Universidade Federal de Alagoas - UFAL
Instituto de Computação - IC
Programa de Pós-graduação em Informática - PPGI
Data analysis

Técnicas de classificação

Docente: Baldoino Fonseca dos Santos Neto
Discente: Williams Lourenço de Alcantara

• Nearest Neighbor (k-NN)

Dataset: Philadelphia Real Estate
Link: https://www.kaggle.com/harry007/philly-real-estate-data-set-sample/data
Descrição: Conjunto de imóveis da Philadelphia mapeados por localidades e que possuem os índices de crimes e de qualidade da educação em cada localidade

Objetivo: Classificar se um imóvel é violento ou não
Variável utilizada: Violent.Crime.Rate (Taxa de violência)

- Resultados
  Total de observações: 210

Matriz de confusão

	data_test_pred			
test_labels	High	Low	Medium	Row Total
High	6	3	15	24
	0.250	0.125	0.625	0.114
	0.600	0.038	0.124	
	0.029	0.014	0.071	
Low	1	64	51	116
	0.009	0.552	0.440	0.552
	0.100	0.810	0.421	
	0.005	0.305	0.243	
Medium	3	12	55	70
	0.043	0.171	0.786	0.333
	0.300	0.152	0.455	
	0.014	0.057	0.262	
Column Total	10	79	121	210
	0.048	0.376	0.576	

High	
(TP) **6**	(FN) **18**
(FP) **4**	(TN) **119**

Medium	
(TP) **55**	(FN) **15**
(FP) **66**	(TN) **70**

Low	
(TP) **64**	(FN) **52**
(FP) **15**	(TN) **61**

Recall

$$Recall_{High} = \frac{TP}{TP+FN} = \frac{6}{6+18} = \frac{6}{24} = 0.25$$
$$Recal_{Medium} = \frac{TP}{TP+FN} = \frac{55}{55+15} = \frac{55}{70} = 0.785$$
$$Recall_{Low} = \frac{TP}{TP+FN} = \frac{64}{64+52} = \frac{64}{116} = 0.551$$

$$Recall = \frac{0.25 + 0.785 + 0.551}{3} = 0.528$$

Precision

$$Precision_{High} = \frac{TP}{TP+FP} = \frac{6}{6+4} = \frac{6}{10} = 0.6$$
$$Precision_{Medium} = \frac{TP}{TP+FP} = \frac{55}{55+66} = \frac{55}{121} = 0.454$$
$$Precision_{Low} = \frac{TP}{TP+FP} = \frac{64}{64+15} = \frac{64}{79} = 0.810$$

$$Precision = \frac{0.6 + 0.454 + 0.810}{3} = 0.621$$

F-measure (F1)

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} = 2 * \frac{0.621 * 0.528}{0.621 + 0.528} = 2 * \frac{0.327}{1.149} = 0.569$$

Informedness

$$Informedness_{High} = \frac{TP}{TP+FN} + \frac{TN}{FP+TN} - 1 = \frac{6}{6+18} + \frac{119}{4+119} - 1 = \frac{6}{24} + \frac{119}{123} - 1 = 0,217$$
$$Informedness_{Medium} = \frac{TP}{TP+FN} + \frac{TN}{FP+TN} - 1 = \frac{55}{55+15} + \frac{70}{66+70} - 1 = \frac{55}{70} + \frac{70}{136} - 1 = 0,30$$
$$Informedness_{Low} = \frac{TP}{TP+FN} + \frac{TN}{FP+TN} - 1 = \frac{64}{64+52} + \frac{61}{15+61} - 1 = \frac{64}{116} + \frac{61}{76} - 1 = 0,354$$

$$Informedness = \frac{0,217 + 0,30 + 0,354}{3} = 0,29$$

Markedness

$$Markedness_{High} = \frac{TP}{TP+FP} + \frac{TN}{FN+TN} - 1 = \frac{6}{6+4} + \frac{119}{18+119} - 1 = \frac{6}{10} + \frac{119}{137} - 1 = 0,468$$
$$Markedness_{Medium} = \frac{TP}{TP+FP} + \frac{TN}{FN+TN} - 1 = \frac{55}{55+66} + \frac{70}{15+70} - 1 = \frac{55}{121} + \frac{70}{85} - 1 = 0,278$$
$$Markedness_{Low} = \frac{TP}{TP+FP} + \frac{TN}{FN+TN} - 1 = \frac{64}{64+15} + \frac{61}{52+61} - 1 = \frac{64}{79} + \frac{61}{113} - 1 = 0,349$$

$$Markedness = \frac{0,468 + 0,278 + 0,349}{3} = 0,365$$

Discussão

Os valores das métricas são baixos, isto quer dizer que este algoritmo de classificação não obteve um bom resultado tanto para predições positivas quanto negativas, apesar de apresentar mais acertos para predições positivas.

