Atividade Prática 4

Matheus Brito Faria - 2017074386

I. Breast Cancer Wisconsin (Diagnostic) Data Set

Para esta base de dados foram disponibilizados uma tabela com 32 colunas e 569 amostras, ao analisar esses dados foi removido a coluna de identificação e obtivemos 30 features para serem utilizadas como entrada e um vector com os rótulos de cada amostra.

Os rótulos foram transformados em dados numéricos de maneira binária, usando uns e zeros, isso foi uma forma de contornar a forma como o ANFIS funciona no MatLab onde ele é usado para solucionar problemas de regressão. Dessa forma ao chegar em um resultado numérico, esse mesmo é truncado em zero para valores menores que 0.5 e um caso contrário.

Cada coluna foi normalizada entre zero e um para manter certa integridade dos dados e para que possíveis amplitudes elevadas não venham a enviesar o modelo.

Foi utilizado a validação cruzada para avaliar o quão bem o modelo pode se ajustar as dados, para isso foram separados cinco folds com os dados dispostos de forma aleatório e posteriormente eles foram divididos em 70% para treino e 30% para teste.

Depois desse passo começa o treinamento de cada fold. Os dados de treino são submetidos ao Fuzzy C-Means para os parâmetros da gaussina de cada regra seja iniciada baseada nele. Posteriormente o modelo é treinado visando o valor ótimo do erro quadrático médio e o comportamento do erro em cada fold pode ser observado nas imagens 1, 2, 3, 4 e 5.

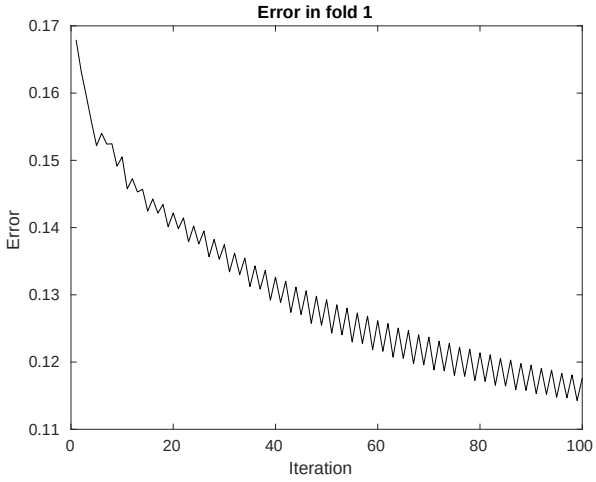

Figura 1. Erro durante o treinamento do dataset Breast Cancer Wisconsin (Diagnostic) no fold 1.

Um ponto importante de ser realçado é que o erro mostrado nos folds é o erro médio quadrático e não o erro real da classificação. Por isso, posteriormente foi calculado para os dados de teste a acurácia juntamente de sua média e de seu desvio padrão para análise como pode ser visto na tabela I.

Como pode ser observado todos os folds resultaram em acurácias bem altas e bem próximas entre si, o que nos leva a concluir que os modelos generalizaram bem para os dados existentes.

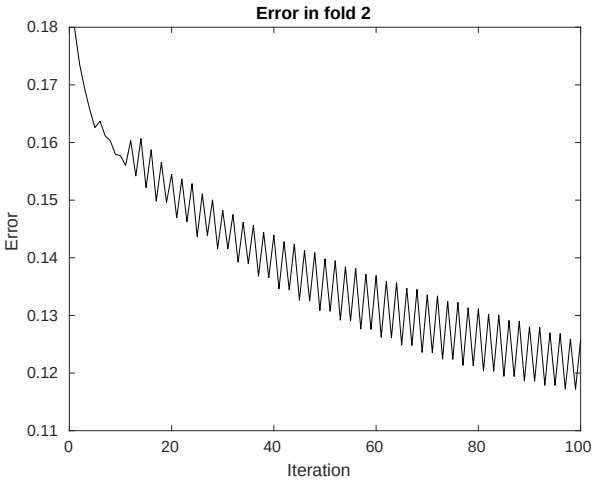

Figura 2. Erro durante o treinamento do dataset Breast Cancer Wisconsin (Diagnostic) no fold 2.

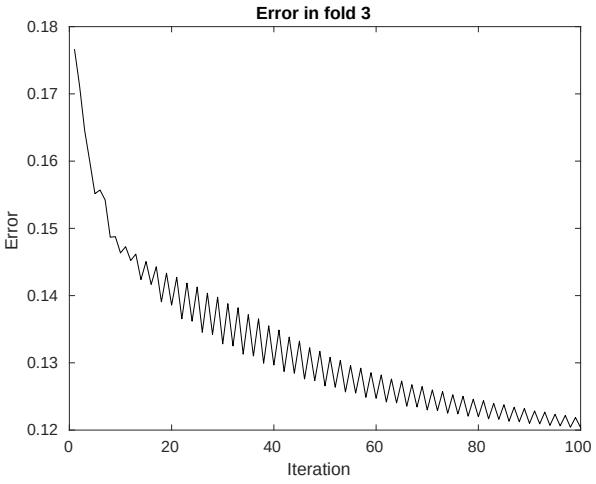

Figura 3. Erro durante o treinamento do dataset Breast Cancer Wisconsin (Diagnostic) no fold 3.

