

Processo de Classificação - Projeto sm@rtbee

Rhaniel Magalhães

Universidade Federal do Ceará

rhaniel@alu.ufc.br

18 de Setembro de 2018

Sumário

No início

Análise dos Dados

Algoritmos

kNN

Random Forest

Naive Bayes

Resultados

Acurácia

Matriz de Confusão

Orange Canvas

Próximos Passos

No início

► Plots Simples dos Dados

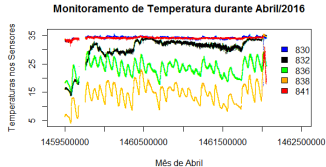


Figura: Plot da Temperatura ao longo de Abril/2016 Tempo

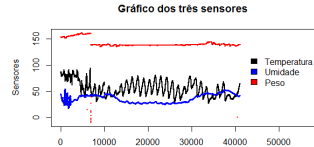


Figura: Plots das três características em um mesmo sensor

Algoritmos

k Nearest Neighbors

Algoritmo de classificação *lazy learning*.

Baseia-se nos vizinhos mais próximos para classificar uma nova entrada de dados.

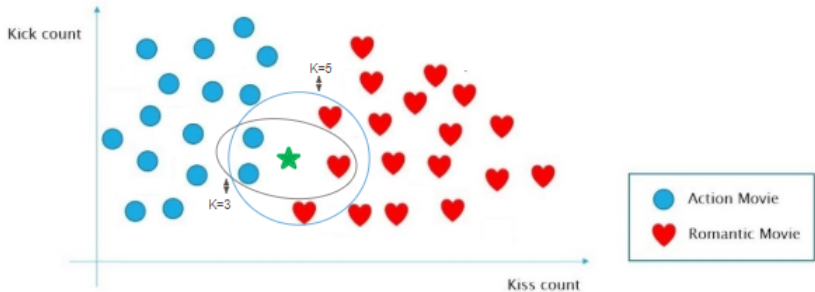


Figura: Classificação de um filme como ação ou Romance

Algoritmos

Random Forest

Algoritmo de classificação baseado em árvores de decisão.

Cada árvore cria regras de decisão baseado em seus dados a fim de classificar novas amostras.

O *Random Forest* gera uma "floresta" com várias dessas árvores a fim de diminuir a variância e aumentar a estabilidade do modelo.

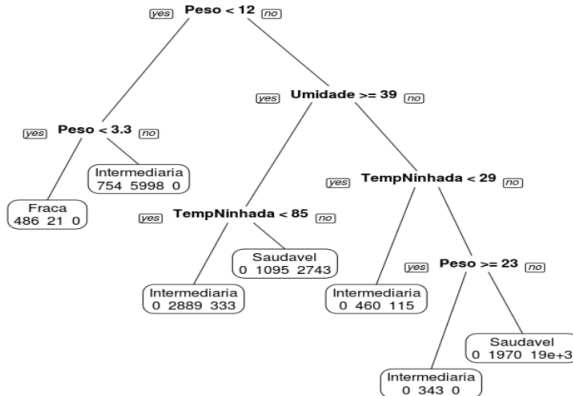


Figura: Árvore de Decisão

Naive Bayes

Baseado na probabilidade bayesiana, o *Naive Bayes* irá calcular as probabilidades das novas amostras pertencerem a cada uma das classes existentes e irá atribuí-las à classe de maior probabilidade.

The diagram shows the Naive Bayes formula for posterior probability:
$$P(c | x) = \frac{P(x | c)P(c)}{P(x)}$$
 Four blue arrows point from the formula to descriptive labels:

- An arrow from $P(c | x)$ points to "Probabilidade posterior" (Posterior Probability).
- An arrow from $P(x | c)$ points to "Probabilidade" (Probability).
- An arrow from $P(c)$ points to "Probabilidade original da Classe" (Original Class Probability).
- An arrow from $P(x)$ points to "Preditor da probabilidade posterior" (Posterior Probability Predictor).

Figura: Probabilidade Bayesiana para o Naive Bayes

Resultados

Foram utilizados os datasets da Arnas para grande parte da análise apresentada.

Algoritmo	Arnas 1º período	Arnas 2º período	Emil 1º período	Emil 2º período
N. Bayes	93.71 $\pm$ 0.31	88.41 $\pm$ 0.38	96.76 $\pm$ 0.19	92.04 $\pm$ 0.29
k-NN	98.96 $\pm$ 0.25	96.53 $\pm$ 0.59	99.63 $\pm$ 0.33	99.30 $\pm$ 0.28
R. Forest	99.13 $\pm$ 0.28	97.07 $\pm$ 0.54	99.67 $\pm$ 0.21	99.24 $\pm$ 0.43

Tabela: Taxas de sucesso dos algoritmos de classificação

Acurácia

	HT101
kNN	66.67 $\pm$ 43.83

Tabela: Acurácia para a colmeia HT101

k-Nearest Neighbors																		
Emil										Arnas								
	1° Período			C0	C1	2° Período				C0	1° Período				C0	2° Período		
	C0	C1	C2			C2	C3	C4	C5		C1	C2	C3	C4		C1	C2	C2
C0	2445	41	16	696	61	64	0	3	0	2273	22	0	3	2	2026	64	129	
C1	66	3087	65	0	1273	8	0	0	0	0	1488	0	0	11	0	1214	1	
C2	96	0	1581	22	258	1112	0	84	0	0	0	2184	0	5	290	102	3060	
C3				0	0	31	963	0	137	3	2	0	1200	0				
C4				13	179	112	0	1174	0	0	20	0	0	446				
C5				0	64	409	0	0	1250									

Tabela: Matrizes de Confusão para o algoritmo *k-Nearest Neighbors*

Bayer

	Sem Cross-Validation			
	C0	C1	C2	C3
C0	0	0	0	1
C1	0	0	1	0
C2	0	0	17	3
C3	0	0	9	5

Tabela: Matriz de Confusão para o kNN sem CV

	Com Cross-Validation			
	C0	C1	C2	C3
C0	0	0	1	0
C1	0	0	1	0
C2	0	0	12	8
C3	0	0	0	14

Tabela: Matriz de Confusão para o kNN com CV

Testes no Orange

- ▶ Normais
- ▶ Precision
- ▶ Recall
- ▶ F1

- ▶ Aprofundamento no Bayer
- ▶ Uso de novos preditores

Obrigado