UNIVERSIDADE FEDERAL DE MINAS GERAIS - UFMG
ELE075 - Sistemas Nebulosos - 2018/2

Professor: Cristiano Leite de Castro

Trabalho Computacional 3

Aluno: Rafael Carneiro de Castro **Matrícula:** 2013030210
Curso: Engenharia de Sistemas

1 - Introdução:

Este trabalho consiste da resolução de alguns problemas por sistemas nebulosos adaptativos (ANFIS). O primeiro deles é a aproximação da função *seno* entre 0 e 2π, utilizando um sistema do tipo *Takagi-Sugeno* de ordem 1, com um conjunto de 3 regras nebulosas. Em seguida será projetado um classificador de padrões, que será testado com três bases: a primeira delas é uma base 2D disponibilizada pelo professor; a segunda é a base *Breast Cancer Wisconsin (Diagnostic) Data Set*; e a terceira é a base *Iris Species Data Set*.

2 - Aproximação da Função *seno*:

Como já mencionado, o primeiro exercício consiste na aproximação da função *seno* no intervalo $[0;\ 2\pi]$. Esta função está representada na Figura 1.

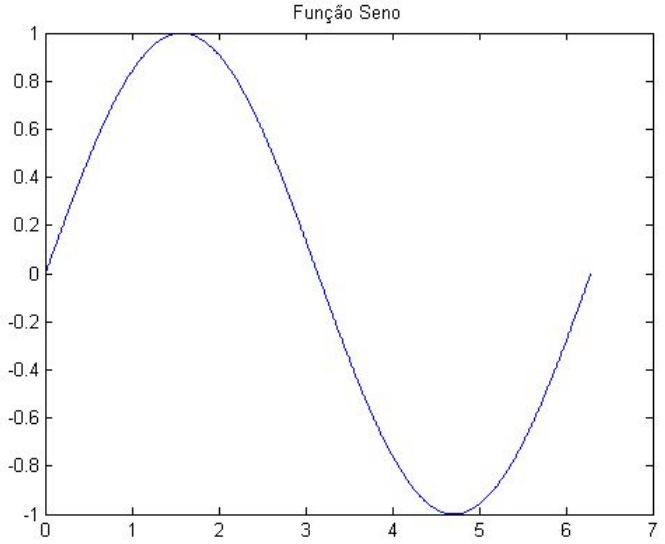

Figura 1 - Função Seno

Para a definição das regras iniciais, utilizou-se como auxílio a *toolbox Fuzzy* do *MatLAB*. Por esta ferramenta foi possível definir as regras iniciais. Conforme especificação, os antecedentes delas devem ter funções de pertinência do tipo Gaussiana. Os consequentes foram definidos segundo as regras utilizadas no Trabalho Computacional 2, apenas deslocando as função para se ajustarem ao intervalo deste problema para a função *seno*. Sendo assim, as funções utilizadas foram:

$$(1)\quad f1 = \frac{2}{\pi}x + 0$$

$$(2)\quad f2 = \frac{-2}{\pi}x + 2$$

$$(3)\quad f3 = \frac{2}{\pi}x - 4$$

sendo que a equação (1) é aplicada no intervalo $[0;\ \pi/2]$, a equação (2) no intervalo $[\pi/2;\ 3\pi/2]$ e a equação (3) no intervalo $[3\pi/2;\ 2\pi]$. A Figura 2 representa visualmente as três funções em seus intervalos, bem como a função *seno* para contraste.

