*threshold*) não precisa ser 0.5
- Podemos definir um valor menor ou maior se for mais conveniente para a situação
- Há mais uma relação importante entre $p(\mathbf{x})$ e $f(\mathbf{x})$:

$$\ln\left(\frac{p(\mathbf{x})}{1-p(\mathbf{x})}\right) = f(\mathbf{x})$$

- Isso implica que $p(\mathbf{x}) = 0.5$ quando $f(\mathbf{x}) = 0$ e que a saída prevista é 1
  - Essa igualdade explica porque $f(\mathbf{x})$ é o logit.

- A regressão logística de variável única (univariada) é o caso mais direto de regressão logística

- A regressão logística de variável única (univariada) é o caso mais direto de regressão logística
- Existe apenas uma variável independente: $\mathbf{x} = x$

- A regressão logística de variável única (univariada) é o caso mais direto de regressão logística
- Existe apenas uma variável independente: $\mathbf{x} = x$
- A regressão logística encontra os pesos b_0 e b_1 que correspondem ao LLF máximo

Regressão Logística

Regressão Logística Univariada

- A regressão logística de variável única (univariada) é o caso mais direto de regressão logística
- Existe apenas uma variável independente: $\mathbf{x} = x$
- A regressão logística encontra os pesos b_0 e b_1 que correspondem ao LLF máximo
- Esses pesos definem o logit $f(x) = b_0 + b_1 x_1$

Regressão Logística

Regressão Logística Univariada

- A regressão logística de variável única (univariada) é o caso mais direto de regressão logística
- Existe apenas uma variável independente: $\mathbf{x} = x$
- A regressão logística encontra os pesos b_0 e b_1 que correspondem ao LLF máximo
- Esses pesos definem o logit $f(x) = b_0 + b_1 x_1$
- Os pesos também definem a probabilidade prevista $p(x) = \frac{1}{1+\exp(1-f(x))}$

Regressão Logística

Regressão Logística Multivariada

- A regressão logística multivariada possui mais de uma variável de entrada

- A regressão logística multivariada possui mais de uma variável de entrada
- No caso de duas variáveis, temos que determinar os pesos b_0, b_1 e b_2

- A regressão logística multivariada possui mais de uma variável de entrada
- No caso de duas variáveis, temos que determinar os pesos b_0, b_1 e b_2
- Assim:
  - O logit é dado por $f(x_1, x_2) = b_0 + b_1x_1 + b_2x_2$
  - As probabilidades são $p(x_1, x_2) = \frac{1}{1+\exp(1-f(x_1,x_2))}$

Regressão Logística

Desempenho do classificador

A classificação binária possui quatro tipos de resultados possíveis:

A classificação binária possui quatro tipos de resultados possíveis:

Verdadeiros negativos : negativos previstos corretamente (zeros)

A classificação binária possui quatro tipos de resultados possíveis:

Verdadeiros negativos : negativos previstos corretamente (zeros)

Verdadeiros positivos : positivos (uns) previstos corretamente

A classificação binária possui quatro tipos de resultados possíveis:

Verdadeiros negativos : negativos previstos corretamente (zeros)

Verdadeiros positivos : positivos (uns) previstos corretamente

Falsos negativos : negativos previstos incorretamente (zeros)

Regressão Logística

Desempenho do classificador

A classificação binária possui quatro tipos de resultados possíveis:

Verdadeiros negativos : negativos previstos corretamente (zeros)

Verdadeiros positivos : positivos (uns) previstos corretamente

Falsos negativos : negativos previstos incorretamente (zeros)

Falsos positivos : positivos (uns) previstos incorretamente

Regressão Logística

Desempenho do classificador

- Geralmente avaliamos o desempenho do classificador comparando os resultados reais e previstos e contando as previsões corretas e incorretas

- Geralmente avaliamos o desempenho do classificador comparando os resultados reais e previstos e contando as previsões corretas e incorretas
- O indicador mais adequado depende do problema de interesse

- Geralmente avaliamos o desempenho do classificador comparando os resultados reais e previstos e contando as previsões corretas e incorretas
- O indicador mais adequado depende do problema de interesse
- O indicador mais direto da precisão da classificação é a razão entre o número de previsões corretas e o número total de previsões (ou observações)

Outros indicadores de classificadores binários incluem:

Outros indicadores de classificadores binários incluem:

Valor preditivo positivo : a razão entre o número de verdadeiros positivos e a soma dos números de verdadeiros e falsos positivos

