Universidade Federal de Campina Grande
Departamento de Sistemas e Computação
Programa FRH-Analytics UFCG/HP
Disciplina: Fundamentos de Analytics

Lab03 - Tópicos de probabilidade para *data analytics*

Aula para tirar dúvidas: **28/02/2012**
Limite de entrega: **01/03/2012 às 23:59**
Disponibilização do gabarito: **02/03/2012 às 00:00**

Exercício

Instruções gerais

Neste lab usaremos dados de 19.548 usuários obtidos na comunidade Ask Ubuntu (http://askubuntu.com/) em setembro de 2011. Os dados brutos podem ser baixados na página da disciplina. Cada coluna do arquivo de dados é uma variável aleatória e cada linha apresenta as seguintes informações para um usuário.

- **display_name** *[Character]*: Nome escolhido pelo usuário para ser apresentado em seu perfil na comunidade;
- **reputation** [*integer*]: Reputação acumulada pelo usuário na comunidade;
- **views** [*integer*]: Número de visualizações do perfil do usuário;
- **up_votes** [integer]: Número de votos positivos atribuídos pelo usuário a *posts* na comunidade;
- **down_votes** [integer]: Número de votos negativos atribuídos pelos usuários a *posts* na comunidade;

Para cada questão deste exercício, teremos dois tipos de resposta: uma em formato **.R* e outra em formato **.pdf*. O nome dos arquivos de resposta devem seguir os seguintes formatos, onde X é o número da questão:

- *seunome-lab03-questaoX.R* para respostas no formato de scripts *R*
- *seunome-lab03-questaoX.pdf* para respostas discursivas em formato texto.

Todos os processamentos estatísticos e os gráficos podem ser feitos utilizando o *R padrão* (o gabarito está implementando dessa forma). Entretanto, caso você queira utilizar alguma biblioteca do R, recomenda-se a biblioteca “ggplot2” para geração dos gráficos e as bibliotecas “stats” e “Hmisc” para os cálculos estatísticos.

Questões:

1) Escreva um *script* R que gera o gráfico da Função de Distribuição Acumulada (FDA), da Função de Densidade de Probabilidade (FDP), para variáveis contínuas, e da Função de Massa de Probabilidade (FMP), para variáveis discretas. Os dados que devem ser utilizados para gerar os gráficos são: *reputation*, *views*, *up_votes* e *down_votes*. Apenas para melhorar a visualização dos dados, limite o eixo x dos gráficos entre 0 e o valor do 99 percentil. Devem ser geradas duas figuras no formato *.png: uma figura contendo os gráficos FDP e/ou FDM e outra figura contendo os gráficos FDA. Cada gráfico deve ser devidamente identificado com o nome da variável aleatória a que se refere. Escreva um comentário com sua interpretação dos gráficos em termos da concentração e dispersão dos dados.

2) Suponha que você deseja analisar o número de visualizações dos perfis dos usuários (coluna *views*) na comunidade Ask Ubuntu. Dada a diversidade de valores que esse atributo pode obter, você decide classificar os usuários usando uma medida estatística para, então, identificar a probabilidade de ocorrência de usuários em cada classe. Neste contexto, pede-se:

- Escreva um *script* R que classifica os usuários segundo o número de visualizações do perfil. Para isso, gere uma nova coluna com a *classe* do usuário. Classifique as visualizações do perfil do usuário em “Muito baixo”, “Baixo”, “Alto” e “Muito Alto”, utilizando os quartis 0.25, 0.5 e 0.75. Por exemplo, se o número de visualizações de um usuário for menor ou igual ao valor do 0.25 quartil o atributo (coluna) “classe” desse usuário deve receber o valor “Muito Baixo”. Após isso, o *script* deve gerar uma figura em formato *.png contendo a FDA e abaixo dela o histograma gerados com os dados da coluna *classe*.

- Avalie outros métodos de classificação, por exemplo, mudando os valores dos limiares (quartis). Escreva um relato comentando (i) se você gostou da classificação usando os quartis, fundamente sua resposta com uma análise dos gráficos; (ii) qual dos método de classificação que você avaliou e que se mostrou mais adequado, defenda o método; (iii) quais relações você observa entre a FDA e o histograma; (iv) quais os impacto do método de classificação no formato da FDA e do histograma.

3) Suponha que você precisa realizar simulações utilizando dados de reputação de 25.000 usuários da comunidade Ask Ubuntu. Para tanto, você decide usar os dados brutos de reputação referente aos 19.548 usuários. Com esses dados você pode gerar uma distribuição de probabilidades empírica e extrair dessa distribuição a amostra da reputação de 25.000 usuários que você precisa. Deve-se observar que o processo de geração dessa amostra será considerado válido apenas se a amostra gerada mantiver as mesmas propriedades dos dados brutos. Para atingir esse objetivo, siga as instruções abaixo:

- Implemente um script R que gera a distribuição de probabilidades empirica dos dados da coluna *reputation*. O script também deve extrair dessa distribuição uma amostra contendo a reputação de 25.000 usuários e salvá-la em um arquivo no formato *.txt, com cada usuário em uma linha do arquivo. Para permitir a análise dos dados, o script também deve: (i) gerar uma figura em formato *.png contendo um único gráfico FDA com os dados brutos e a amostra gerada; (ii) imprimir os valores: min, max, média, mediana, 1º e 3º quartis. **Dica**: estude com cuidado o help das funções *sample*, *ecdf* e *hist* no R. Essas funções podem tornar esse exercício mais fácil.

- Analise as informações estatísticas obtidas com a execução do script e comente se a nova amostra mantém as propriedades do dados brutos. Nessa análise, estabeleça comparações entre as estatísticas dos dois conjuntos de dados. Se você conhece algum teste estatístico que permite verificar a similaridade entre dois conjuntos de dados, comente sobre ele.

4) A distribuição **normal** é famosa por descrever diversos comportamentos na natureza. Por outro lado, estudos mais recentes têm mostrado que muitos comportamentos em sistemas computacionais e em redes sociais podem ser descritos por uma distribuição **lognormal**, por exemplo: o tamanho de arquivos na web, o intervalo entre sessões de usuários no Orkut, a popularidade de mensagem no Twitter, de notícias no Digg e de arquivos em comunidades BitTorrent. Neste contexto, coloque em perspectiva as diferenças entre uma distribuição normal e uma distribuição lognormal. Para isso, siga as instruções abaixo:

- Implemente um *script* R que permite gerar um conjunto de dados que seguem uma distribuição normal com uma média e um desvio padrão recebidos como parâmetro. O *script* deve converter os dados normais em lognormais e gerar as FDPs de ambos os dados em um único gráfico. As FDPs devem ser devidamente identificadas e salvas em um único arquivo em formato *.png.

- Utilizando o seu *script*, dada uma distribuição normal padrão, gere a lognormal correspondente e suas respectivas FDPs. Comente sobre os principais impactos da transformação de uma distribuição normal em uma lognormal em termos do formato do gráfico FDP, concentração dos dados, média e desvio padrão.