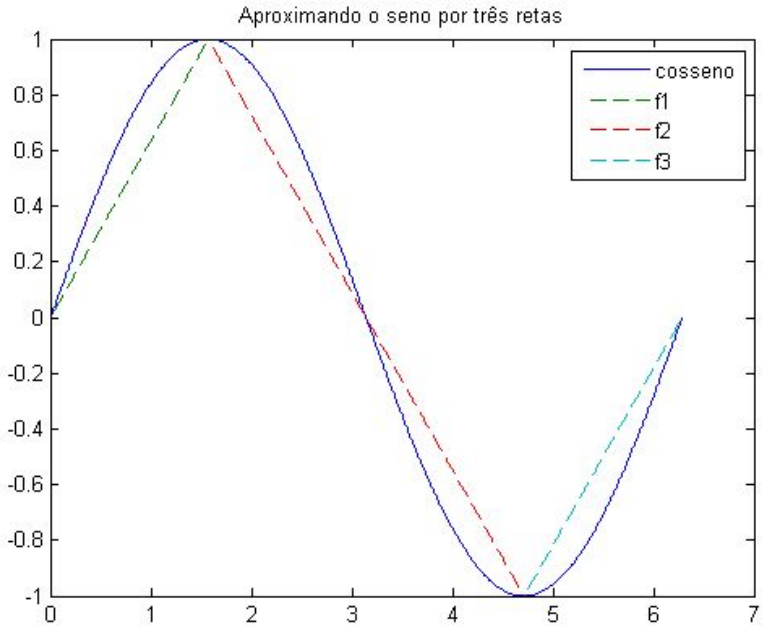

Figura 2 - Aproximando o seno por três retas

Desta forma, as três regras *fuzzy*, que possuem tais funções como consequente, ficam iguais a:

- *Se x é A1, y é f1*
- *Se x é A2, y é f2*
- *Se x é A3, y é f3*

No arquivo *sugeno_inicial.fis*, anexo junto a esta documentação, existe a definição destes parâmetros iniciais para a aproximação da função *seno* pelo ANFIS, com os antecedentes utilizando Gaussianas e com os consequentes iguais às funções *f1, f2* e *f3.* A

Figura 3 ilustra as configurações do arquivo *sugeno_inicial.fis*, construído atravéz do *toolbox fuzzy* do *MatLAB.*

