



Aprendizado Supervisionado

Classificação

Prof. Raphael Carvalho

Classificação



Classificação

- É uma técnica de aprendizado supervisionado para

distribuir um conjunto de dados de entrada em categorias ou classes pré-definidas de saída.

- No exemplo anterior, podemos utilizar um algoritmo de classificação binária para decidir se uma mariposa é da espécie **Imperatriz** (imagem da esquerda) ou **Luna** (imagem da direita)

Como o algoritmo pode decidir algo assim?



Como o algoritmo pode decidir algo assim?

- Supervisor pode escolher um conjunto de *features* (também chamados de características ou qualidades). ● Essas *features* são basicamente valores que caracterizam de forma útil as coisas que desejamos classificar. ● Para o nosso exemplo, vamos utilizar duas features: envergadura e massa.
- Para treinar nosso algoritmo de classificação para fazer boas previsões, vamos precisar de dados de treinamento

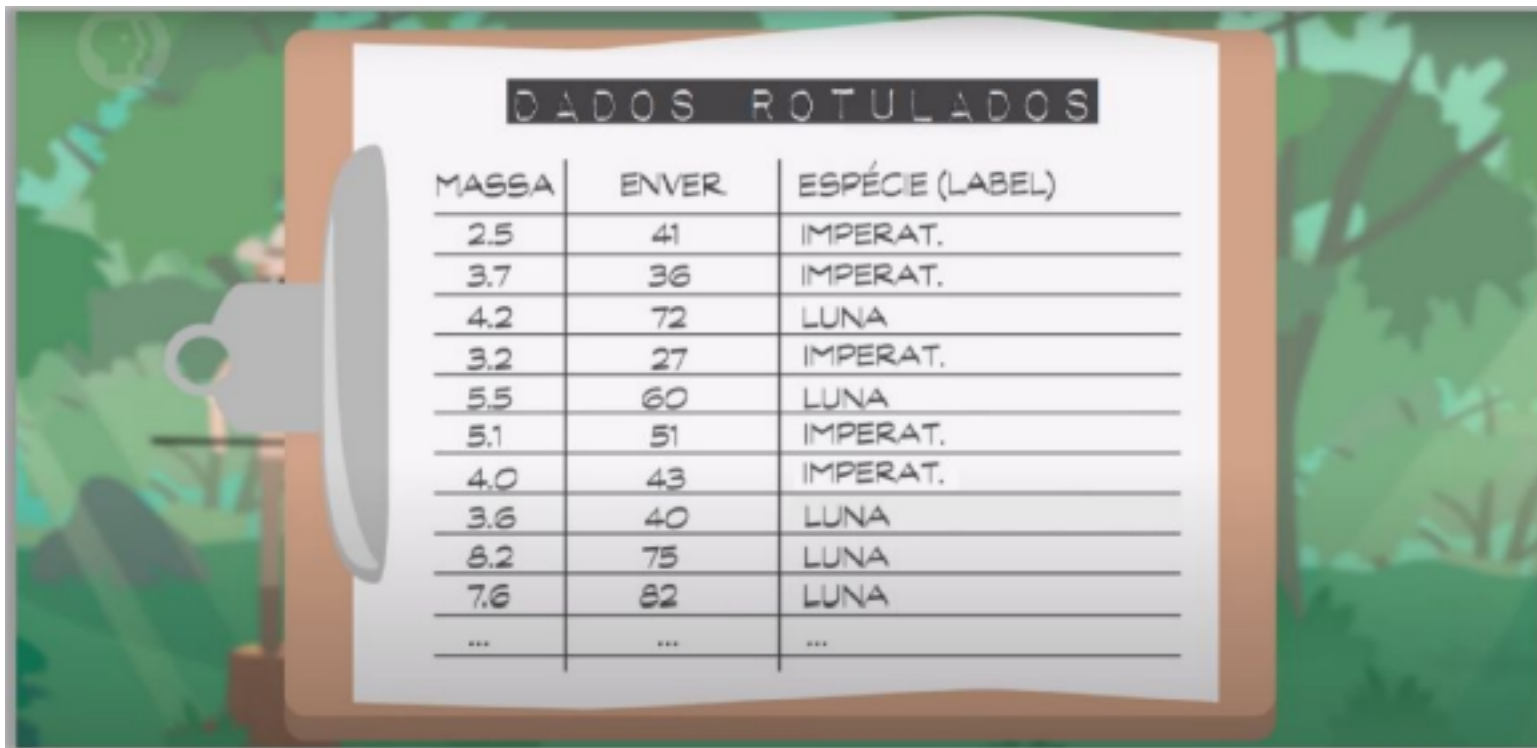
Aprendizado de Máquina

- Uma forma de tentar resolver esse problema é por meio do aprendizado de máquina.
