PCS5024 – Classificação da base de dados Adult

Marcelo Monari Baccaro – 8989262

Este trabalho de classificação foi feito em linguagem R. Antes de aplicar os algoritmos de ajuste de hiperparâmetros de cada modelo estatístico de classificação, os dados da base *Adult* (tanto os dados de treino quanto os de teste) passaram por um pré-processamento, em que a coluna *education* foi removida pois *education_num* contém a mesma informação porém em dados numéricos ordenados, o que simplifica o tratamento dos dados categóricos da primeira coluna. Além disso, todos as colunas de dados numéricos foram escalonadas para ter média zero e desvio padrão unitário. E todos os dados categóricos (exceto o de educação como mencionado antes) foram codificados com *dummy variables* (semelhante ao *hot one*, mas uma coluna é removida), em que a coluna com valores mais frequentes foi removida. Como a função que gera variáveis *dummy* em R retorna apenas as colunas correspondentes aos dados, ocorre um desencontro entre as colunas de treino e de teste, porém isto foi resolvido com a remoção das colunas que não estivem presentes em ambos, afinal isto indica que são casos raros. Por fim, os *missing values* foram tratados com a simples remoção destes, afinal a quantidade de linhas que os tem é menor que 8% tanto nos dados de treino quanto nos de teste.

A tabela abaixo apresenta os resultados de acurácia, precisão e revocação (*recall* ou sensibilidade) para os modelos estatísticos de classificação: kNN (*k-Nearest Neighbor*) para o hiperparâmetro *k* de 1 até 30; Regressão Logística; *Naive Bayes*; Linear (*Least Squares Linear Regression*); Regressão de Poisson; Árvore de Regressão; Floresta Aleatória com $\sqrt{p}$ (em que p = número de variáveis de entrada) variáveis aleatoriamente amostradas como candidatas para cada separação; *Boosted Trees* com 400 árvores e profundidade $\sqrt{p}$; SVM (*Support Vector Machine*) com kernel linear e custo 2; e MLP (*Multilayer Perceptor*) com 2 camadas escondidas com 10 e 4 neurônios e função de ativação logística.

Modelo		Acurácia	Precisão	Revocação
kNN	1	0.794672512	0.862163114	0.865134777
	2	0.793543244	0.858816276	0.866293532
	3	0.818519995	0.888409371	0.873182133
	4	0.818519995	0.887528624	0.873829344
	5	0.828683406	0.897833363	0.877809352
	6	0.828683406	0.897569139	0.878004652
	7	0.833001196	0.900915977	0.880444138
	8	0.833931181	0.901708649	0.880915505
	9	0.835193304	0.902060948	0.882094565
	10	0.834728311	0.902237097	0.881431767
	11	0.834861166	0.901268275	0.882307294
	12	0.834661884	0.901092126	0.882210917
	13	0.837252557	0.902589396	0.884057971
	14	0.837252557	0.902501321	0.884124245
	15	0.83771755	0.902941695	0.884326749
	16	0.83745184	0.901884798	0.884818111
	17	0.839046101	0.903382068	0.885521886
	18	0.840308224	0.903558217	0.886843015
	19	0.840108941	0.90417474	0.886145878
	20	0.838116115	0.90267747	0.884984026
	21	0.840308224	0.904438964	0.886175354
	22	0.839245383	0.902237097	0.886619353

	23	0.839909659	0.903646292	0.886316517
	24	0.838713963	0.903910516	0.884741379
	25	0.839046101	0.903822441	0.885189338
	26	0.837385413	0.902413246	0.884343173
	27	0.838647536	0.904967412	0.883870968
	28	0.838780391	0.906024309	0.883231733
	29	0.839112528	0.905407786	0.884072927
	30	0.839112528	0.905407786	0.884072927
Regressão Logística		0.846884548957088	0.926369561388057	0.877450571452407
Naive Bayes		0.582635844293875	0.469702307556808	0.953172475424486
Linear (LSLR)		0.838248970373323	0.938083494803593	0.860130824517484
Regressão de Poisson		0.829746246844692	0.943015677294346	0.848213578388656
Árvore de Regressão		0.835126876577654	0.93059714638013	0.861827079934747
Floresta Aleatória		0.856915105619769	0.940285361986965	0.878538512179065
Boosted Trees		0.852265178690049	0.941342258234983	0.872774783602809
SVM		0.781187724192905	0.978950149726968	0.784403669724771
MLP		0.851534475886808	0.931742117315484	0.878727469058892

O pior modelo no geral é o Naive Bayes, embora ela tenha a melhor revocação, o que indica que o número de falsos negativos foi o menor. Os três melhores modelos são: Floresta Aleatória, Boosted Trees e MLP (redes neurais). Entretanto, os dois primeiros demoram bem mais para treinar do que a rede neural e não apresentam um resultado tão melhor. Os três modelos mais rápidos de treinar são: Regressão Logística, Linear e Árvore de Regressão; em que a Regressão Logística tem as melhores medidas no geral.

Como mostra a figura abaixo (além da tabela acima), o técnica de *grid search* para o hiperparâmetro *k* do modelo kNN mostra que, no domínio de 1 a 30, o valor ótimo é de $k = 18$, pois é o menor parâmetro em que a maior acurácia é obtida, enquanto que as demais medidas (precisão e revocação) estão próximas dos respectivos valores de saturação.