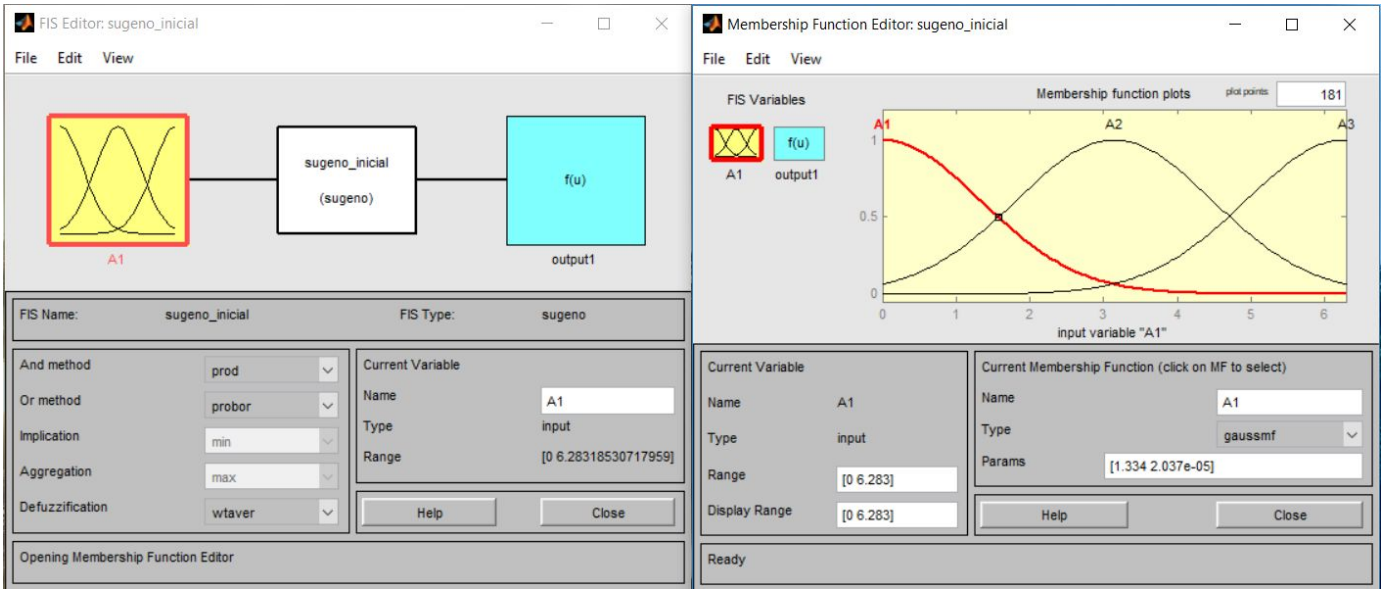

Figura 3 - Configuração do sugeno_inicial.fis

Com esta configuração inicial, utilizou-se a função *anfis* do *MatLAB* para se fazer a aproximação final. Foi passado como parâmetro os dados de treinamento, o sistema inicial configurado no arquivo *sugeno_inicial.fis* (lido com a função *readfis* do *MatLAB*) e uma quantidade de épocas máximas igual a 20. Todos estes procedimentos no *MatLAB* foram escritos no arquivo *main.m*, anexo junto a esta documentação. A Figura 4 mostra a aproximação final. Como se pode notar, a função alcançada pelo ANFIS projetado ficou muito próxima à função *seno* real.

Para uma análise quantitativa dos resultados, pode-se calcular o Erro Quadrático Médio, dado pela fórmula:

$$EQM = \frac{1}{N}\sum_{i=1}^{N}(yd_i - yv_i)^2$$

onde yd_i é a saída real da função *seno* e yv_i é a saída obtida pelo ANFIS projetado, ou seja, a aproximação. Este cálculo foi feito com o auxílio do *MatLAB*, também no script *main.m*, e o resultado obtido foi:

$$EQM = 0.000320$$

que é um valor muito pequeno, menor inclusive do que o alcançado no Trabalho Computacional 2 para a função *cosseno*, com EQM = 0.001283, representando o baixo erro da aproximação final.

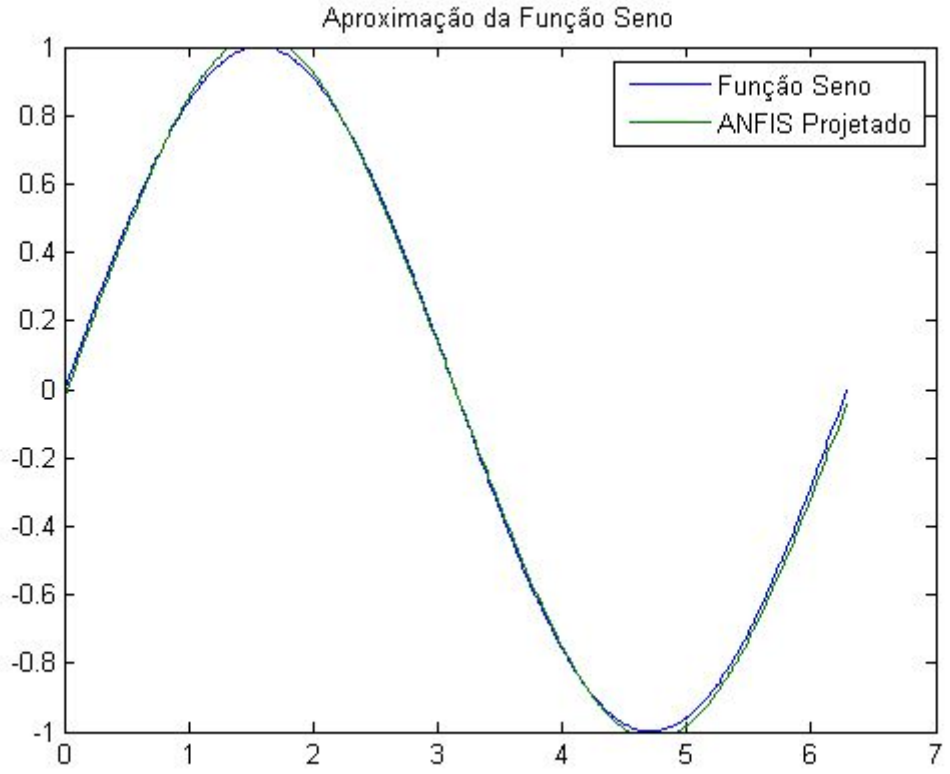

Figura 4 - Aproximação Final da Função Seno

3 - Classificador de Padrões ANFIS:

O próximo exercício consiste do projeto de um classificador de padrões por um sistema nebuloso adaptativo (ANFIS). O sistema projetado é do tipo *Takagi-Sugeno* de ordem 1. O classificador será testado, primeiramente, com uma base de dados disponibilizada pelo professor, contida no arquivo *dataset_2d.mat*. Em seguida ele também será testado com duas bases reais, a do problema *Breast Cancer Wisconsin*, contida no arquivo *breast_cancer.csv*, e a do problema *Iris Species Data Set*, contida no arquivo *iris.csv*. Todos os arquivos das bases de dados estão anexos a este relatório.

