IA353 – Exercícios de Fixação de Conceitos
EFC 2 – 1s2017

Questão 5) (1,0 pontos) Data de entrega: 08/05/2017

Emprego da metodologia *wrapper* com *backward elimination* para seleção de variáveis em duas tarefas de regressão e em uma tarefa de classificação. Comparação com a abordagem baseada em filtro.

(1) Recorra ao paper [Guyon, I.; Elisseeff, A. "An introduction to variable and feature selection", Journal of Machine Learning Research, vol. 3, pp. 1157-1182, 2003] para explicar a diferença entre filtro e *wrapper*.

(2) Explique como funcionam as abordagens *forward selection* e *backward elimination* e apresente a razão pela qual elas não garantem encontrar a melhor combinação de entradas para a tarefa. Aponte vantagens e desvantagens da abordagem *backward elimination*.

(3) Caso de estudo 1: Série temporal SUNSPOT, do ano de 1749 ao ano de 2014 [http://www.esrl.noaa.gov/psd/gcos_wgsp/Timeseries/Data/sunspot.long.data].
São considerados 20 atrasos candidatos (a entrada 1 corresponde ao atraso 20 e a entrada 20 corresponde ao atraso 1), são incluídas 5 entradas aleatórias e deve ser empregado 10-folds cross-validation. Todas as variáveis excursionam no intervalo [0,+0.2]. Operação com preditor linear e preditor ELM, ambos com regularização.

   a. Use os programas do diretório [wrapper_sunspot]. Comente acerca da série temporal sunspot: a que seus valores se referem, qual é o período estimado desta série, qual é a periodicidade dos valores medidos, etc.
   b. Use o programa [pre_proc.m], que vai carregar os dados do arquivo [sunspot.txt]. Defina o número de atrasos candidatos como sendo 20 e o número de entradas aleatórias como sendo 5. A execução de [pre_proc.m] vai preparar os dados de treinamento, salvos no arquivo [train.mat].
   c. Use o programa [lin_filter.m] com argumento 'train' de modo a obter as correlações lineares de Pearson entre as 25 entradas candidatas e a saída (predição a ser realizada). Comente os resultados obtidos.
   d. Use o programa [nlin_filter.m] com argumento 'train' de modo a obter as correlações não-lineares de Spearman entre as 25 entradas candidatas e a saída (predição a ser realizada). Comente os resultados obtidos e compare com o item (3c).
   e. Use o programa [gen_k_folds.m], com arquivo de entrada 'train' e número de pastas para validação cruzada igual a 10. Do único arquivo de treinamento, serão produzidas 10 pastas (arquivos [train1.mat] até [train10.mat]).
   f. Use o programa [wrapper_lin_regr.m], fornecendo as mesmas informações dos programas anteriores, de modo a realizar a seleção de variáveis empregando wrapper com backward elimination. Comente os resultados obtidos e compare com os itens (3c) e (3d). Não deixe de apresentar as entradas que produzem o menor erro médio RMS para os 10 conjuntos de validação.
   g. Apresente o diagrama de barras com a frequência de escolha dos 7 valores definidos para a regularização do modelo linear. É necessário regularizar um modelo linear? Justifique.

h. Não é necessário verificar na prática, mas se você executar repetidas vezes o item (3e) seguido do item (3f), a cada execução a sequência de entradas selecionada para sair do modelo pode variar. Explique o motivo.
i. Explique por que nem sempre as entradas de menor correlação (Pearson ou Spearman) são as primeiras a serem podadas na abordagem *backward selection*, sendo algumas dessas entradas de baixa correlação até selecionadas para compor a entrada do modelo, ao final.
j. Use o programa [wrapper_nlin_regr.m], fornecendo as mesmas informações dos programas anteriores, de modo a realizar a seleção de variáveis empregando wrapper com backward elimination. O modelo não-linear deve ser uma ELM com um número de neurônios em torno de 100. Comente os resultados obtidos e compare com o item (3f). Não deixe de apresentar as entradas que produzem o menor erro médio RMS para os 10 conjuntos de validação.
k. Apresente o diagrama de barras com a frequência de escolha dos 7 valores definidos para a regularização do modelo não-linear. Comente.