Outros indicadores de classificadores binários incluem:

Valor preditivo positivo : a razão entre o número de verdadeiros positivos e a soma dos números de verdadeiros e falsos positivos

Valor preditivo negativo : a razão entre o número de verdadeiros negativos e a soma do número de verdadeiros e falsos negativos

Outros indicadores de classificadores binários incluem:

Outros indicadores de classificadores binários incluem:

Sensibilidade (*recall* ou taxa de verdadeiros positivos): razão entre o número de verdadeiros positivos e o número de verdadeiros positivos

Regressão Logística

Desempenho do classificador

Outros indicadores de classificadores binários incluem:

Sensibilidade (*recall* ou taxa de verdadeiros positivos): razão entre o número de verdadeiros positivos e o número de verdadeiros positivos

Especificidade (taxa de verdadeiros negativos): a razão entre o número de verdadeiros negativos e o número de verdadeiros negativos

Regressão Logística

Regularização

- Um dos principais problemas encontrados no aprendizado de máquina é o *overfitting* ("sobre-ajuste")

Regressão Logística

Regularização

- Um dos principais problemas encontrados no aprendizado de máquina é o *overfitting* ("sobre-ajuste")
- Ocorre quando um modelo aprende muito bem os dados de treinamento
    - O modelo aprende não apenas as relações entre os dados, mas também o ruído no conjunto de dados

Regressão Logística

Regularização

- Um dos principais problemas encontrados no aprendizado de máquina é o *overfitting* (“sobre-ajuste“)
- Ocorre quando um modelo aprende muito bem os dados de treinamento
  - O modelo aprende não apenas as relações entre os dados, mas também o ruído no conjunto de dados
- Modelos superajustados tendem a ter bom desempenho com os dados usados para ajustá-los (os dados de treinamento)

Regressão Logística

Regularização

- Um dos principais problemas encontrados no aprendizado de máquina é o *overfitting* (“sobre-ajuste“)
- Ocorre quando um modelo aprende muito bem os dados de treinamento
  - O modelo aprende não apenas as relações entre os dados, mas também o ruído no conjunto de dados
- Modelos superajustados tendem a ter bom desempenho com os dados usados para ajustá-los (os dados de treinamento)
- Porém, tendem a se comportam mal com dados não vistos (ou dados de teste)

- Um dos principais problemas encontrados no aprendizado de máquina é o *overfitting* (“sobre-ajuste“)
- Ocorre quando um modelo aprende muito bem os dados de treinamento
  - O modelo aprende não apenas as relações entre os dados, mas também o ruído no conjunto de dados
- Modelos superajustados tendem a ter bom desempenho com os dados usados para ajustá-los (os dados de treinamento)
- Porém, tendem a se comportam mal com dados não vistos (ou dados de teste)
- O *overfitting* geralmente ocorre com modelos complexos

Regressão Logística

Regularização

- Na regressão logística, a regularização é uma técnica usada para prevenir o *overfitting*

Regressão Logística

Regularização

- Na regressão logística, a regularização é uma técnica usada para prevenir o *overfitting*
- A regularização adiciona uma penalidade à função de custo
  - Ou seja, penaliza a função que o modelo tenta minimizar durante o treinamento

- Na regressão logística, a regularização é uma técnica usada para prevenir o *overfitting*
- A regularização adiciona uma penalidade à função de custo
  - Ou seja, penaliza a função que o modelo tenta minimizar durante o treinamento
- Essa penalidade desestimula o modelo de atribuir pesos (coeficientes) muito altos às variáveis independentes

Regressão Logística
Regularização

As duas formas mais comuns de penalização são:

1. **L1**, ou *Lasso*
2. **L2**, ou *Ridge*

Regressão Logística

Penalidade L2 (Ridge)

- A penalidade L2 adiciona a soma dos quadrados dos coeficientes à função de custo

Regressão Logística

Penalidade L2 (Ridge)

- A penalidade L2 adiciona a soma dos quadrados dos coeficientes à função de custo
- Essa penalidade tende a reduzir os coeficientes sem necessariamente zerá-los

Regressão Logística

Penalidade L2 (Ridge)

- A penalidade L2 adiciona a soma dos quadrados dos coeficientes à função de custo
- Essa penalidade tende a reduzir os coeficientes sem necessariamente zerá-los
- Isso significa que ela distribui a penalização por todos os coeficientes, tornando o modelo mais robusto e menos suscetível ao *overfitting*