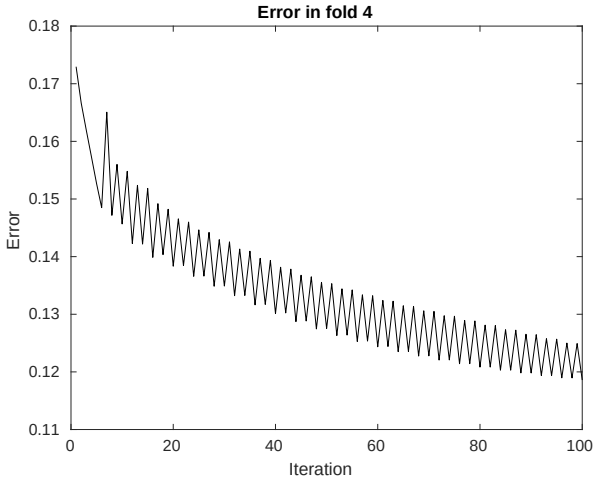

Figura 4. Erro durante o treinamento do dataset Breast Cancer Wisconsin (Diagnostic) no fold 4.

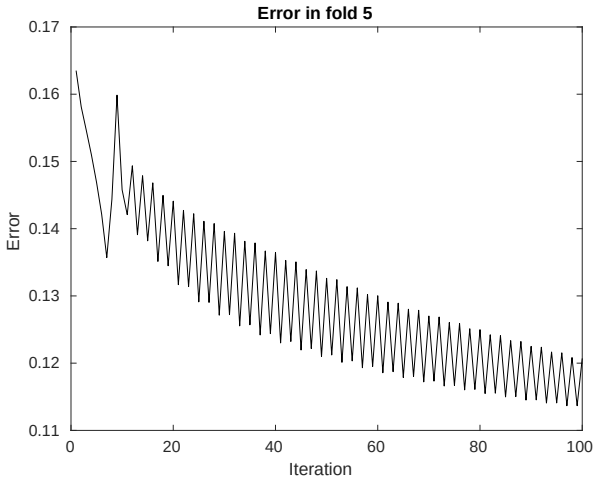

Figura 5. Erro durante o treinamento do dataset Breast Cancer Wisconsin (Diagnostic) no fold 5.

II. Iris Species Data Set

Esta base de dados foi tratada exatamente como a anterior, porém nesse caso quatro features e 150 amostras e apesar de existirem três rótulos, os rótulos foram analisados de maneira a distinguir a espécie “Iris-setosa"das outras.

Efetuando os mesmo processos do exercício anterior foi possível observar o comportamento dos erros em cada fold nas imagens 6, 7, 8, 9 e 10. Nelas apesar do comportamento oscilatório e inconstante, mostra um erro já quase saturado, porém o modelo ainda não atingiu o overfitting visto que foi possível classificar corretamente todas as amostras de teste, obtendo 100% de aproveitamento. Vale ressaltar que o modelo começa a ser treinado com os valores definidos pelo Fuzzy C-Means, o que permite que o treinamento seja mais simples.

Tabela I
Valores das acurácias de teste nos cinco folds junto da média e do desvio padrão.

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Média	Desvio Padrão
0.9706	**0.9765**	0.9647	**0.9765**	0.9706	0.9718	0.0049

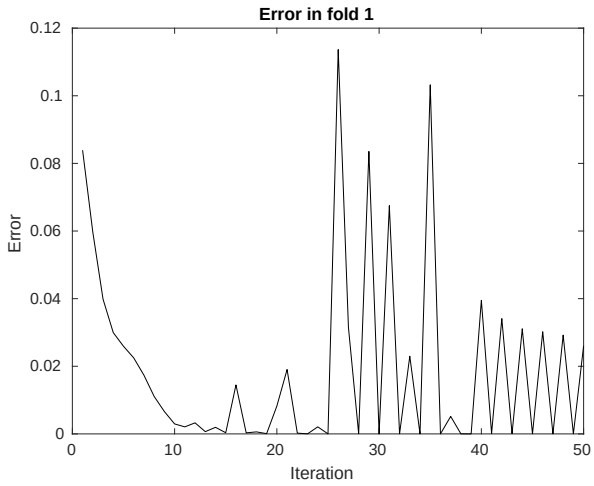

Figura 6. Erro durante o treinamento do dataset Iris Species no fold 1.

