Universidade Federal de Campina Grande
Projeto HP-FRH-Analytics

Lab 2 - Introdução às Redes Neurais

Objetivo: O objetivo deste Lab é verificar na prática conceitos básicos de Redes Neurais do tipo Perceptron de múltiplas camadas (*Multilayer Perceptron*) treinadas com o algoritmo *backpropagation*. Como objetivo adicional, deseja-se estimular o aprendizado da ferramenta Weka. Pode ser utilizada a interface gráfica do Weka ou criar scripts a partir da importação das bibliotecas em Java. Neste segundo caso, deve-se enviar também o código Java para cada questão desenvolvida.

Entrega: Até as 23h59min do dia 26/04/2012.
Formato da entrega: Idêntico ao Lab anterior. Ao escrever seu relatório, lembre-se de: (1) incluir artefatos (imagens, dados, etc.) que comprovem sua resposta, (2) tornar os passos utilizados fáceis de entender e replicar.

Warm-up - Reproduza os resultados obtido nas questões 1, 2 e 3 **do Lab 01** utilizando, desta vez, o ferramental do Weka.

- **Esta resposta deve constar no seu documento de análise**.
- **O modelo gerado (extensão do arquivo: .model) pela ferramenta Weka também deve ser entregue**.

Dados: A base SPAMBASE foi criada pelo HP-Labs com o intuito de identificar se uma determinada mensagem de correio eletrônico é ou não um SPAM. A maioria dos atributos diz respeito ao percentual de ocorrências de uma determinada palavra ou caractere no corpo do e-mail.

1 - Após a instalação do **Weka**, carregue a base de dados **SPAMBASE**, sobre a qual você deve responder a cada uma das seguintes perguntas.

- **Esta resposta deve constar no seu documento de análise**. Para cada pergunta, devem ser apresentados os resultados do treinamento acompanhados das explicações dos resultados encontrados.

a. Qual o impacto, no resultado do treinamento[1], da variação do número de neurônios na **camada escondida**?
b. No Weka, é possível utilizar como conjunto de testes: (a) o conjunto de treinamento inteiro, (b) um conjunto de dados a ser carregado, (c)

[1] *Por resultado do treinamento, neste lab, entenda como sendo a propagação do conjunto de* ***testes*** *usando a rede treinada.*

validação-cruzada e (d) divisão proporcional do conjunto. Qual o impacto da na escolha dos métodos a, c ou d?

c. Qual o impacto da seleção de parte do conjunto de treinamento para ser utilizado como conjunto de validação (não confundir com validação-cruzada)?
d. Como você poderia, com 95% de confiança, verificar se as diferenças ou semelhanças apontadas em a) e b) são significativas? **Basta descrever** a metodologia.
e. O problema em questão é **linearmente separável**? Justifique!
f. Qual o impacto nos resultados da alteração da **semente de treinamento**?
g. Qual o impacto do balanceamento dos dados das classes no conjunto de treinamento? Decida qual método de balanceamento usar e justifique.

2 - Após todos os experimentos realizados na primeira questão, qual o melhor resultado obtido? A partir das métricas obtidas (a saber, TP, FP, TN, FN, Precision, Recall, F-measure, área sob a ROC e matriz de confusão) como se pode analisar o resultado? Use apenas as métricas que julgar necessário para dar suporte a sua argumentação.

- **Esta resposta deve constar no seu documento de análise**.
- **O modelo gerado pelo weka para a melhor rede deve ser entregue**.

3 - Caso 99% dos dados sejam utilizados para o **treinamento** o percentual de acerto pode chegar a mais de 90%. Comente (1) se isso é possível, (2) o(s) motivo(s), (3) a consequência disto.

- **Esta resposta deve constar no seu documento de análise**.

4 - Sobre os melhores resultados, aplique a Análise de Componentes Principais (PCA) e analise se houve algum benefício da aplicação desta técnica nos resultados de classificação. Indique também a quantidade de elementos gerados pelo PCA, após a transformação, para uma cobertura de 95%.

- **Esta resposta deve constar no seu documento de análise**.
- **O modelo gerado pelo weka após o treinamento usando PCA deve ser entregue junto ao arquivo cujos dados correspondem ao resultado da aplicação do PCA na base original de dados**.

5 - Em um problema de classificação de SPAMs, qual o erro (FP ou FN) é o mais indesejável? Levando sua resposta em consideração, você mudaria a resposta da **questão 3**? Justifique!

- **Esta resposta deve constar no seu documento de análise**.