UNIVERSIDADE FEDERAL DO CEARÁ
CAMPUS SOBRAL
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA E DE COMPUTAÇÃO

Reconhecimento de Padrões

Borges, C.D.B

PCA, RBF E K-MEANS

SOBRAL

2019

QUESTÃO 1

Aplique o método Principal Component Analysis (PCA) na base de dados iris_log.dat. A seguir, aplique a rede neural ELM nos n primeiros componentes principais usando a estratégia de validação leave-one-out. Apresente as acurácias obtidas para diferentes quantidades de componentes (n) e de neurônios ocultos (Q).

A solução dessa questão encontra-se no script rp_pca_elm_loo.py. O código do script contém implementações da técnica de redução de dimensionalidade PCA e do classificador ELM. Na Tabela 1, são exibidas as acurácias obtidas com a ELM e validação leave-one-out, sem normalização e sem utilização do PCA. Os resultados da validação com PCA são mostrados na Tabela 2, para diferentes valores de n e Q.

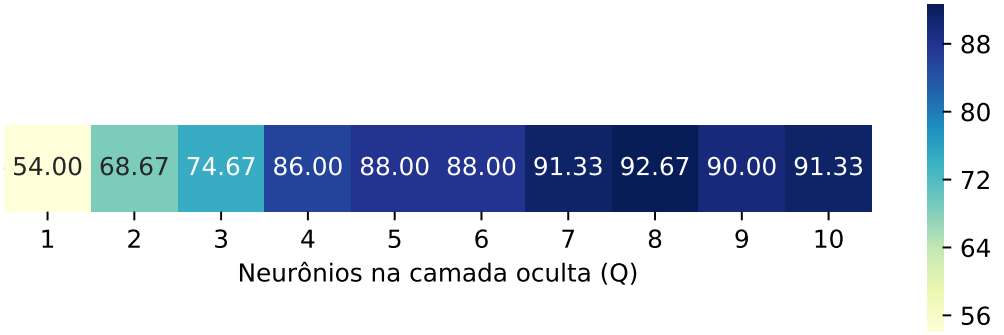

Tabela 1 – Acurácias obtidas com ELM, sem PCA e sem normalização zscore.

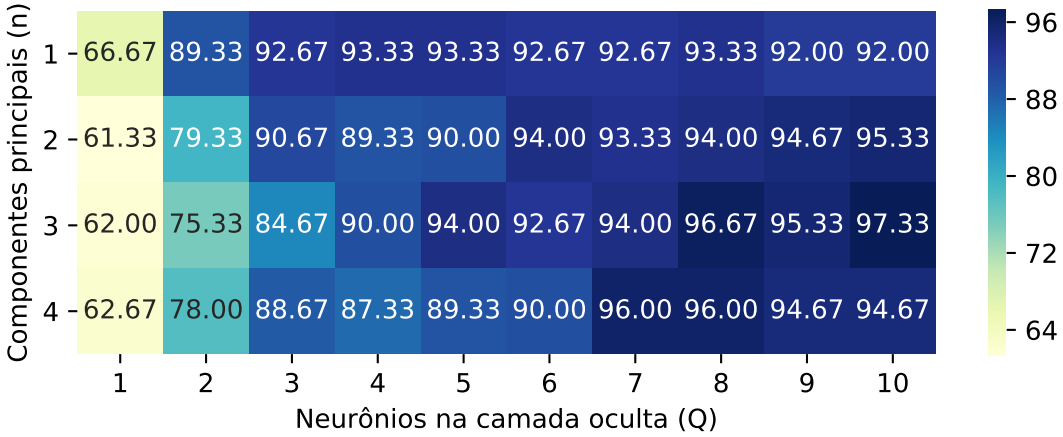

Tabela 2 – Acurácias obtidas com ELM, com PCA e sem normalização zscore.

