Advanced Institute for Artificial Intelligence

Classificação

Advanced Institute for Artificial Intelligence

https://advancedinstitute.ai

Classificação

Agenda

- O que é classificação?
- Classificador linear
- Avaliação de um classificador

Classificação

- Uma coluna na base de dados rotula cada instância da base de modo qualitativo
- Cada instância pode possuir dois ou mais rótulos, que são chamados de classe
- Um algoritmo de classificação busca descobrir para uma instância nova, a qual classe essa instância pertence, a partir de variáveis preditoras
- A saído do modelo pode ser também uma distribuição de probabilidade associada a cada possível classe da base de dados

Classificação

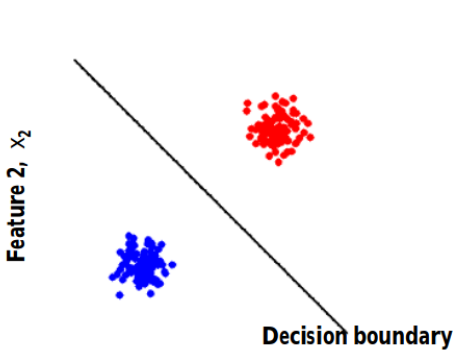
Exemplo de classificação:

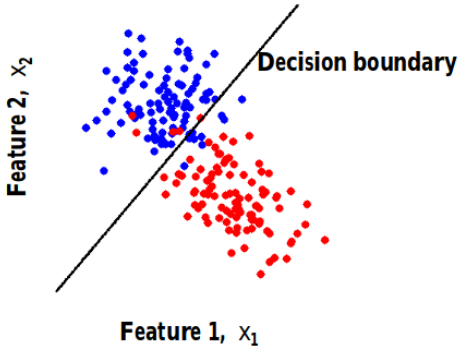
- Diagnóstico médico
- Identificar se um atleta olímpico é halterofilista ou jogador de basquete olhando apenas sua altura e peso
- Detecção de fraude em cartões de crédito
- Filtragem de spam em e-mails
- Bioinformática (sequências de DNA)

Classificação

Um conjunto de dados é separável por um modelo se :

- Existe alguma instância desse aluno que prevê corretamente todos os pontos de dados
- Dados separáveis linearmente
- Podem separar as duas classes usando uma linha reta no espaço de características
- em 2 dimensões o limite de decisão é um linha reta


Feature 2, x_2
Decision boundary


Feature 2, x_2
Decision boundary
Feature 1, x_1

Classificação

Matriz de confusão

- medida efetiva do modelo de classificação
- mostra o número de classificações corretas versus as classificações preditas para cada classe

$$M(C_i, C_j) = \sum_{\{\forall (x,y) \in T : y = C_i\}} \left\| h(x) = C_j \right\|$$

Classe	predita C_1	predita C_2	$\cdots$	predita C_1
verdadeira C_1	$M(C_1, C_1)$	$M(C_1, C_2)$	$\cdots$	$M(C_1, C_k)$
verdadeira C_2	$M(C_2, C_1)$	$M(C_2, C_2)$	$\cdots$	$M(C_2, C_k)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
verdadeira C_k	$M(C_k, C_1)$	$M(C_k, C_2)$	$\cdots$	$M(C_k, C_k)$

Classificação

- O número de acertos, para cada classe, se localiza na diagonal principal M(Ci,Ci) da matriz
- Os demais elementos M(Ci,Cj), para i j, representam erros na classificação
- A matriz de confusão de um classificador ideal possui todos esses elementos iguais a zero uma vez que ele não comete erros

Classe	predita C_+	predita C_-	Taxa de Erro da Classe	Taxa de Erro Total
verdadeira C_+	T_P	F_N	$\frac{F_N}{T_P + F_N}$	$\frac{F_P + F_N}{n}$
verdadeira C_-	F_P	T_N	$\frac{F_P}{F_P + T_N}$	

T_P = Verdadeiro Positivo (True Positive)
F_N = Falso Negativo (False Negative)
F_P = Falso Positivo (False Positive)
T_N = Verdadeiro Negativo (True Negative)
n = (T_P+F_N+F_P+T_N)

Classificação

Matriz de confusão

- medida efetiva do modelo de classificação
- mostra o número de classificações corretas versus as classificações preditas para cada classe

Classificação

- Precisão : TP / (TP + FP)
- Porcentagem de previsões positivas corretas
- Recall: TP / (TP + FN)
- Porcentagem de instâncias rotuladas positivamente, também previstas como positivas
- Acurácia: (TP + TN) / (TP + TN + FP + FN)
- Porcentagem de previsões corretas
- f1 score: média harmonica de precision e recall
- F1 score próximo de 1 indica melhor qualidade