

UFPA PPGCC: Aprendizado de Máquina

Lista de Exercício Final - Valor 10 pts

1) [2.0 pts] Use os dados Breast Cancer Wisconsin (Diagnostic) Data Set do UCI Machine Learning Repository. Use validação cruzada para avaliar qual dos algoritmos tem maior acurácia nos dados:

- SVM Linear
- SVM RBF

Decida que tipo de padronização (normalização) dos dados você usará para cada algoritmo (ou nenhuma). Justifique.

Em princípio, fiz um escalado dos dados, normalizando eles entre o intervalo (0,1), dado que assim o *accuracy* do modelo é incrementado. A seguinte fórmula deixa ver como foi a normalização, dado que x é um vector de características:

$$x = \frac{x - \min(x)}{\max(x) - \min(x)}$$

	2	3	4	5	6	7	8	9	10	11	...
0	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	0.2419	0.07871	...
1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	...
2	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	0.2069	0.05999	...
3	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	0.2597	0.09744	...
4	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	0.1809	0.05883	...

5 rows × 30 columns



	2	3	4	5	6	7	8	9	10	11	...
0	0.521037	0.022658	0.545989	0.363733	0.593753	0.792037	0.703140	0.731113	0.686364	0.605518	...
1	0.643144	0.272574	0.615783	0.501591	0.289880	0.181768	0.203608	0.348757	0.379798	0.141323	...
2	0.601496	0.390260	0.595743	0.449417	0.514309	0.431017	0.462512	0.635686	0.509596	0.211247	...
3	0.210090	0.360839	0.233501	0.102906	0.811321	0.811361	0.565604	0.522863	0.776263	1.000000	...
4	0.629893	0.156578	0.630986	0.489290	0.430351	0.347893	0.463918	0.518390	0.378283	0.186816	...

5 rows × 30 columns

A continuação, fiz uma validação cruzada, com 10 folds para o *SVM Linear* e o *SVM RBF*, as figuras 1 e 2 mostram os resultados por separado e logo a figura 3 mostra a comparação entre deles.

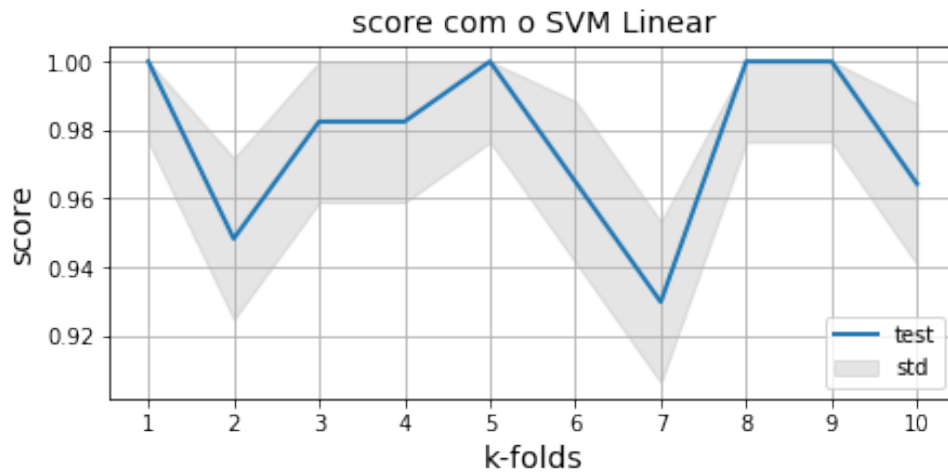


Figura 1: *Accuracy* do modelo *SVM Linear* com sua desviação estándar para cada k-fold.

