Advanced Institute for Artificial Intelligence

Classificação

Advanced Institute for Artificial Intelligence

https://advancedinstitute.ai

Classificação

Agenda

- O que é classificação?
- Classificador linear
- Naive Bayes
- Matrix de Confusão
- Regressão Logística

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

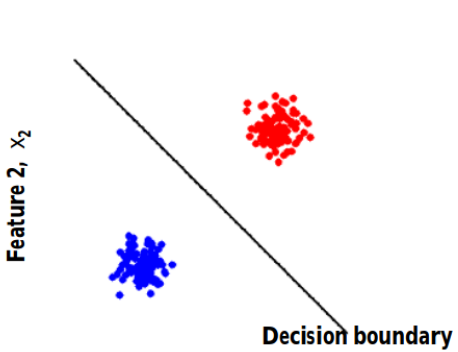
Exemplo de classificação:

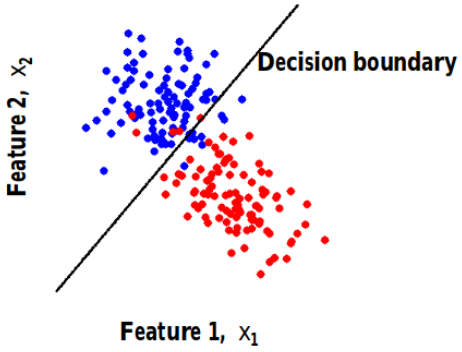
- Diagnóstico médico
- Identificar se um atleta olímpico é halterofilista ou jogador de basquete olhando apenas sua altura e peso
- Detecção de fraude em cartões de crédito
- Filtragem de spam em e-mails
- Bioinformática (sequências de DNA)

Classificação

Um conjunto de dados é separável por um modelo se :

- Existe alguma instância desse conjunto de dados 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

- Para separar as classes linearmente é necessário um algoritmo
- O algoritmo utilizará a relação entre as características e as classes para definir um critério de clasificação
- A predição de novos dados será feita com base no classificador construído com os dados de treinamento

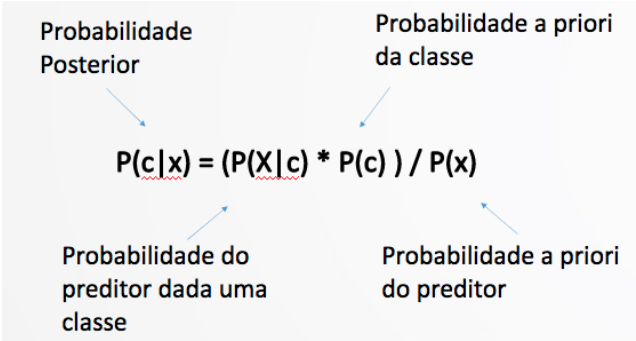
Naive Bayes

Métodos de classificação Bayesianos (teorema de Bayes)

- Baseado na suposição de que as quantidades de interesse são reguladas por distribuições de probabilidades.
- A classificação bayesiana busca definir probabilidade de um rótulo, dados algumas observações
- Métodos Bayesianos requerem o conhecimento inicial de várias probabilidades.
- Quando não conhecidas, podem ser estimadas:
  - a partir de conhecimento prévio
  - dados previamente disponíveis
  - suposições a respeito da forma da distribuição.

Naive Bayes

Teorema de Bayes em aprendizagem de máquina define como obter a probabilidade de uma classe ocorrer, dado uma característica



Naive Bayes

- $P(c|x)$ é a probabilidade posterior da classe (c, destino) dado preditor (x, atributos).
- $P(c)$ é a probabilidade a priori da classe.
- $P(x|c)$ é a probabilidade que é a probabilidade do preditor dada classe.
- $P(x)$ é a probabilidade a priori do preditor.

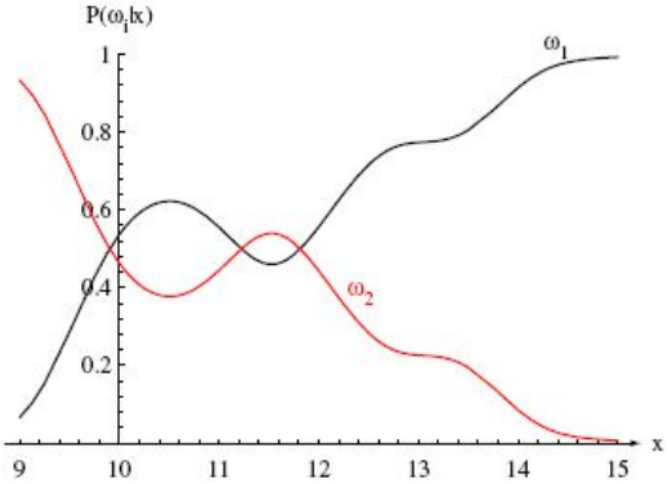
Naive Bayes

- Classe W_1
- Probabilidades a priori $P(W_1)$: Conhecimento a priori que se tem sobre o problema, ou seja, conhecimento a priori sobre a aparição de exemplos das classes do problema.
- Função de Densidade Probabilidade $P(x)$: Frequência com a qual encontramos uma determinada característica (Evidências)

Naive Bayes

- Densidade de Probabilidade Condicional
- $P(X|W_j)$ (Likelihood) - Verossimilhança
- Frequência com que encontramos uma determinada característica x dado que a mesma pertence a classe W_j
- Densidade de duas classes em que x representa uma característica qualquer

Naive Bayes



Naive Bayes

- Probabilidades a posteriori para um valor de $x = 14$,
- a probabilidade do padrão pertencer a W_1 é de 0,92,
- a probabilidade do padrão pertencer a W_2 é de 0,08.
- Para cada x, as probabilidades a posteriori somam 1.

