Breast Cancer Wisconsin (Diagnostic) Data Set

Predict whether the cancer is benign or malignant

Henrique Aparecido Laureano
est171 - Aprendizado de Máquina
henriquelaureano@outlook.com

Departamento de Estatística
Universidade Federal de Minas Gerais

Dados

Dataset fornecido pela UCI Machine Learning e hospedado no **Kaggle** [1]. Temos 569 pacientes, 357 (63%) com câncer benigno e 212 (37%) com câncer maligno.

Através da imagem digitalizada de um tipo de biópsia de mama não-cirúrgica chamada de FNA (*fine needle aspiration*), aspiração por agulha fina, dez características são mensuradas para cadanúcleo celular:

1. raio (média das distâncias do centro aos pontos no perímetro
2. textura (desvio padrão dos valores da escala de cinza)
3. perímetro
4. área
5. suavidade (variação local dos comprimentos do raio)
6. compacidade (perímetro^2 / área - 1)
7. concavidade (gravidade das porções côncavas do contorno)
8. pontos côncavos (número de porções côncavas do contorno)
9. simetria
10. dimensão fractal (aproximação *coastline* - 1)

Para cada imagem, e consequentemente para cada paciente, a média, erro padrão e "pior" ou maior (média dos três maiores valores) valor dessas características são computadas, resultando em 30 variáveis. A base de dados foi dividida em treino (454 pacientes, 80%) e teste (115 pacientes, 20%, 72 benignos e 43 malignos).

Objetivos

Com o objetivo de verificar com quais dessas medidas (médias, erros padrão ou maiores) conseguimos melhor classificar o câncer em benigno ou maligno, quatorze algoritmos de classificação foram utilizados, com cada um sendo utilizado três vezes, uma vez considerando apenas as variáveis correspondentes as médias, outra com os erros padrões e outra com os "piores" (*worst*) valores.

Resultados

Na Tabela 1 temos o risco estimado e a área abaixo da curva, AUC, da curva ROC para os quatorze classificadores. Para o algoritmo KNN foi obtido e utilizado um número de vizinhos igual a 5, 18 e 27, respectivamente.

De maneira geral, o classificador com melhor desempenho foi a regressão logística e piores resultados foram obtidos utilizando as variáveis de erro padrão.

	Mean		Standard error		Worst	
	Risco	AUC	Risco	AUC	Risco	AUC
Regressão logística	0.0435	0.9948	0.0609	0.9855	**0.0087**	**0.9994**
Regressão linear	0.0522	0.9939	0.0869	0.9609	**0.0174**	**0.9958**
Naive Bayes	**0.0783**	0.9094	0.1043	0.8651	**0.0783**	**0.9141**
Discriminante linear	**0.0435**	**0.9465**	0.1478	0.8117	0.0522	0.9349
Discriminante quadrático	0.0696	0.921	0.113	0.8723	**0.0435**	**0.9559**
Discriminante regularizado	**0.1043**	**0.8698**	0.113	0.8488	**0.1043**	0.8651
KNN	0.0696	0.9448	0.0609	0.9629	**0.0522**	**0.9905**
SVM linear	0.0435	0.9465	0.087	0.9025	**0.0261**	**0.9698**
SVM polinomial	**0.0435**	**0.9512**	0.1217	0.8513	0.0522	0.9349
SVM radial	**0.0348**	**0.9582**	0.113	0.8676	0.0435	0.9465
SVM sigmoidal	0.0435	0.9465	0.1043	0.8698	**0.0348**	**0.9535**
Árvore de classificação	**0.0522**	**0.9536**	0.0783	0.9141	0.087	0.9165
Bagging	**0.0348**	**0.9675**	0.0609	0.9373	0.0609	0.9443
Boosting	**0.0261**	**0.9792**	0.0783	0.9141	0.0522	0.9443

Tabela 1: Riscos e AUC dos algoritmos ajustados a base de dados

Na Tabela 2 são apresentados os erros e acertos de classificação dos melhores algoritmos para cada grupo de variáveis.