- # Naive Bayes

  Dataset: Titanic
  Link: https://www.kaggle.com/c/titanic
  Objetivo: Predizer as integrantes do navio Titanic que sobreviveram
  Variável utilizada: Survived

- Resultados
  Total de observações: 2201

Matriz de confusão

(TP) 1364	(FN) 362
(FP) 126	(TN) 349

Recall

$$Recall = \frac{TP}{TP+FN} = \frac{1364}{1364+362} = \frac{1364}{1726} = 0.790$$

Precision

$$Precision_{High} = \frac{TP}{TP+FP} = \frac{1364}{1364+126} = \frac{1364}{1490} = 0,915$$

F-measure (F1)

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} = 2 * \frac{0.915 * 0.790}{0.915 + 0.790} = 2 * \frac{0,722}{1,705} = 0,846$$

Informedness

$$Informedness = \frac{TP}{TP+FN} + \frac{TN}{FP+TN} - 1 = \frac{1364}{1364+362} + \frac{349}{126+349} - 1 = \frac{1364}{1726} + \frac{349}{475} - 1 = 0,525 = 52,5\%$$

Markedness

$$Markedness = \frac{TP}{TP+FP} + \frac{TN}{FN+TN} - 1 = \frac{1364}{1364+126} + \frac{349}{362+349} - 1 = \frac{1364}{1490} + \frac{349}{711} - 1 = 0,406 = 40,6\%$$

Discussão

Considerando o conjunto de predições positivas (sobreviventes), a métrica *Precision (0,79)* indica que possui um bom índice de acerto quando afirmar que um indivíduo *sobreviveu*, e a *Recall* (0,915) indica que possui um bom índice de percepção de sobreviventes. Ao analisar as estas duas métricas juntamente, através de F-measure, o índice mantém-se alto (0,846). Isto quer dizer que uma boa predição quando indica valores positivos (sobreviventes).

Considerando o conjunto de predições positivas (sobreviventes) e negativas (não sobreviventes), Informedness indica que possui uma baixa probabilidade (52,5%) de fazer uma predição correta (tanto para sobrevivente quanto para não sobrevivente). Enquanto que Markedness indica uma baixa probabilidade (40,6%) de detectar indivíduos sobreviventes.

● Decision Tree

Dataset: Breast Cancer Wisconsin (Diagnostic)
Link: https://www.kaggle.com/arpisinanyan/naive-bayes-on-diagnosis/data

Objetivo: Prever o tipo do câncer (Maligno ou Benigno)

- Resultados
  Total de observações: 177

Matriz de confusão

(TP) 101	(FN) 1
(FP) 6	(TN) 69

Recall

$$Recall = \frac{TP}{TP+FN} = \frac{101}{101+1} = \frac{101}{102} = 0.99$$

Precision

$$Precision = \frac{TP}{TP+FP} = \frac{101}{101+6} = \frac{101}{107} = 0.943$$

F-measure (F1)

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} = 2 * \frac{0.943 * 0.99}{0.943 + 0.99} = 2 * \frac{0.933}{1.933} = 0,965$$

Informedness

$$Informedness = \frac{TP}{TP+FN} + \frac{TN}{FP+TN} - 1 = \frac{101}{101+1} + \frac{69}{6+69} - 1 = \frac{101}{102} + \frac{69}{75} - 1 = 0,91$$

Markedness

$$Markedness = \frac{TP}{TP+FP} + \frac{TN}{FN+TN} - 1 = \frac{101}{101+6} + \frac{69}{1+69} - 1 = \frac{101}{107} + \frac{69}{70} - 1 = 0,929$$

Discussão

Os valores das métricas *Recall*, *Precision, F-measure, Markedness e Informedness* são altos, superiores a 0,91, isto quer dizer que este algoritmo obteve um bom resultado para predições positivas e negativas, além de possuir uma alta capacidade de detecção de indivíduos que apresentam o câncer.