(4) Caso de estudo 2: Conjunto de dados *Wine Quality* do *UCI Machine Learning Repository* [http://archive.ics.uci.edu/ml/datasets/Wine+Quality]. Foi considerado o caso com 1599 amostras e foram eliminadas as linhas 1361, 1364, 1441, 1443, 1477 e 1516, por conterem dados espúrios. São considerados 11 atributos de entrada, devem ser incluídas 5 entradas aleatórias e deve ser empregado 10-folds cross-validation. Todas as variáveis excursionam no intervalo [–0,2;+0,2]. Operação com preditor linear e preditor ELM, ambos com regularização.
a. Use os programas do diretório [wrapper_wineq].
b. Use o programa [pre_proc.m], que vai carregar os dados do arquivo [wineq.txt]. Defina o número de entradas aleatórias como sendo 5. A execução de [pre_proc.m] vai preparar os dados de treinamento, salvos no arquivo [train.mat].
c. Use o programa [lin_filter.m] com argumento 'train' de modo a obter as correlações lineares de Pearson entre as 16 entradas candidatas e a saída (predição a ser realizada). Comente os resultados obtidos.
d. Use o programa [nlin_filter.m] com argumento 'train' de modo a obter as correlações não-lineares de Spearman entre as 16 entradas candidatas e a saída (predição a ser realizada). Comente os resultados obtidos e compare com o item (4c).
e. Use o programa [gen_k_folds.m], com arquivo de entrada 'train' e número de pastas para validação cruzada igual a 10. Do único arquivo de treinamento, serão produzidas 10 pastas (arquivos [train1.mat] até [train10.mat]).
f. Use o programa [wrapper_lin_regr.m], fornecendo as mesmas informações dos programas anteriores, de modo a realizar a seleção de variáveis empregando wrapper com backward elimination. Comente os resultados obtidos e compare com os itens (4c) e (4d). Não deixe de apresentar as entradas que produzem o menor erro médio RMS para os 10 conjuntos de validação. Você vê diferenças significativas entre o que ocorreu até aqui e o que havia sido observado no item (3f)? Justifique.
g. Apresente o diagrama de barras com a frequência de escolha dos 7 valores definidos para a regularização do modelo linear.
h. Use o programa [wrapper_nlin_regr.m], fornecendo as mesmas informações dos programas anteriores, de modo a realizar a seleção de

variáveis empregando wrapper com backward elimination. O modelo não-linear deve ser uma ELM com um número de neurônios em torno de 80. Comente os resultados obtidos e compare com o item (4f). Não deixe de apresentar as entradas que produzem o menor erro médio RMS para os 10 conjuntos de validação.

i. Apresente o diagrama de barras com a frequência de escolha dos 7 valores definidos para a regularização do modelo não-linear.

(5) Caso de estudo 3: Conjunto de dados *Wisconsin Diagnostic Breast Cancer* do *UCI Machine Learning Repository* [http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.data]. O qual dispõe de 30 atributos de entrada e duas classes (*malignant* – Classe 1 e *benign* – Classe 2). Não considere aqui os filtros, executando apenas os wrappers linear e não-linear. Devem ser incluídas 5 entradas aleatórias e deve ser empregado 5-folds cross-validation. Todas as variáveis excursionam no intervalo [−0,2;+0,2]. Operação com preditor linear e preditor ELM, ambos com regularização. Execute o wrapper linear e o não linear 5 vezes, para partições em pastas diferentes a cada vez. Para tanto, após rodar uma única vez [pre_proc.m], rode [gen_k_folds.m] e os dois wrappers cinco vezes, nesta sequência: [gen_k_folds.m] + dois wrappers. Colete os resultados das entradas selecionadas numa tabela, compare os resultados e comente.

(6) Caso de estudo 4: Conjunto de dados *Image Segmentation data* do *UCI Machine Learning Repository* [http://archive.ics.uci.edu/ml/datasets/Image+Segmentation]. O qual dispõe de 18 atributos de entrada e sete classes (*brickface* – Classe 1, *sky* – Classe 2, *foliage* – Classe 3, *cement* – Classe 4, *window* – Classe 5, *path* – Classe 6, *grass* – Classe 7). Originalmente, são 19 atributos, mas um deles assume um único valor e foi descartado. Não considere aqui os filtros, executando apenas os wrappers linear e não-linear. Devem ser incluídas 5 entradas aleatórias e deve ser empregado 5-folds cross-validation. Todas as variáveis excursionam no intervalo [−0,2;+0,2]. Operação com preditor linear e preditor ELM, ambos com regularização. Execute o wrapper linear e o não linear 5 vezes, para partições em pastas diferentes a cada vez. Para tanto, após rodar uma única vez [pre_proc.m], rode [gen_k_folds.m] e os dois wrappers cinco vezes, nesta sequência: [gen_k_folds.m] + dois wrappers. Este conjunto contém dados de teste, que tem 10 vezes mais amostras que os dados de treinamento. Repita todas as etapas para os dados de teste no lugar dos dados de treinamento. Para tanto, realize alterações no programa [pre_proc.m]. Colete os resultados das entradas selecionadas numa tabela, compare os resultados e comente.