MCC - DC/UEL - Tópicos Especiais em Reconhecimento de Padrões
Primeiro Semestre de 2016 Avaliação 1 **03/05/2016**
Nome: _____
Prof. Dr. Sylvio Barbon Jr

Projeto de Consistência de Dados

Descrição "Consistent data are technically correct data that are fit for statistical analysis. They are data in which missing values, special values, (obvious) errors and outliers are either removed, corrected or imputed. The data are consistent with constraints based on real-world knowledge about the subject that the data describe. Consistency can be understood to include in-record consistency, meaning that no contradictory information is stored in a single record, and cross-record consistency, meaning that statistical summaries of different variables do not conflict with each other. Finally, one can include cross-dataset consistency, meaning that the dataset that is currently analyzed is consistent with other datasets pertaining to the same subject matter. In this tutorial we mainly focus on methods dealing with in-record consistency, with the exception of outlier handling which can be considered a cross-record consistency issue. The process towards consistent data always involves the following three steps.

- Detection of an inconsistency. That is, one establishes which constraints are violated. For example, an age variable is constrained to non-negative values.
- Selection of the field or fields causing the inconsistency. This is trivial in the case of a univariate demand as in the previous step, but may be more cumbersome when cross-variable relations are expected to hold. For example the marital status of a child must be unmarried. In the case of a violation it is not immediately clear whether age, maritalstatus or both are wrong.
- Correction of the fields that are deemed erroneous by the selection method. This may be done through deterministic (model-based) or stochastic methods."

1. (10 points) Considere as seguintes regras:
   - As espécies devem estar entre: setosa, versicolor ou virginica;
   - Todos os atributos devem ser positivos;
   - O "petal length" deve ter pelo menos duas vezes o "petal width";
   - O "sepal length" não pode ultrapassar "30 cm";

   Utilizando o dataset irisModificado.csv, determine:

   1. Determine quantas amostras não seguem as regras e Plote os resultados totais e por classe;
   2. Qual o percentual dos dados não apresenta erros?
   3. Quantas instâncias apresentam o "sepal length" fora do padrão?
   4. Encontre os outliers e identifique-os;
   5. Correções:
      (a) Ajuste os valores negativos e justifique;
      (b) Substitua os valores ausentes (NA) usando uma regra adequada e justifique;
   6. Imputações:
      (a) Use o *"kNN" imputation* (VIM) para imputar valores ausentes;
      (b) Use o *hotdeck imputation* para imputar o Petal.Width ordenando o dataset pela classe.

Leitura recomendada:
de Jonge, Edwin, and Mark van der Loo. "An introduction to data cleaning with R." Statistics Netherlands, The Hague (2013): 53.