3.1 - Separação do Conjunto de Dados:

O primeiro passo para a criação e teste do algoritmo é a separação dos dados. Conforme especificação, o conjunto de dados X com dimensões $n \times d$ (onde n é o número de amostras e d é a dimensão do espaço de entrada) e a matriz Y com dimensões $n \times 1$, correspondendo aos rótulos das amostras, deve ser dividido entre treinamento e validação. 70% dos dados do conjunto devem ser utilizados para o treinamento, enquanto que os outros 30% devem ser utilizados para validar o modelo obtido.

Neste trabalho optou-se por fazer uma separação dos dados de forma balanceada. Desta forma, tanto o conjunto de treinamento quanto o conjunto de validação terão aproximadamente a mesma proporção de amostras de cada classe. Este ponto é importante para evitar viés de treinamento ou de validação provocado pelo fato de um dos grupos ter apenas amostras de uma classe, ou proporcionalmente muito mais amostras de uma classe do que de outra. Vale salientar que o conjunto de dados é separado de forma aleatória entre treinamento e validação.

Foi utilizada também, durante treinamento do sistema, a Validação Cruzada com 10 *folds*. Neste método, o conjunto de dados para treino é divido em 10 partes (*folds)*, e o sistema é treinado 10 vezes, sendo que em cada uma das vezes uma parte é utilizada para validação, enquanto que as outras nove são de fato utilizadas para treino. Desta forma, após os 10 treinos, as 10 partes serão utilizadas para validação, e os parâmetros do sistema podem ser melhor refinados.

3.2 - O Algoritmo do Classificador:

Após a parte da separação dos dados, é possível se fazer a lógica do treinamento do algoritmo. Um modelo inicial foi gerado utilizando o agrupamento do algoritmo *Fuzzy C-Means* para a inicialização dos parâmetros das funções de pertinência do antecedente. A quantidade de regras deste modelo inicial é um hiperparâmetro. O modelo foi gerado utilizando como auxílio a função *genfis3* do *MatLAB*. Em seguida, a função *anfis* foi utilizada para treinar o modelo final, recebendo como parâmetros os dados de treino (lembrando que 1/10 das amostras foram separadas para a validação cruzada, em cada iteração), o modelo inicial, e um máximo de 20 épocas. Com o modelo pronto, é necessário criar um limiar, que será discutido numa sessão mais a frente. No fim da iteração, é possível validar o modelo com o *fold* de dados separado para validação. Todo este procedimento é feito 10 vezes, para os 10 *folds.*

3.3 - Limiar de Saída do Modelo:

O modelo *fuzzy Takagi-Sugeno* de ordem 1, utilizado conforme especificação, foi inicialmente projetado para resolver problemas de regressão. Contudo, como o objetivo aqui é criar um classificador, é necessário estabelecer uma abordagem para adaptar o sistema ANFIS, já que ele retorna um número real para cada amostra, quando o que se quer é uma classificação binária, 0 ou 1. Uma abordagem simples, e adotada aqui, é a da definição de um limiar de saída do modelo. Neste caso, o limiar adotado é 0,5, conforme mostrado na equação a seguir, onde $\widehat{y}$ é a saída final do classificador.

$$\hat{y} = \begin{cases} 1 & \text{se } \hat{y} \geq 0.5 \\ 0 & \text{caso contrario} \end{cases}$$

3.4 - As Bases de Dados:

A primeira base de dados utilizada para testes, contida no arquivo *dataset_2d.mat,* já estava num bom formato para o treino do modelo, e não precisou passar por nenhuma intervenção. Já as bases *Breast Cancer Wisconsin* e *Iris Species Data Set* precisaram de adaptações. No primeiro caso, a coluna *ID Number* foi removida, por não ser uma informação relevante para o treino, e os rótulos estavam definidos como *M* (presença de tecido maligno) ou *B* (presença de tecido benigno), que foram mudados para 1 e 0, respectivamente. Já no caso da base *Iris Species Data Set*, a coluna *ID* também foi ignorada, e esta base possui a peculiaridade de que os rótulos são definidos como *Iris-setosa, Iris-versicolor* e *Iris-virginica*. Por especificação, o objetivo do classificador neste caso é classificar as amostras da espécie *Iris-setosa.* Desta forma, os rótulos de tais amostras foram mudados para 1, enquanto todos os outros foram mudados para 0. Vale salientar que os dados contidos nos arquivos *breast_cancer.csv* e *iris.csv* já estão preparados e prontos para uso, conforme o explicado nesta seção.

