Resumo da atividade da disciplina exploração e mineração de dados (Técnicas)

Randerson Douglas R. Santos

Universidade Federal de Alagoas (UFAL)
Instituto de Computação (IC)

rdrs@ic.ufal.br

1. Dataset Iris

   Inclui três espécies de íris com 50 amostras cada, bem como algumas propriedades sobre cada flor. Uma espécie de flor é linearmente separável das outras duas, mas as outras duas não são linearmente separáveis uma da outra.

   Variáveis:

   - Id
   - SepalLengthCm
   - SepalWidthCm
   - PetalLengthCm
   - PetalWidthCm
   - Species

   Link dataset: Iris

   a. Técnica KNN com R

      Link que contém a implementação do teste: Dataset Iris - Técnica KNN.

      Resultado obtido através de uma tabela de contingência:

Total Observations in Table: 40				
	iris_pred			
iris.testLabels	setosa	versicolor	virginica	Row Total
setosa	12 1.000 1.000 0.300	0 0.000 0.000 0.000	0 0.000 0.000 0.000	12 0.300
versicolor	0 0.000 0.000	12 1.000 1.000	0 0.000 0.000	12 0.300

	0.000	0.300	0.000	
virginica	0 0.000 0.000 0.000	1 0.062 0.077 0.025	15 0.938 1.000 0.375	16 0.400
Column Total	12 0.300	13 0.325	15 0.375	40

Da tabela acima, lemos que nosso modelo fez apenas uma previsão errada: para uma observação de espécie virginica, o modelo prediziu que era de espécie versicolor.

Para todas as outras 39 observações do conjunto teste, o nosso modelo fez a previsão correta. Poderíamos concluir que a performance do modelo é bastante satisfatória.

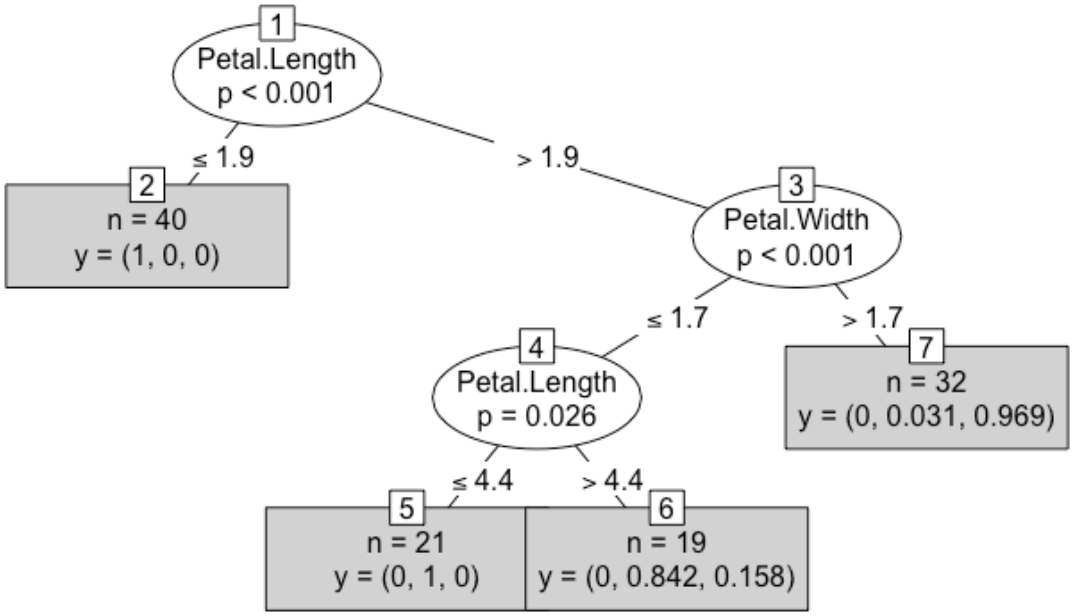
b. Técnica Naïve Bayes com R

Link que contém a implementação do teste: Dataset Iris - Técnica NAIVE BAYES.

Total Observations in Table: 40				
	actual			
Predicted	Iris-setosa	Iris-versicolor	Iris-virginica	Row Total
Iris-setosa	10 1.000	0 0.000	0 0.000	10
Iris-versicolor	0 0.000	12 1.000	0 0.000	12
Iris-virginica	0 0.000	0 0.000	18 1.000	18
Column Total	10 0.250	12 0.300	18 0.450	40

c. Técnica Decision Tree com R

Link que contém a implementação do teste: Dataset Iris – Técnica DECISION TREE.



Total Observations in Table: 112				
	actual			
Predicted	Iris-setosa	Iris-versicolor	Iris-virginica	Row Total
Iris-setosa	40 1.000	0 0.000	0 0.000	40
Iris-versicolor	0 0.000	37 0.974	3 0.088	40
Iris-virginica	0 0.000	1 0.026	31 0.912	32
Column Total	40 0.357	38 0.339	34 0.304	112

d. Conclusão

Pode-se ver na aplicação da técnica KNN, avaliando no que diz respeito a variável precision não houve algo ruim, onde somente um caso saiu como erro, mas houve uma boa significância com o valor de predição positiva de 1 ou 100% para os eventos setosa e vernicolor, onde para o ultimo evento virginica houve uma diminuição no valor da predição positiva, sendo um valor de 0,9375 ou 93,75%.

