Mini-Projeto 4 - Séries Temporais

1. Introdução e Descrição dos Dados

Vamos analisar as finanças de nosso querido país! Temos uma base de dados relativos à **taxa de juros básica**[1] estabelecida pelo Banco Central diariamente. Leiam sobre o assunto no link fornecido e aproveitem para aprender sobre um dos componentes mais importantes da economia de um país. Essa base tem dados de 04/01/1999 até 03/02/2014.

O seu trabalho nesse mini-projeto consistirá em predizer o futuro da taxa de juros do país! Para tal você receberá uma parcela dos dados para treino e outra parcela para teste, o teste, como de praxe, deverá ser utilizado para predição com o melhor modelo selecionado nos dados de treino. O erro deve ser calculado em termos do *Erro Absoluto*.

2. Fases

Esse mini-projeto é dividido nas 3 fases descritas abaixo. Todos os processamentos deverão acontecer sobre os dados de treino. Os dados de teste disponibilizados serão utilizados apenas para predição.

2.1. Fase 1: Análise e Preparação

Descrição

A 1ª fase requer que você estude os dados, ou seja, busque dados muito diferentes do restante (possíveis outliers), busque dados faltantes (onde deveria haver dados e não tem), visualize o padrão temporal dos dados, realize transformações de escala (dia, mês, trimestre, ano) e quaisquer outras análises que ajudem-o a entender/descrever os dados.

Após a análise inicial deve-se preparar a base de dados para que seja possível treinar um modelo de predição. Nessa fase deve-se realizar dois janelamentos dos dados (visto em sala de aula). Um terá um horizonte de curto prazo (2 dias) e outro horizonte de longo prazo (15 dias), o tamanho das janelas deverão ser definidos por você (tamanho máximo de 60 dias).

Após ter a série janelada você irá executar funções de extração de características (também estudadas em sala ou outras que desejar) sobre os dados de cada janela. Assim, cada janela terá um vetor de características que será unido com a váriável de saída (valor dos juros após 2 dias/15 dias) e formará uma tabela de dados para predição. Com essa nova tabela o problema de predição de séries temporais reduz para um problema de regressão e quaisquer algoritmos poderão ser utilizados (em especial os que foram estudados nos mini-projetos anteriores).

Exemplo do processo

Dados de entrada:

dia *valor*

[1] http://pt.wikipedia.org/wiki/Juro#Taxa_b.C3.A1sica_de_juros

01-01-2010	*1.0*
02-01-2010	*1.1*
03-01-2010	*0.8*
04-01-2010	*0.7*
05-01-2010	*1.3*
06-01-2010	*1.6*
07-01-2010	*0.2*
08-01-2010	*0.3*
09-01-2010	*0.4*
10-01-2010	*0.5*

Dados janelados (janela de tamanho 3 e horizonte 2):

dia_horizonte	*jan1*	*jan2*	*jan3*	*horizonte2*
05-01-2010	*1.0*	*1.1*	*0.8*	*1.3*
06-01-2010	*1.1*	*0.8*	*0.7*	*1.6*
07-01-2010	*0.8*	*0.7*	*1.3*	*0.2*
08-01-2010	*0.7*	*1.3*	*1.6*	*0.3*
09-01-2010	*1.3*	*1.6*	*0.2*	*0.4*
10-01-2010	*1.6*	*0.2*	*0.3*	*0.5*

As **características** serão geradas utilizando os dados das janelas (no exemplo: *jan1*, *jan2* e *jan3*) e você escolherá quais atributos usar, as características mais simples a serem usadas seriam os próprios dados das janelas, gerando assim a mesma tabela acima. A coluna *dia_horizonte* corresponde ao dia que será previsto pelo modelo (ou seja dia do *horizonte2*), não ao dia da predição (que no caso do exemplo seria o dia da *jan3*).