3.5 - Resultados:

O algoritmo do classificador está presente no arquivo *classifier.m*. Este arquivo define uma função que recebe como parâmetro as variáveis de entrada x, os rótulos y, a quantidade de regras K e uma variável $plotGraphs$ que pode ser usada para problemas com amostras de duas variáveis, para plotar os gráficos do resultado final da classificação do grupo de teste. Esta função retorna a acurácia do classificador final, do treino e da validação. A divisão dos dados nos 10 *folds* da validação cruzada foi implementada no arquivo *get_10_fold.m,* que contém uma função que recebe os dados e o *fold* atual que se deseja obter. Todos estes arquivos estão anexos junto a esta documentação.

Em primeiro momento, no arquivo *main.m*, anexo junto a esta documentação, a função do classificador é chamada 7 vezes, para a base *dataset_2d.mat*, para a quantidade de regras indo de 2 até 8. Foram plotadas as regiões de separação para 2, 3 e 8 regras,

que podem ser vistos na Figura 5. Para 2 regras, à esquerda, é possível notar que existiram mais erros do que nos outros casos, erros estes mostrados em “x”. Para 3 regras, à direita, o algoritmo chegou a 100% de acerto, e para 8 regras, a precisão caiu um pouco, com alguns erros. A Figura 6 mostra o gráfico das precisões de treino e teste nesta base de dados.