Percebesse que a taxa de sensibilidade (recall) segue o mesmo principio da medida precision, onde somente em um dos casos (virginica – caso 3) o valor foi abaixo de 100%, ficando em 0,923 ou 92,3%.

No tocante ao F1 Score, foi feita a divisão da multiplicação da precision e o recall pela soma e multiplicando por 2 no final, obtendo assim um valor de 0,93 ou 93%. Dessa forma percebemos que há uma acurácia de 100% aproveitamento.

Ao aplicarmos a técnica naive bayes, foi visto que todas as métricas de efetividade foram satisfatórias, sendo assim um aproveitamento de 100%.

Analisando o método de decision tree, foi visto que houveram 2 erros, durante a segunda e terceira verificação (Iris-versicolor e Iris-virginica). De modo que o valor da predição positiva (precision) é 0,925 ou 92,5% (no segundo caso, versicolor) e a taxa de sensibilidade (recall) igual à 0,973 ou 97,3%.

2. Dataset Titanic

   Contém informações de pessoas da tripulação que estavam a bordo no navio, informando sexo, idade, embarcado, classe e sobrevivência.

   Link dataset: Titanic

   a. Técnica KNN com R

Total Observations in Table: 178			
	KNN_pred		
t(validation_set_clean[, 1])	0	1	Row Total
0	95 0.864 0.819 0.534	15 0.136 0.242 0.084	110 0.618
1	21 0.309 0.181 0.118	47 0.691 0.758 0.264	68 0.382
Column Total	116 0.652	62 0.348	178

Link que contém a implementação do teste: Dataset Titanic - Técnica KNN.

O valor da predição positiva é de 79,78%, o valor da predição de para a sobrevivência é de 69,12% enquanto a taxa de predição para não sobrevivência é de 86,36%.

b. Técnica Naïve Bayes com R

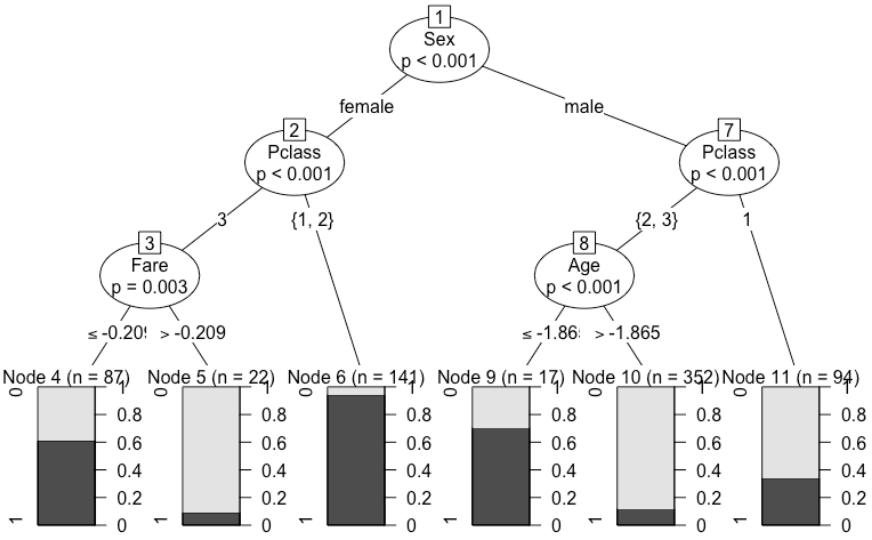
Link que contém a implementação do teste: Dataset Titanic – Técnica NAIVES BAYES

Total Observations in Table: 178			
	NBayes_pred		
t(validation_set_clean[, 1])	0	1	Row Total
0	97 0.882 0.808 0.545	13 0.118 0.224 0.073	110 0.618
1	23 0.338 0.192 0.129	45 0.662 0.776 0.253	68 0.382
Column Total	120 0.674	58 0.326	178

O valor da predição positiva é de 79,78%, o valor da predição de para a sobrevivência é de 66,18% enquanto a taxa de predição para não sobrevivência é de 88,18%.

c. Técnica Decision Tree

Link que contém a implementação do teste: Dataset Titanic - Técnica Decision Tree



Total Observations in Table: 178			
	Dtree_pred		
t(validation_set_clean[, 1])	0	1	Row Total
0	95 0.864 0.792 0.534	15 0.136 0.259 0.084	110 0.618
1	25 0.368 0.208 0.140	43 0.632 0.741 0.242	68 0.382
Column Total	120 0.674	58 0.326	178

O valor da predição positiva é de 77,53%, o valor da predição de para a sobrevivência é de 63,24% enquanto a taxa de predição para não sobrevivência é de 86,36%.