Mineração de Dados **Nome:** _____
2 semestre de 2016
Relatório 6 - Otimização de parâmetro
19/10/2016
André C. P. L. F. de Carvalho
Victor H. Barella

O objetivo desta prática é otimizar o parâmetro k do *k-means* e gerar um agrupamento com o k escolhido. Espera-se que o aluno consiga aplicar o *k-means* com diversos ks de forma inteligente, analisar os resultados em uma medida para agrupamentos e gerar um arquivo com o agrupamento desejado.

Para esta prática você vai precisar:

- RapidMiner
- Editor de texto com suporte para figuras (ex: Microsoft Office Word, LibreOffice Writer, LaTex)
- Base de dados sobre vias metabólicas

A seguir são apresentados 4 tópicos: entendendo a base, preparação, otimização de parâmetro e agrupamento final.

O relatório deve estar no formato pdf para submissão. Ao final da atividade o aluno terá uma base com exemplos agrupados. Esta base também deverá ser submetida e deverá estar no formato csv.

Em Aprendizado de Máquina, utilizamos técnicas de agrupamento quando desejamos encontrar alguma relação entre os exemplos de uma base de dados. Esta não é uma tarefa trivial. O desempenho da maioria das técnicas é extremamente dependente dos parâmetros que são escolhidos. A técnica *k-means*, por exemplo, define o número de conjuntos produzidos através do parâmetro k. Nesta prática, você irá aplicar ferramentas que auxiliarão na escolha do melhor k para um casa específico.

1. **Entendendo a base**: A fim de entender melhor a base de dados, responda os seguintes questionamentos
   - Qual o contexto em que a base de dados se encontra?
   - Quantos exemplos e quantos atributos a base possui?
   - O que cada instância/exemplo representa?
   - O que os atributos representam? Quais o(s) atributo(s) alvo(s)? Quais os atributos preditivos?

2. **Preparação**: A base originalmente possui mais de 50000 exemplos. Para esta prática, se torna inviável processar todos os exemplos.

   - Faça uma amostragem dos dados.
   - Remova os atributos que variam pouco (por exemplo, atributos com desvio menor que 0.1)

3. **Otimização de parâmetros**:

   - Otimize o parâmetro k. Para isso você pode utilizar o operador *Loop Parameters*
   - Escolha uma medida de avaliação apropriada para selecionar o melhor agrupamento gerado.

4. **Agrupamento final**:

   - Utilize o k escolhido para gerar o agrupamento final

   Utilize o operador *Write csv* para salvar a base de dados agrupada. Ela deverá ser enviada junto com o relatório.