III. Erro de treino e teste e número de regras

O número de regras definidas inicialmente durante um treinamento é uma variável que pode causar alguns problemas. Isso ocorre pois como o objetivo do treinamento é fazer a minimização de alguma função objetivo, quanto mais graus de liberdade existir mais próximos dos dado o modelo vai se tornar, e isso acarreta em alguns problemas visto que os dados podem possuir ruido que e o modelo pode tentar modelar esse ruido, ou então o modelo se aperfeiçoar tanto para um certo conjunto de dados que se torna impossível classificar outra coisa a não ser aquele mesmo conjunto.

O número de regras nesse caso afeta diretamente nos graus de liberdade, ou seja se foram definidas n funções de pertinência para cada entrada, o modelo vai passar a se superajustar a esses dados, perdendo sua generalização para solução de algum problema.

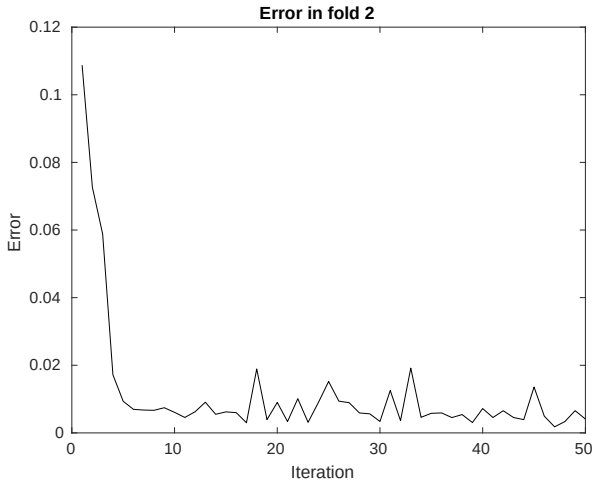

Figura 7. Erro durante o treinamento do dataset Iris Species no fold 2.

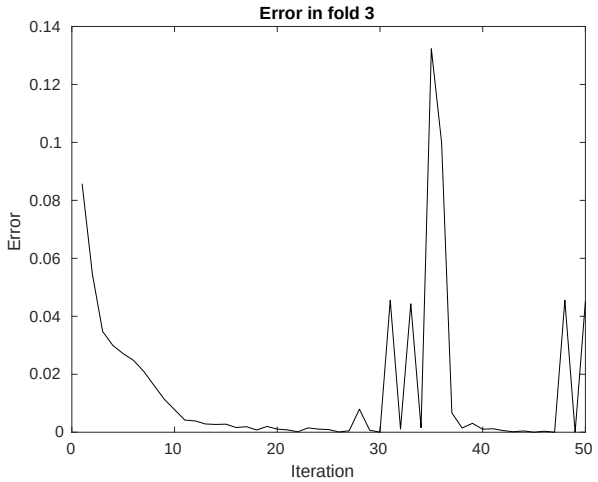

Figura 8. Erro durante o treinamento do dataset Iris Species no fold 3.

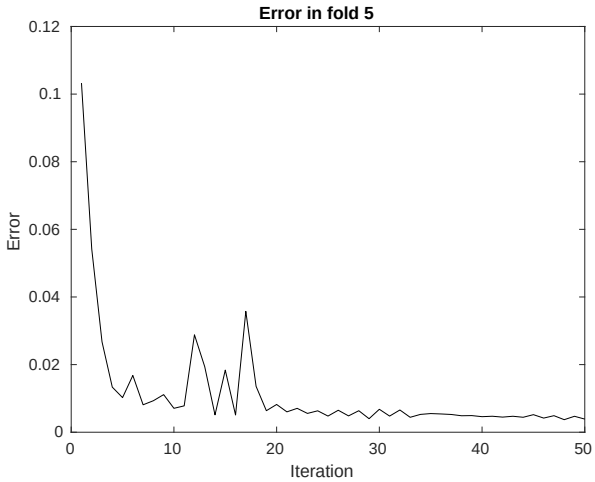

Figura 10. Erro durante o treinamento do dataset Iris Species no fold 5.

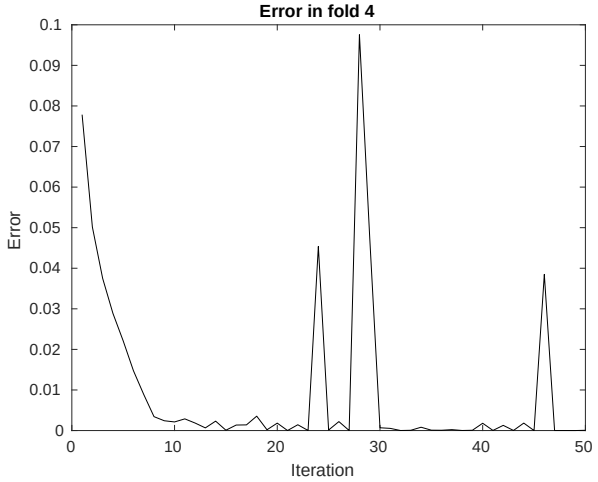

Figura 9. Erro durante o treinamento do dataset Iris Species no fold 4.