- Ocorre que funciona de uma maneira praticamente inversa à programação tradicional.
- Continuamos entrando com dados, mas – em vez de um programador criar manualmente as regras – são inseridos exemplos de resultados passados.
- A saída de um algoritmo de aprendizado de máquina são justamente as regras.

E como conseguimos?

- Podemos enviar um entomologista a uma floresta para capturar diversas mariposas de ambas as espécies, examiná-las e registrar os valores das duas features que nós escolhemos em uma tabela.
- A tabela final resultante contendo todas as mariposas catalogadas com seu rótulo de espécie, envergadura e massa é também chamado de dados rotulados

E como conseguimos?

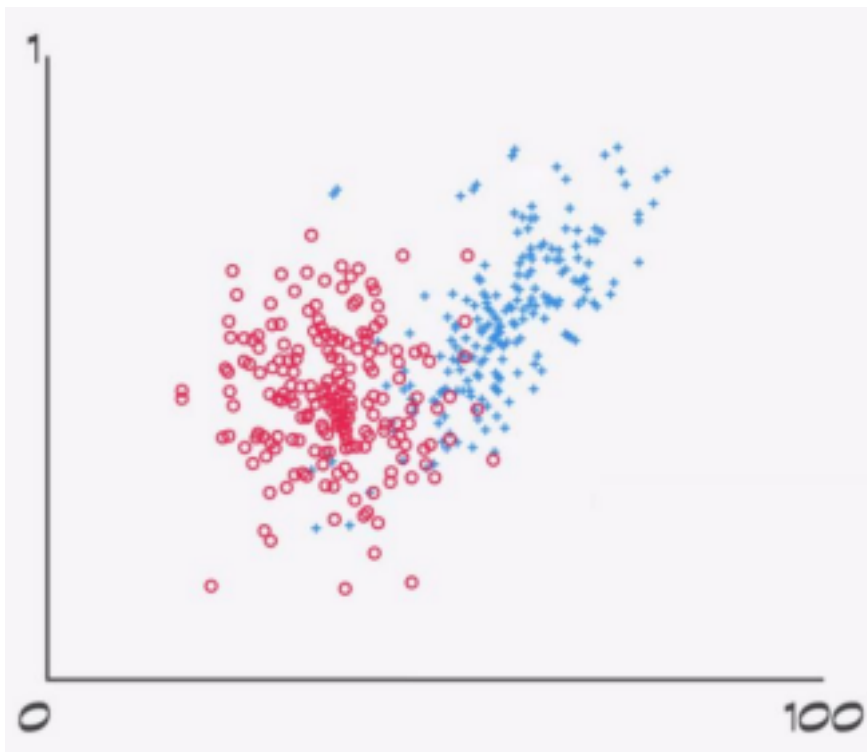


| MASSA | ENVER | ESPÉCIE (LABEL) |
|-------|-------|-----------------|
| 2.5 | 41 | IMPERAT. |
| 3.7 | 36 | IMPERAT. |
| 4.2 | 72 | LUNA |
| 3.2 | 27 | IMPERAT. |
| 5.5 | 60 | LUNA |
| 5.1 | 51 | IMPERAT. |
| 4.0 | 43 | IMPERAT. |
| 3.6 | 40 | LUNA |
| 8.2 | 75 | LUNA |
| 7.6 | 82 | LUNA |
| ... | ... | ... |