A análise dos gráficos das Figuras 1 e 2 permite visualizar a importância de uma técnica de redução de dimensionalidade como o PCA. Na base Iris, a utilização do PCA juntamente à ELM resultou em melhorias consideráveis na taxa de acurácia, com a vantagem adicional de uso de representações de dados mais compactas.

QUESTÃO 2

Usando o conjunto de dados 2-D disponível no arquivo twomoons.dat, trace a superfície de decisão obtida com a rede neural RBF treinada com todas as amostras.

A solução dessa questão encontra-se no arquivo rp_rbf_split.py, que contém a implementação de uma RBF e sua aplicação ao dataset twomoons. Para traçar a superfície de decisão, foi definida uma malha de coordenadas $x_0 \in [-2, 9]$ e $x_1 \in [0, 8]$. Todas as coordenadas (x_0, x_1) foram utilizadas como entrada para uma RBF com 10 neurônios ocultos e um neurônio de saída. Os valores esperados na saída para as classes são -1 e 1. Dessa maneira, pode-se interpretar valores intermediários, próximos a 0, como instâncias de incerteza na decisão. Um limiar de incerteza ϵ foi definido, de tal maneira que coordenadas que produzam saídas dentro do intervalo $[-\epsilon, \epsilon]$ são denominadas regiões de incerteza, bordas das superfícies de classificação. A Figura 1 mostra superfícies de decisão com base nas regiões de incerteza (em vermelho) para $\epsilon = 0.005$ e $\epsilon = 0.2$. A acurácia obtida foi 99.7%

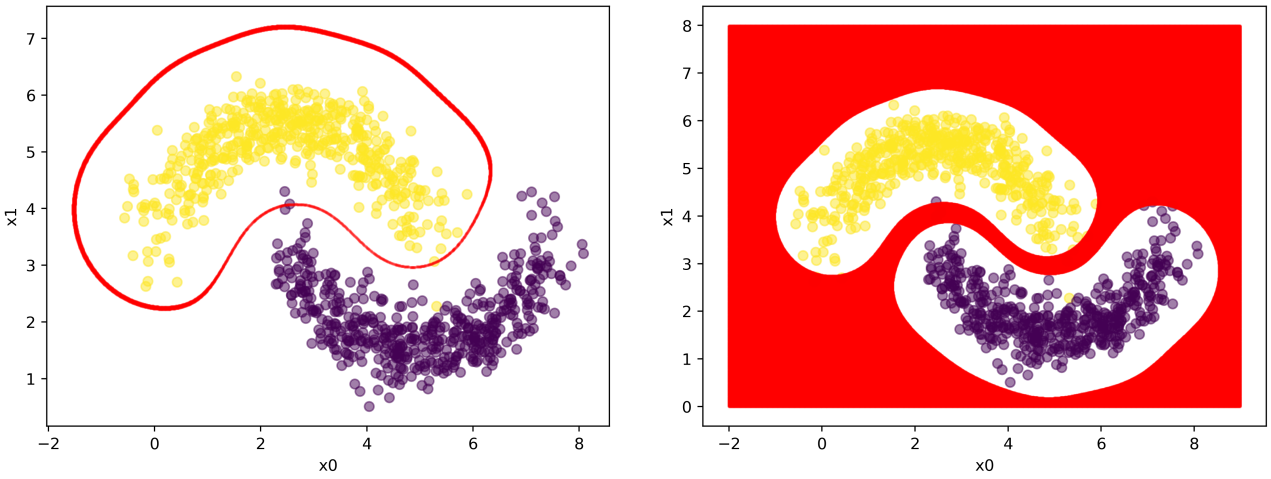

Figura 1 – Superfícies de decisão construídas com $\epsilon = 0.005$ (esquerda) e $\epsilon = 0.2$ (direita).

Vê-se na Figura 1 que o treinamento da RBF torna a ativação em sua saída bem definida nas regiões determinadas pelas amostras de treinamento. No entanto, coordenadas fora dessas regiões não apresentam ativação significante, produzindo regiões de incerteza, fato que pode ser visto claramente na superfície de decisão obtida com $\epsilon = 0.2$.