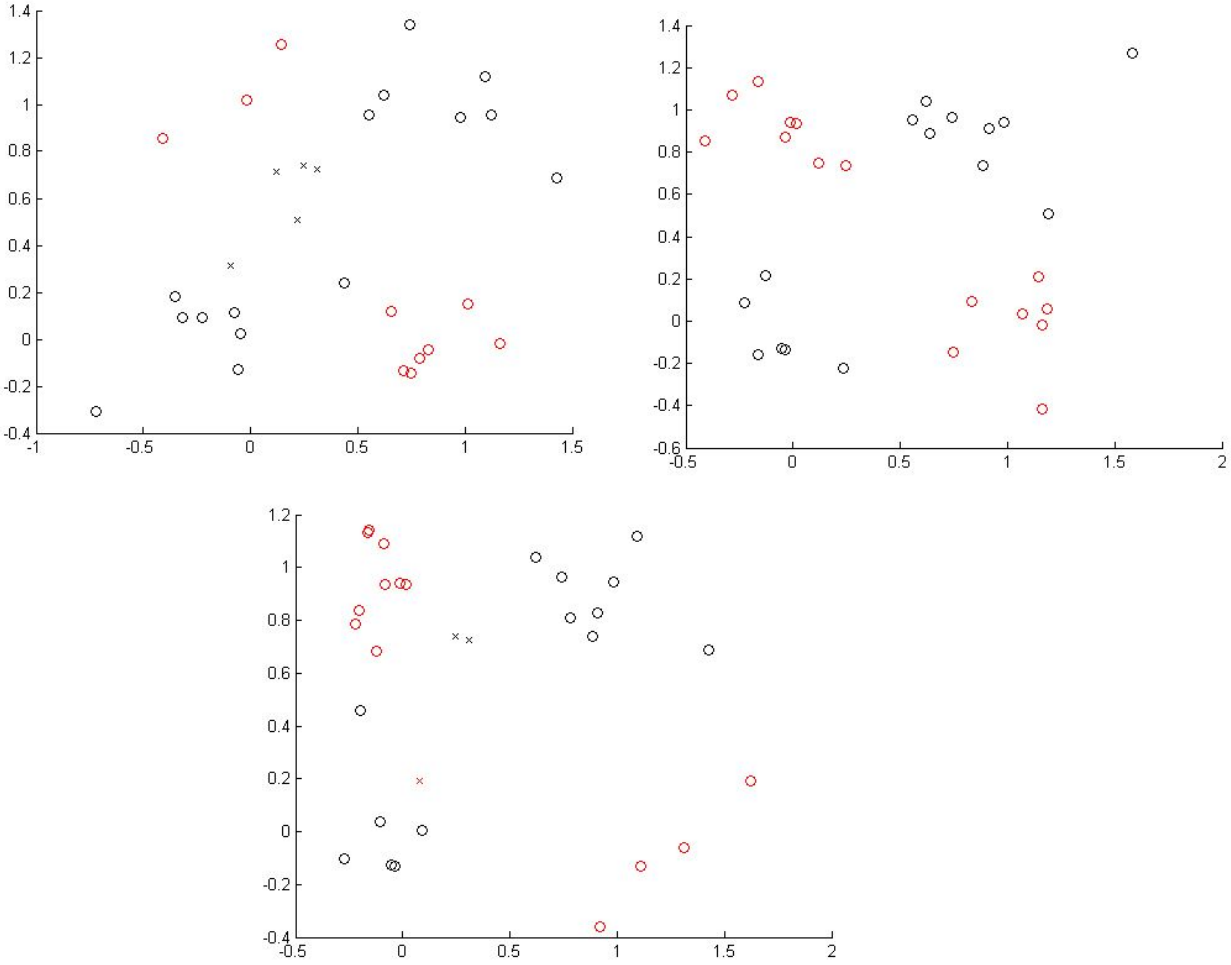

Figura 5 - Classificação dos pontos de *dataset_2d.mat* para 2, 3 e 8 regras

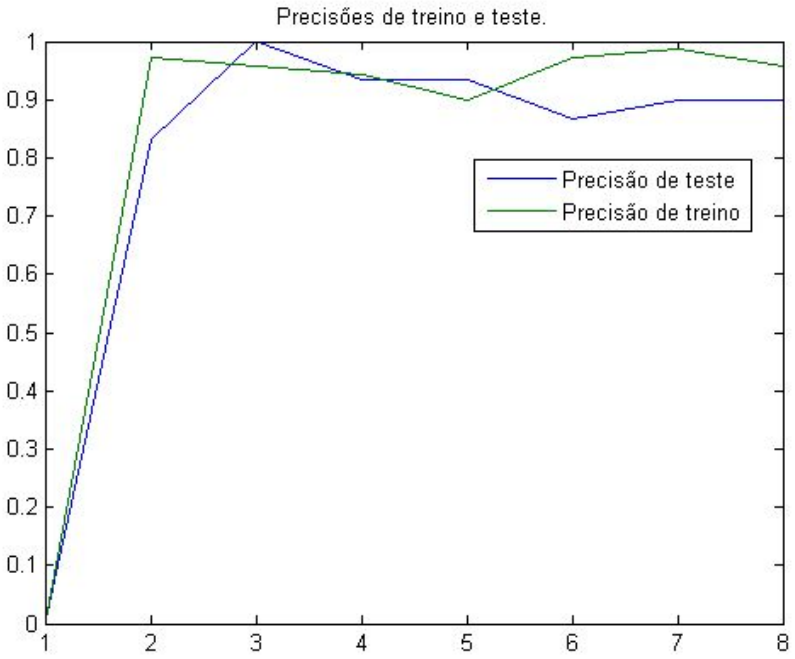

Figura 6 - Precisões de treino e teste para *dataset_2d.mat*

Em seguida, neste mesmo arquivo *main.m*, o mesmo procedimento foi feito com as bases *Breast Cancer Wisconsin* e *Iris Species Data Set*, com a quantidade de regras variando de 2 até 8. A precisão do treino e do teste para ambos os casos estão mostradas na Figura 7.

