Universidade Federal de Campina Grande
Projeto HP-FRH-Analytics

Lab 4

Objetivo: O objetivo deste Lab é a compreensão dos conceitos de **redes recorrentes** e **séries temporais**.

Entrega: Até as 23h59min do dia 17/05/2012.

> **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.

Bases de Dados: Neste exercício será utilizada a base de dados House Sales, a qual descreve, segundo Makridakis, Wheelwright and Hyndman (1998), a venda de casas novas para uma família a cada mês nos EUA desde 1973. Essa base de dados é composta por uma única coluna que indica a quantidade de casas vendidas em um dado mês.

Conhecimento prévio: Neste lab estaremos discutindo problemas de séries temporais e a base utilizada é composta por dados numéricos. Como métrica do erro, calcule as medidas **SSE** e **RMSE**. **SSE** é a soma quadrática do erro e o **RMSE** é a raiz do erro médio quadrático, ambas podendo ser definidas, na linguagem R (aliado à utlilização da biblioteca RSNNS), por:

```
sse = sum((model$fittedTestValues-patterns$targetsTest)^2)
rmse = sqrt(mean((model$fittedTestValues-patterns$targetsTest)^2))
```

Para todas as questões, a análise da qualidade dos resultados deve ser realizada com base nos valores do SSE e do RMSE.

Construção da base dos dados: Em geral, nos problemas de **séries temporais**, os dados precisam ser preparados **antes** do treinamento. O objetivo é **prever** o valor a ser obtido em um período subsequente com base nos períodos anteriores. Logo, tendo os dados de janeiro, fevereiro, março e abril, deseja-se prever o mês de maio. Isso é feito a partir de uma **janela**. Por exemplo, considere os seguintes dados hipotéticos:

Mês	***Medição***
Jan	10
Fev	20
Mar	30
Abr	40
Mai	20
Jun	30

Suponha que você decida por uma janela de **3 meses**. Assim, você utilizará 3 meses como entrada para a rede para a previsão do quarto mês. Logo, na primeira linha serão dispostos os dados de janeiro, fevereiro e março para prever o de abril; na segunda linha os dados de fevereiro, março, abril para prever o de maio e assim por diante. Dessa forma, a base ficará assim:

med_mes1	**med_mes2**	**med_mes3**	**previsto**
10	20	30	40
20	30	40	20
30	40	20	30
40	20	30	...

No caso ilustrado acima, as três primeiras colunas serão utilizadas como dados de entrada para a rede e a quarta coluna é a que apresenta os valores esperados para as respectivas entradas. Naturalmente, as últimas j linhas (em que j é o tamanho da janela) não irão dispor de dados em algumas colunas. Descarte tais linhas. Neste laboratório, o procedimento descrito acima, para preparação dos dados, é de **sua responsabilidade**.

Normalização: Os dados precisam ser normalizados antes de serem apresentados à rede neural. As normalizações mais comuns são entre -1 e 1 ou entre 0 e 1. Para normalizar os dados entre 0 e 1, aplique a seguinte fórmula:

$$x_{norm} = \frac{x - min(dados)}{max(dados) - min(dados)}$$

Na linguagem R:

vec_x = vec_x-min(vec_x)/(max(vec_x)-min(vec_x))

Para todas as questões, devem ser entregues, sempre que pertinente, os código produzidos e as análises dos resultados.

1) Preparação da base de dados:

a) Carregue os dados do arquivo [**hsales.dat**].

b) Crie uma variável [**dados.de.venda**], na qual você construirá uma base de dados com uma janela de **3 meses**.

c) Realize a normalização dos dados (nas colunas que julgar pertinente);

2) Treine uma Rede MLP para calcular a qualidade da previsão da série temporal para a base fornecida. Defina os parâmetros da rede a sua escolha.

3) De acordo com a teoria estudada na disciplina, Redes Recorrentes são mais indicadas para a previsão de séries temporais. Compare os resultados do treinamento da rede MLP com as redes de ELMAN e de JORDAN. Você confirma ou refuta a teoria com base nesses dados?

4) "Problemas de regressão podem apresentar melhores resultados se utilizadas redes com duas camadas escondidas, ao invés de uma única". Execute experimentos usando as duas formas (uma ou duas camadas escondidas) e use um método estatístico apropriado para definir se, para confiança de 95%, você confirma ou refuta essa afirmação.

Experimentos adicionais: Estes experimentos não são obrigatórios, não valem ponto, mas podem ser realizados por aqueles que terminarem o lab com antecedência como fim de aprimorar os conhecimentos neste tema.

EA1) Realize o treinamento indicado em 2 utilizando a base de dados pura (com apenas uma entrada) ao invés de utilizar a janela deslizante. Como varia o comportamento de uma rede MLP e de uma rede recorrente?

EA2) Analise o impacto do aumento / diminuição do tamanho da janela.