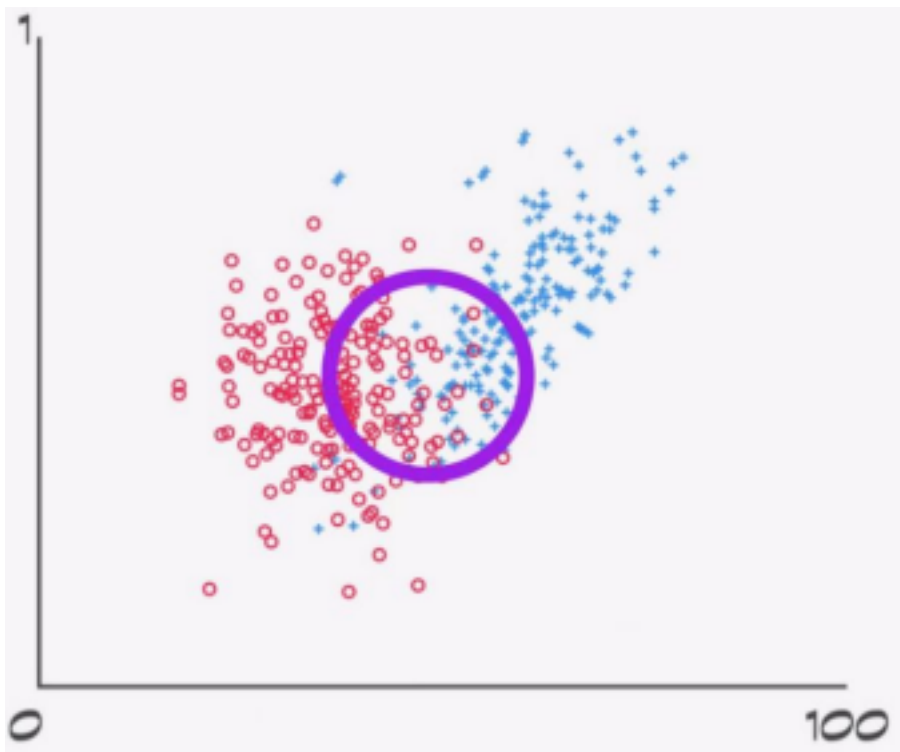
Visualização dos dados

- Como temos apenas duas features, é fácil visualizar esses dados em um gráfico de dispersão.
- Na imagem da esquerda, eu plotei os dados de 100 mariposas imperatriz em vermelho e 100 mariposas luna em azul.
- No eixo horizontal, temos a envergadura em milímetros; no eixo vertical, temos a massa em gramas.