Regressão Logística

Penalidade L2 (Ridge)

- A penalidade L2 adiciona a soma dos quadrados dos coeficientes à função de custo
- Essa penalidade tende a reduzir os coeficientes sem necessariamente zerá-los
- Isso significa que ela distribui a penalização por todos os coeficientes, tornando o modelo mais robusto e menos suscetível ao *overfitting*
- É o valor padrão em muitas implementações, incluindo a regressão logística do `scikit-learn`

REGRESSÃO LOGÍSTICA

PENALIDADE L2 (RIDGE)

Matematicamente:

$$\text{RSS} = \underbrace{\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{r} x_{ij} b_j \right)^2}_{\text{perda}} + \underbrace{\lambda \cdot \sum_{j=1}^{r} b_j^2}_{\text{regularização}}$$

sendo que:

- y_i são as variáveis independentes
- x_{ij} são as variáveis dependentes
- b_j são os coeficientes
- λ é a regularização ("lambda")
- n é o número de observações
- r é o número de atributos

Regressão Logística

Penalidade L1 (Lasso)

- A penalidade L1 adiciona a soma dos valores absolutos dos coeficientes à função de custo

Regressão Logística

Penalidade L1 (Lasso)

- A penalidade L1 adiciona a soma dos valores absolutos dos coeficientes à função de custo
- Essa penalidade tende a forçar alguns coeficientes a serem exatamente zero

Regressão Logística

Penalidade L1 (Lasso)

- A penalidade L1 adiciona a soma dos valores absolutos dos coeficientes à função de custo
- Essa penalidade tende a forçar alguns coeficientes a serem exatamente zero
- Isso é útil para a seleção de características, pois elimina as variáveis menos importantes, simplificando o modelo

Regressão Logística

Penalidade L1 (Lasso)

- A penalidade L1 adiciona a soma dos valores absolutos dos coeficientes à função de custo
- Essa penalidade tende a forçar alguns coeficientes a serem exatamente zero
- Isso é útil para a seleção de características, pois elimina as variáveis menos importantes, simplificando o modelo
- É útil quando se espera que apenas algumas variáveis sejam relevantes, ajudando na criação de um modelo mais interpretável

Regressão Logística

Penalidade L1 (Lasso)

Matematicamente:

$$\text{RSS} = \underbrace{\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{r} x_{ij} b_j \right)^2}_{\text{perda}} + \underbrace{\lambda \cdot \sum_{j=1}^{r} |b_j|}_{\text{regularização}}$$

sendo que:

- y_i são as variáveis independentes
- x_{ij} são as variáveis dependentes
- b_j são os coeficientes
- λ é a regularização ("lambda")
- n é o número de observações
- r é o número de atributos

- Além das penalidades L1 e L2, também existe a combinação das duas, chamada Elastic Net

- Além das penalidades L1 e L2, também existe a combinação das duas, chamada Elastic Net
- A Elastic Net traz benefícios tanto de L1 quanto de L2, ajudando na seleção de características e na redução dos coeficientes

ELASTIC NET

- Além das penalidades L1 e L2, também existe a combinação das duas, chamada Elastic Net
- A Elastic Net traz benefícios tanto de L1 quanto de L2, ajudando na seleção de características e na redução dos coeficientes
- É útil quando se tem um grande número de variáveis correlacionadas

Regressão Logística
Elastic Net

Matematicamente:

$$\text{RSS} = \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{r} x_{ij}b_j\right)^2 + \lambda_1 \cdot \sum_{j=1}^{r}|b_j| + \lambda_2 \cdot \sum_{j=1}^{r} b_j^2$$

sendo que:

- y_i são as variáveis independentes
- x_{ij} são as variáveis dependentes
- b_j são os coeficientes
- λ_1 é a regularização L1
- λ_2 é a regularização L2
- n é o número de observações
- r é o número de atributos

- A regressão logística é um modelo fundamental para a classificação binária, funcionando por meio da função sigmoide
- Oferece uma maneira direta e interpretável de modelar a probabilidade de uma classe
- É especialmente útil quando se deseja entender a influência das variáveis independentes

Leitura Recomendada

Fernandes, A. A. T., Figueiredo Filho, D. B., Rocha, E. C. da ., & Nascimento, W. da S.. (2020). Read this paper if you want to learn logistic regression. *Revista de Sociologia e Política*, 28(74), 006. https://doi.org/10.1590/1678-987320287406en