	Boosting		Regressão logística			
	Mean		Standard error		Worst	
	Benigno	Maligno	Benigno	Maligno	Benigno	Maligno
Benigno	69	3	67	5	69	3
Maligno	0	43	2	41	0	43

Tabela 2: Matrizes de confusão para os melhores classificadores ajustados a base de dados

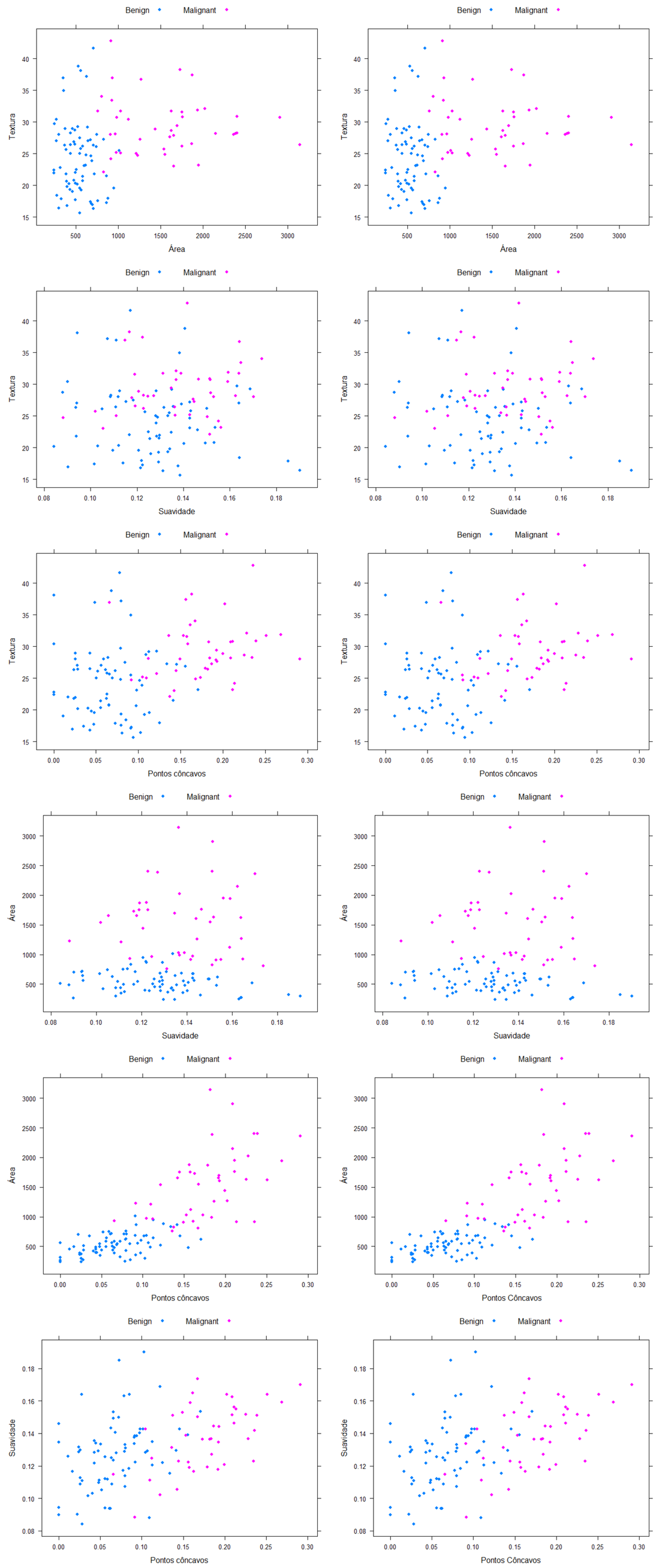

Figura 1: Gráficos de dispersão das variáveis significativas na regressão logística considerando apenas as variáveis correspondentes aos maiores ou piores (*worst*) valores das características. Coluna da esquerda: Diagnóstico (câncer benigno ou maligno) observado; Coluna da direita: Diagnóstico predito

Referências

[1] https://www.kaggle.com/uciml/breast-cancer-wisconsin-data