Visualização dos dados



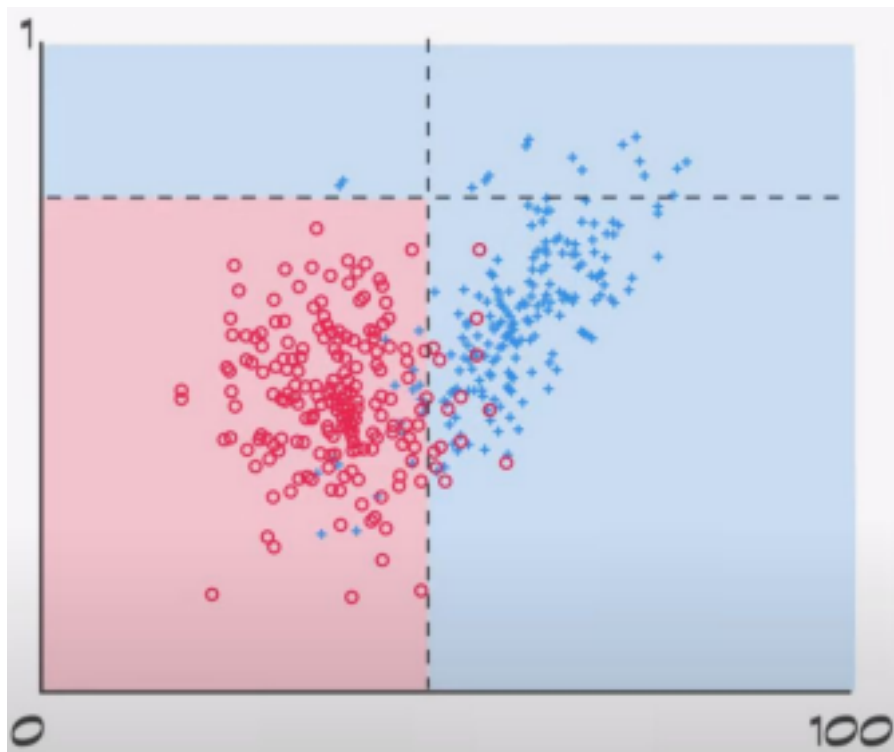
Visualização dos dados



Visualização dos dados

- Essa sobreposição indica que não é tão óbvio separar dois grupos.
- É aqui que entram os algoritmos de aprendizado de máquina: eles são capazes de encontrar uma divisão ideal entre os dois grupos
- Vamos explicar isso melhor!

Entendo a separação dos dados



Entendendo os resultados

- Essas linhas que dividem o espaço de decisão são chamadas de limites de decisão porque auxiliam a indicar qual será o classificador sugerido.
- Após essa separação, devemos comparar os dados rotulados na tabela com os dados resultantes do gráfico de dispersão para verificar se as linhas sugeridas fizeram uma divisão satisfatória ou não para identificar as espécies de mariposa
- Esse é uma das formas de avaliar se o processo de classificação foi satisfatório ou não

Entendendo os resultados

- Se pudéssemos contar, iremos verificar que 86 mariposas imperatriz (vermelho) terminaram de forma correta dentro da região de decisão (em vermelho), mas 14 delas acabaram de forma incorreta no território da mariposa luna (em azul).
- Por outro lado, 82 mariposas luna (azul) foram classificadas corretamente (em azul), com 18 caindo para o lado errado (em vermelho).

Matriz de Confusão

- É uma tabela utilizada para avaliar a qualidade de um modelo que mostra as frequências de classificação para cada classificador/rótulo do modelo
- Trata-se geralmente de uma tabela com duas linhas e duas colunas que exhibe a quantidade de erros e acertos de classificação de uma amostra de dados
- Possui dois eixos:
 - Horizontal: indica o valor previsto ou esperado
 - Vertical: indica o valor real

Matriz de Confusão

| | | VALOR PREVISTO | |
|------------|------------|----------------|------|
| | | IMPERATRIZ | LUNA |
| VALOR REAL | IMPERATRIZ | 86 | 14 |
| | LUNA | 18 | 82 |

Como posso interpretar essa matriz?

- 86 mariposas das 100 coletadas como mariposa imperatriz foram classificadas corretamente
- 14 mariposas das 100 coletadas como mariposa imperatriz foram classificadas incorretamente como luna
- 18 mariposas das 100 coletadas como mariposa luna foram classificadas incorretamente como mariposa imperatriz
- 82 mariposas das 100 coletadas como mariposa luna foram classificadas corretamente

Analizando os resultados

- Não há como desenharmos linhas que nos forneçam 100% de acurácia!
 - Se reduzirmos o valor limite de decisão da envergadura da mariposa (eixo horizontal), classificaremos erroneamente mais mariposas imperatriz como mariposas luna
- Por outro lado, se o aumentarmos, classificaremos incorretamente mais mariposas luna.

Analizando os resultados



- Trabalho dos algoritmos de aprendizado de máquina é tentar maximizar as classificações corretas enquanto minimiza seus erros
- Como podemos medir o desempenho do modelo? ● Podemos fazer utilizando a métrica de acurácia, que é a divisão do número de acertos pelo total de predições. ● No exemplo:
 - Tivemos $82+86 = 168$ acertos em uma amostra de 200 mariposas.
 - Acurácia foi de $168/200 = 84\%$.