QUESTÃO 3

Efetue o agrupamento (clustering) da base iris_log.dat por meio do método K-means com diferentes valores de K e apresente as larguras de silhueta correspondentes.

A solução dessa questão encontra-se no arquivo rp_kmeans.py, que contém uma implementação do K-means e do cálculo de largura média de silhueta, um método para avaliação de qualidade de agrupamentos. Para visualizar o comportamento das larguras de silhueta S em relação à quantidade de clusters K, foram geradas a Tabela 3 e a Figura 2. A Tabela 3 mostra os valores de S obtidos para agrupamentos usando K entre 2 e 10. Pode-se notar que os maiores valores de S pertencem a agrupamentos com 2, 3 e 4 clusters. Uma observação interessante, quando se considera que o valor real de classes presentes na base Iris é 3. No entanto, é notável, através da Figura 2, que o valor de S, por si só, não pode ser usado como uma medida confiável da quantidade de classes presentes nos dados. A partir de um determinado K $\approx$ 30, S apresenta uma tendência crescente até atingir o valor máximo de 1.0, quando K é igual ao número de amostras do dataset.

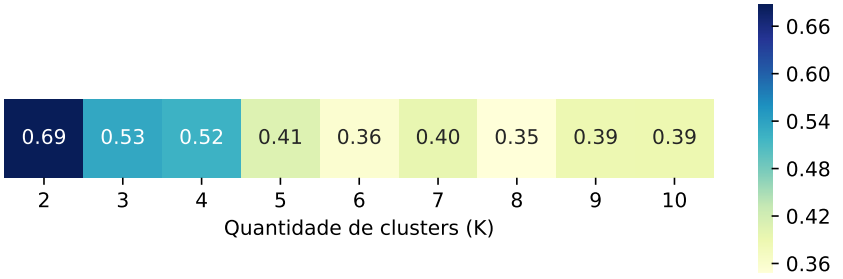

Tabela 3 – Larguras médias de silhueta S para valores baixos de K.

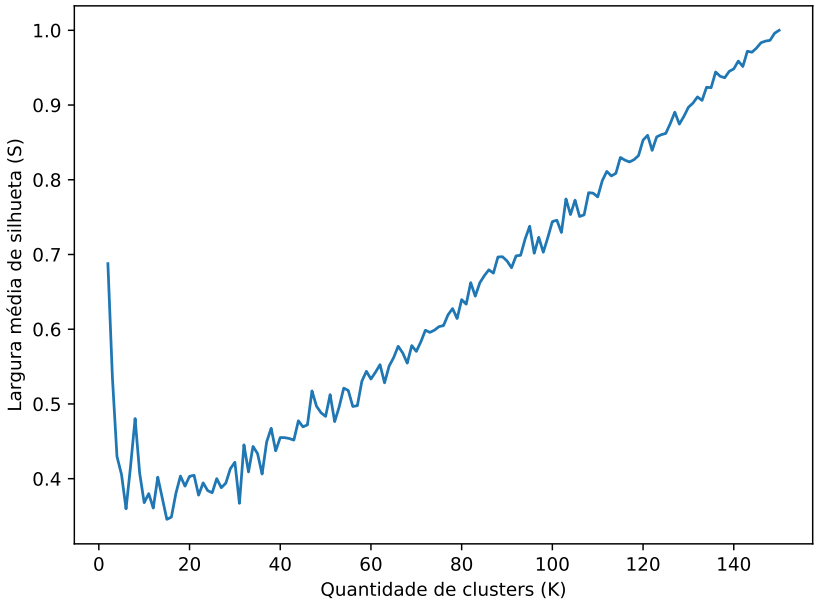

Figura 2 – Larguras médias de silhueta S para valores de K até 150.