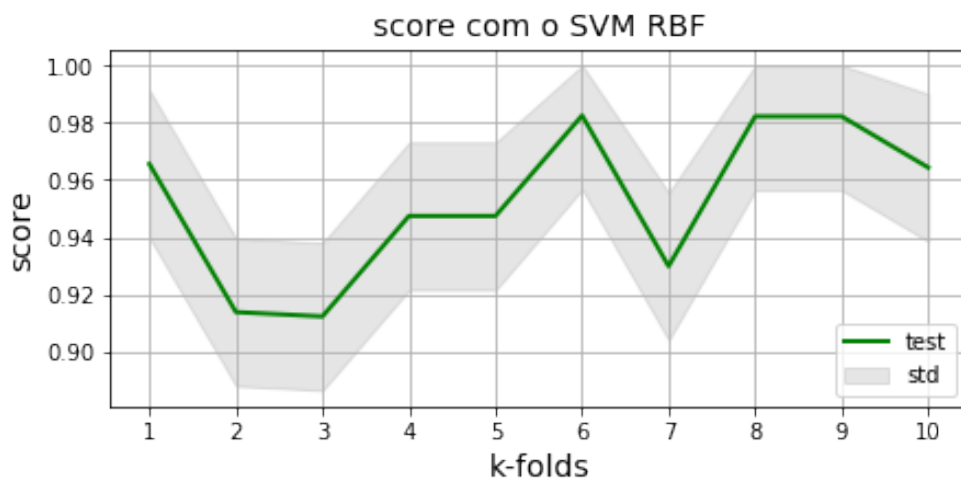


Figura 2: *Accuracy* do modelo *SVM RBF* com sua desviação estándar para cada k-fold.

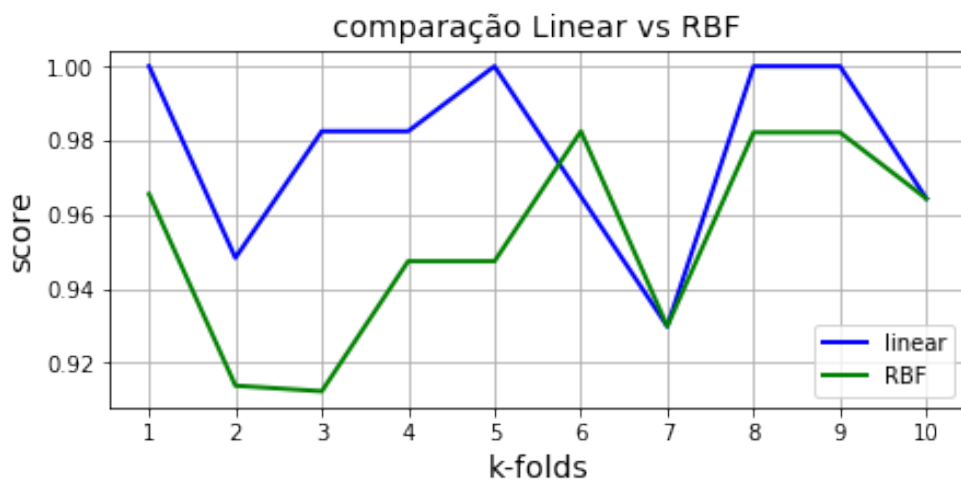


Figura 3: Comparação do *accuracy* dos modelos.

Ao final se pode ver que o *SVM Linear* tem um melhor rendimento que o *SVM RBF* já que ele foi igual o melhor em cada k fold, o *Linear* apresenta um score promedio de 97.722% e o *RBF* apresenta um score de 95.271%.

2) [2.0 pts] Implemente em uma linguagem de programação de sua escolha uma Rede Neural Artificial Multilayer Perceptron treinada com o algoritmo backpropagation que resolva o problema do OU EXCLUSIVO. Para validar sua implementação utilize a arquitetura apresentada na aula sobre Redes Neurais. Mostre que os resultados foram os mesmos. Avalie como ficaria a solução se considerarmos uma arquitetura com 2 neurônios na camada escondida e um neurônio na camada de saída. Lembrando que a função de ativação do neurônio deve ser sigmoide.

3) [2.0 pts] Dado o conjunto de dados abaixo:

a) aplique o método de agrupamento aglomerativo utilizando a métrica single-link e o critério de dissimilaridade distância Euclidiana.

b) aplique o algoritmo K-means utilizando distância Euclidiana considerando $K = 3$. O Algoritmo deve parar caso não apresente convergência após 5 iterações. Considere também que os centros iniciais são: cliente1 e cliente4.

c) Avalie qual melhor solução de clusterização considerando 3 grupos.

	X1	X2	X3	X4	X5
C_1	7,000	10,000	9,000	7,000	10,000
C_2	9,000	9,000	8,000	9,000	9,000
C_3	5,000	5,000	6,000	7,000	7,000
C_4	6,000	6,000	3,000	3,000	4,000
C_5	1,000	2,000	2,000	1,000	2,000
C_6	4,000	3,000	2,000	3,000	3,000
C_7	2,000	4,000	5,000	2,000	5,000