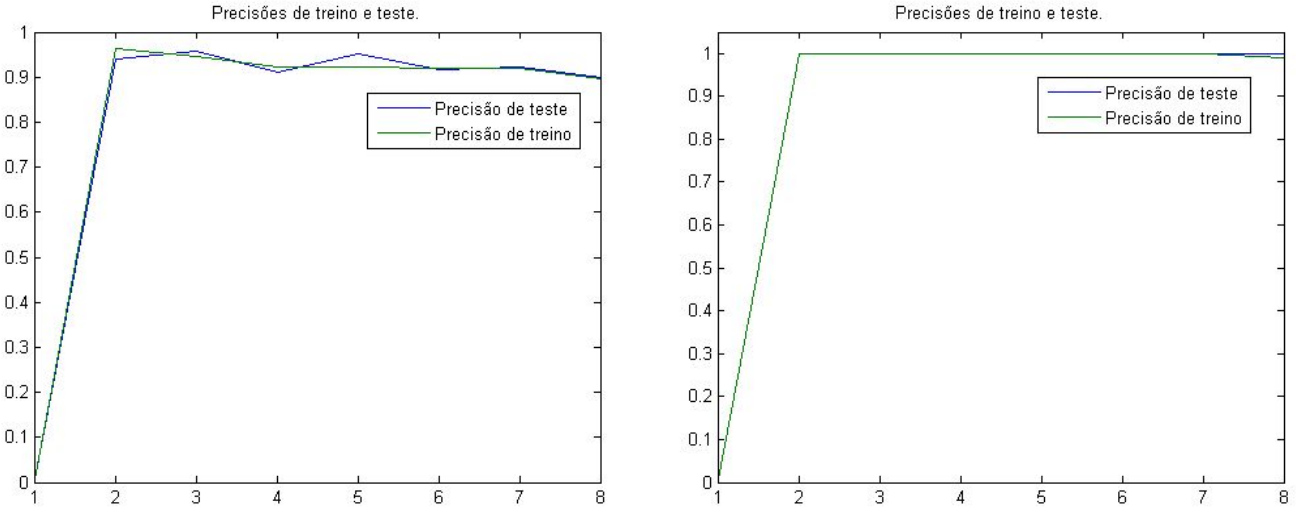

Figura 7 - Precisões de treino e teste para *breast_cancer.csv* (esquerda) e *iris.csv* (direita)

Como se pode notar nos gráficos de precisão, existe uma tendência de aumento da precisão de treino com o aumento da quantidade de regras. Contudo, a precisão de testes tende a se reduzir um pouco. Isto acontece principalmente por causa do efeito do aumento de complexidade do modelo com a adição de mais regras. Com maior complexidade, o modelo tende a “memorizar” a base de treino, passando a errar na classificação de teste. Contudo, em todos os casos analisados, a precisão ficou bem alta, quase sempre acima de 90% para os testes. Para a base *Iris Species Data Set,* devido a suas características favoráveis à classificação, a precisão chegou a 100% nos treinamentos de quase todos os casos analisados.

4 - Conclusão:

Neste trabalho foi apresentada a aproximação da função *seno* a partir dos conhecimentos obtidos na matéria Sistemas Nebulosos, por um sistema nebuloso adaptativo. A partir destes conhecimentos também foi apresentado e testado um classificador, também implementado através de um ANFIS. Estas implementações foram importantes para tornar os conceitos vistos em sala de aula mais palpáveis, por utilizações práticas que promovem a assimilação do conteúdo.

Desafios importantes foram superados, como a definição de um modelo inicial para a aproximação da função *seno*, bem como na utilização da validação cruzada no classificador e no projeto de um sistema inicial. Superados tais obstáculos, conclui-se que o objetivo do trabalho foi alcançado, através da promoção e prática dos conceitos vistos em sala de aula. O erro quadrático médio da aproximação da função *seno* foi muito baixo, menor inclusive do que o encontrado no Trabalho Computacional 2 para a função *cosseno*, e a aproximação ficou de fato muito precisa. As precisões do classificador baseado em sistemas nebulosos adaptativos também ficaram aceitáveis, na maioria dos casos ficando superiores a 90%. Inclusive foi possível ver os efeitos do aumento da complexidade do modelo pela adição de regras. Para a base *Iris Species Data Set,* com técnicas de aprendizado de máquina como o *KNN* com *Petals*, é possível alcançar modelos com 98% de precisão, e para *Breast Cancer Wisconsin* utilizando *Random Forest,* é possível chegar à precisão de 96%. Ambos estes patamares foram alcançados pelo classificador utilizando sistemas nebulosos adaptativos.