Artefatos

Arquivo PDF (nome: *mp4_analises.pdf*) com:

- Resultados da análise inicial descrita acima (pode conter gráficos, etc.). Mais valor terá esse artefato quanto melhor for a descrição dos dados, importância e embasamento dos argumentos e organização.
- Tamanho das janelas para cada janelamento.
- Características implementadas e extraídas para cada janelamento.
- Código Fonte (em R)

2.2. Fase 2: Treino e Predição a Curto Prazo

Descrição

Nesse momento, será selecionado o melhor modelo de regressão para ser utilizado na predição com os dados de teste.

Seleção de Modelos por Validação

O processo de seleção deverá otimizar o *Erro Absoluto* (ou seja, minimizá-lo), o qual deverá ser implementado. Ao menos **3 modelos** deverão ser comparados entre si. Será utilizado um conjunto de validação gerado da seguinte forma:

1. Os 60% iniciais dos dados de treino (arredondando para cima) são para treino de fato e os 40% finais para validação
2. Nos dados de validação:
    a. Execute o janelamento
        i. Gerar a seguinte tabela: DIA_HORIZONTE | JANELA | SAÍDA_REAL
    b. Extraia as características dos modelos
        i. Gerar a seguinte tabela: CARACTERÍSTICAS | SAÍDA_REAL

Depois para cada modelo:

3. Usando as características, compute a SAÍDA_PREVISTA
4. Calcule o ERRO ABSOLUTO: SAÍDA_REAL - SAÍDA_PREVISTA

Ao término gere um **arquivo .csv** com 3 colunas:

- NOME DO MODELO | DIA_HORIZONTE (formato ano-mês-dia) | ERRO ABSOLUTO.

Compare os valores do Erro Absoluto (usando o aplicativo AD2_HOME[2], por exemplo) e selecione o melhor modelo.

Predição

Com o melhor modelo gere as predições sobre os dados de teste, o que inclui:

- Gerar tabela de dados janelados com **horizonte de 2 dias**
- Gerar tabela com características
- Gerar tabela com as predições

Ao término gerar um **arquivo .csv** com 2 colunas:

- DIA_HORIZONTE (formato ano-mês-dia) | SAIDA_PREVISTA

Artefatos

- 1 Arquivo **.csv** (mp4_validacao_fase1.csv):
    - Com os **dados de validação** como descrito acima (sem cabeçalho)
- 1 Arquivo **.csv** (mp4_predicao_fase1.csv):
    - Com os **dados da predição** como descrito acima (sem cabeçalho)
- Código Fonte (em R)

[2] analisedados2.lsd.ufcg.edu.br/ad2_home

2.3. Fase 3: Treino e Predição a Longo Prazo

Descrição

Essa fase terá os mesmos passos da anterior, só mudará o tamanho do horizonte de 2 dias para **15 dias**.

Artefatos

- 1 Arquivo **.csv** (mp4_validacao_fase2.csv):
  - Com os **dados de validação** como descrito acima (sem cabeçalho)
- 1 Arquivo **.csv** (mp4_predicao_fase2.csv):
  - Com os **dados da predição** como descrito acima (sem cabeçalho)
- Código Fonte (em R)
- Arquivo PDF (*mp4_analises.pdf):*
  - Complementação do arquivo de análise PDF da Fase 1 com as conclusões/aprendizados durante o processo de treino, validação e predição.

3. Nota

- 3.0 pontos : Fase 1
- 3.0 pontos : Fase 2
- 2.5 pontos : Fase 3
- 1.0 ponto : Presença em sala e apresentação final
- 0.5 ponto : Participação na lista do piazza

Todas as entregas devem ser realizadas até o dia **13/02/2014** às **23h59**.

4. Aulas

Aula 11/02: Terça-Feira no RE-08

Aula teórica sobre o assunto.

Aula 14/02: Sexta-Feira no RE-08

Apresentação do resultado. Atenção! Você será argüido(a) neste momento pelos professores, o que pesará na sua avaliação! Esteja preparado(a)!

5. Lembretes

- Esse mini-projeto é **individual**.
- Todas as aulas terão lista de presença.

Bom trabalho! =)