Analizando os resultados

| | | VALOR PREVISTO | |
|------------|----------|------------------------|------------------------|
| | | CLASSE 1 | CLASSE 2 |
| VALOR REAL | CLASSE 2 | VERDADEIRO CLASSE 1 | FALSO CLASSE 2 |
| | CLASSE 1 | FALSO CLASSE 1 | VERDADEIRO CLASSE 2 |

Analizando os resultados

- Outra forma de representar os resultados é por meio da determinação Positivo ou Negativo
- Em um classificação binária, temos apenas dois rótulos (Classe 1 = Positivo e Classe 2 = Negativo)
- Resultados possíveis:
 - Verdadeiro-Positivo (VP)
 - Falso-Positivo (FP)
 - Verdadeiro-Negativo (VN)
 - Falso-Negativo (FN).

Analisando os resultados

| | | VALOR PREVISTO | |
|------------|----------|------------------------|------------------------|
| | | CLASSE 1 | CLASSE 2 |
| VALOR REAL | CLASSE 2 | VERDADEIRO CLASSE 1 | ERRO TIPO II |
| | CLASSE 1 | ERRO TIPO I | VERDADEIRO CLASSE 2 |

| | | VALOR PREVISTO | |
|------------|----------|------------------------|------------------------|
| | | POSITIVO | NEGATIVO |
| VALOR REAL | POSITIVO | VERDADEIRO POSITIVO | FALSO NEGATIVO |
| | NEGATIVO | FALSO POSITIVO | VERDADEIRO NEGATIVO |

Medindo o desempenho do modelo

- Uma maneira de medir o desempenho de um processo de classificação é por meio da acurácia, mas existem outras formas de medir

Acurácia

- É a métrica mais simples que permite mensurar o percentual de acertos, isto é, a quantidade de previsões corretas dentro do total de previsões possíveis.
- Responde à pergunta: dentre todas as previsões realizadas, quantas o modelo acertou?

$$\frac{VP + VN}{VP + FP + VN + FN}$$

Sensibilidade

- É métrica que permite avaliar a capacidade do classificador de detectar com sucesso resultados positivos
- Também conhecida como recall (revocação). •

Responde à pergunta: dentre os valores realmente positivos, quantos o modelo previu corretamente como

positivo?

$$\frac{VP}{VP + FN}$$

Precisão

- É a métrica que permite mensurar a proporção de previsões positivas corretas sobre a soma de todos os valores positivos.
- Responde à pergunta: dentre os valores previstos como

positivos, quantos o modelo acertou (previu corretamente como positivo)?

$$\frac{VP}{VP + FP}$$

F1-Score

- É a média harmônica calculada com base na precisão e na sensibilidade
- Essa medida tenta condensar em uma única medida um

pouco da precisão e um pouco da sensibilidade.



Para que tantas medidas diferentes?

- Dependendo do contexto, o desempenho pode ser medido de maneira diferente para refletir melhor a efetividade da medição •
- Precisão pode ser utilizada em situações em que falsos-positivos são mais prejudiciais que falsos-negativos

Exemplo: ao classificar ações da bolsa de valores como boas ou ruins, um falso-positivo pode fazer uma pessoa investir em uma ação ruim e ter prejuízos; já um falso-negativo pode fazer uma pessoa não investir em uma ação boa e deixar de ter lucros, mas ela não terá prejuízos, logo é menos prejudicial.

Para que tantas medidas diferentes?

- Recall pode ser utilizado em situações em que falsos-negativos são mais prejudiciais que falsos-positivos

Exemplo: ao classificar uma pessoa com vacinado ou

não-vacinado, um falso-positivo pode fazer uma pessoa saudável não pegar um avião com outras pessoas; já um falso-negativo pode fazer uma pessoa infectada pegar um avião com outras pessoas e infectá-las com seu vírus