● Linear Regression

Dataset: Titanic

Link: https://www.kaggle.com/c/titanic
Objetivo: Predizer as integrantes do navio Titanic que sobreviveram
Variável utilizada: Survived

- Resultados
  Total de observações: 2201

Matriz de confusão

(TP) 44	(FN) 94
(FP) 6	(TN) 22

Recall

$$Recall = \frac{TP}{TP+FN} = \frac{44}{44+94} = \frac{44}{138} = 0.318$$

Precision

$$Precision_{High} = \frac{TP}{TP+FP} = \frac{44}{44+6} = \frac{44}{50} = 0.888$$

F-measure (F1)

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} = 2 * \frac{0.888 * 0.318}{0.888 + 0.318} = 2 * \frac{0,282}{1,206} = 0.47$$

Informedness

$$Informedness = \frac{TP}{TP+FN} + \frac{TN}{FP+TN} - 1 = \frac{44}{44+94} + \frac{22}{6+22} - 1 = \frac{44}{138} + \frac{22}{28} - 1 = 0.103 = 10,3\%$$

Markedness

$$Markedness = \frac{TP}{TP+FP} + \frac{TN}{FN+TN} - 1 = \frac{44}{44+6} + \frac{22}{94+22} - 1 = \frac{44}{50} + \frac{22}{116} - 1 = 0.069 = 6,9\%$$

Discussão

Considerando o conjunto de predições positivas (sobreviventes), a métrica *Precision (0,888)* indica que o modelo da regressão linear (M1) possui um bom índice de acerto quando afirmar que um indivíduo *sobreviveu*, no entanto a métrica *Recall* (0,318) indica que M1 não possui um bom índice de percepção de sobreviventes. No entanto, ao comparar estas duas métricas, através de F-measure, o índice é baixo. Isto indica que este modelo não está adequado.

Considerando o conjunto de predições positivas (sobreviventes) e negativas (não sobreviventes), Informedness indica que M1 possui uma baixa probabilidade (10,3%) de fazer uma predição correta (tanto para sobrevivente quanto para não sobrevivente). Enquanto que Markedness indica uma baixa probabilidade (6,9%) de detectar indivíduos sobreviventes.

● Support Vector Machine

Dataset: Breast Cancer Wisconsin (Diagnostic)

Link: https://www.kaggle.com/arpisinanyan/naive-bayes-on-diagnosis/data

Objetivo:
Variável utilizada: diagnosis

- Resultados
  Total de observações: 169

Matriz de confusão

(TP) 63	(FN) 105
(FP) 1	(TN) 0

Recall

$$Recall = \frac{TP}{TP+FN} = \frac{63}{63+105} = \frac{63}{168} = 0.375$$

Precision

$$Precision_{High} = \frac{TP}{TP+FP} = \frac{63}{63+1} = \frac{63}{64} = 0.984$$

F-measure (F1)

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} = 2 * \frac{0.984 * 0.375}{0.984 + 0.375} = 2 * \frac{0.369}{1.359} = 0.543$$

Informedness

$$Informedness = \frac{TP}{TP+FN} + \frac{TN}{FP+TN} - 1 = \frac{63}{63+105} + \frac{0}{1+0} - 1 = \frac{63}{168} + \frac{0}{1} - 1 = |-0.625| = 0.625 = 62,5\%$$

Markedness

$$Markedness = \frac{TP}{TP+FP} + \frac{TN}{FN+TN} - 1 = \frac{63}{63+1} + \frac{0}{105+0} - 1 = \frac{63}{64} + \frac{0}{105} - 1 = |-0.016| = 0.016 = 1,6\%$$

Discussão

Considerando o conjunto de predições positivas (*maligno*), a métrica *Precision (0,984)* indica que o algoritmo possui um bom índice de acerto quando afirmar que um indivíduo *possui câncer maligno*, no entanto a métrica *Recall* (0,375) indica que não possui um bom índice de percepção de câncer *benigno*. Ao comparar estas duas métricas, através de F-measure, o índice é baixo. Isto indica que este modelo não está adequado.

Considerando o conjunto de predições positivas (maligno) e negativas (benigno), Informedness indica que possui uma probabilidade de 62,5% de fazer uma predição correta (tanto para maligno quanto para benigno). Enquanto que Markedness indica uma baixa probabilidade (1,6%) de detectar câncer benigno.