Naive Bayes

- Um dos algoritmos de aprendizagem mais práticos e utilizados na literatura.
- Denominado Naive (ingênuo) por assumir que os atributos são condicionalmente independentes, ou seja, a informação de um evento não é informativa sobre nenhum outro.
- Apesar dessa premissa, o classificador reporta bom desempenho em diversas tarefas de classificação onde há dependência.

Naive Bayes

- Aplica-se a tarefas de aprendizagem onde cada instância x é descrita por uma conjunção de valores de atributos em que a função alvo, $F(x)$, pode assumir qualquer valor de um conjunto V
- Um conjunto de exemplos de treinamento da função alvo é fornecido. E então uma nova instância é apresentada, descrita pela tupla de valores de atributos $a_1, a_2, ...$
- A tarefa é predizer o valor alvo (ou classificação) para esta nova instância.
- Para atributos contínuos o classificador assume que a distribuição de probabilidades dos atributos é normal

Naive Bayes

Algumas estratégias de Naive Bayes

- Gaussiano: assume que as características seguem uma distribuição normal.
- Multinomial: É usado para contagens discretas.
- Bernoulli: o modelo binomial é útil se seus vetores de características são binários

Classificação

Matriz de confusão

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

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

Exemplo: resultado da classificação para 180 amostras (100 da classe 1 e 80 da classe 2)

$$\begin{matrix} 98 & 2 \\ 4 & 76 \end{matrix}$$

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

- 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

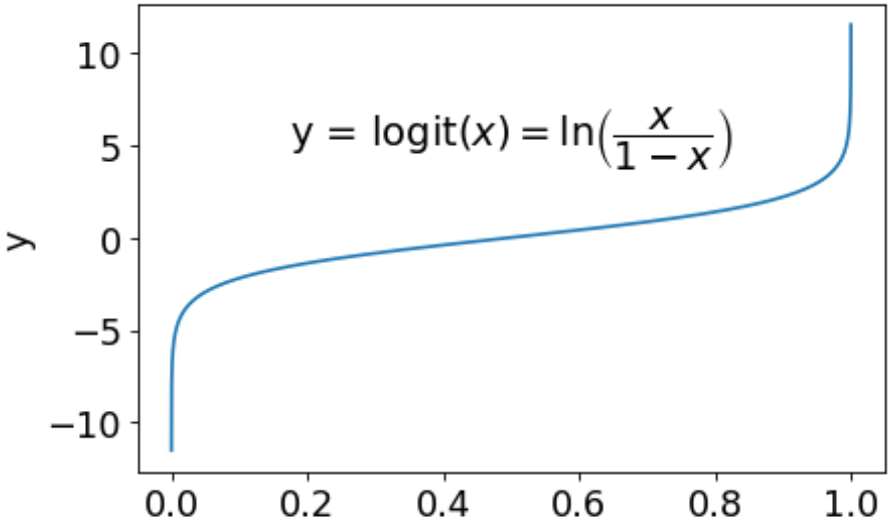
Regressão logística

- A regressão logística binária é um tipo de análise de regressão em que a variável alvo é uma variável qualitativa: 0 ou 1
- Modelar a probabilidade de um evento ocorrer dependendo dos valores das variáveis independentes.
- Distribuição discreta de espaço amostral 0,1 que tem probabilidade de sucesso p e falha q = 1 - p

Regressão logística

- A técnica de Regressão Logística busca um algoritmo que relacione os valores das característica, com uma probabilidade de ocorrência de um evento entre 0 e 1
- Para isso usamos a Função logit como ligação entre a dsitribuição de valores de uma variável e a probabilidade de ocorrência de um valor em y

Regressão logística



Regressão logística

O modelo procurado na regressão logística é:

- $logit(p) = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + ... + + \beta_n * X_n$
- log odds (logit) são linearmente relacionadas às características definidas em X
- Um modelo de regressão logística, modela a probabilidade transformada por logit como uma relação linear com as variáveis preditoras.

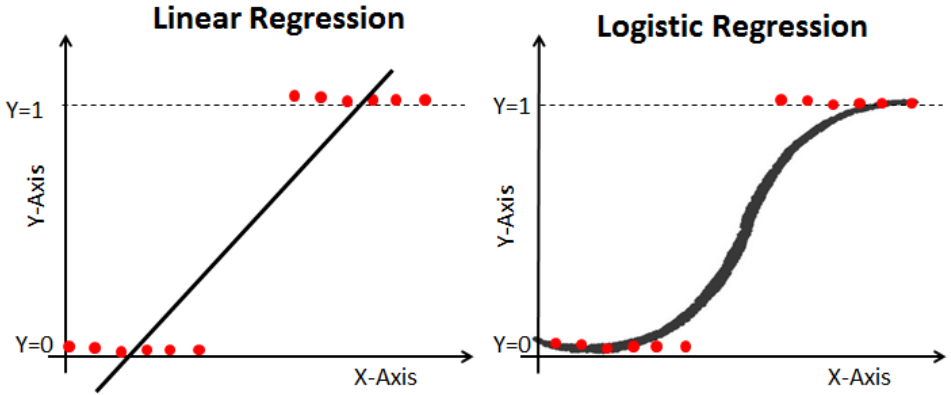
Regressão logística

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

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

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

Regressão logística



Regressão logística

- Na regressão logística estimamos p para qualquer combinação linear das variáveis independentes.
- Isso pode ser alcançado usando o MLE (*Maximum Likehood Estimation*) para estimar os coeficiente do modelo.
- Também pode ser usado descida do gradiente, otimizando os coeficientes para aproximar os valores mais próximos de 0 e 1.