R Notebook

Purpose

This file prepares data sets for predictive analysis. Several machine learning tools are evaluated against their ability to predict normalized dependent variables (y_i) based on event characterisitics (x_i). Normalization is done by dividing the bulk isotope signatures or bulk mass measured in soil transects at of the nearest soil sampling period.

Dependent (normalized) variables include:

- Signatures
- SM masses
- TPs masses
- MEL_{SM} (Mass equivalent loads for S-metolachlor)

Independent variables include:

- Event duration ($t_f - t_i$)
- Volume discharged ($\sum_{i=1}^{N} Q_i \cdot dt_i$, N: no. of measurements within the event)
- Average discharge ($\sum_{i=1}^{N} Q_i / N$)
- Kurtosis (Tailed-ness)
- Skweness (symmetry extent)

Imports:

- **.csv**

Generates:

- ---

Required R-packages:

```
library("plyr")
library("dplyr")
```

Working directory

```
# setwd("D:/Documents/these_pablo/Alteckendorf2016/R")
# setwd("/Users/DayTightChunks/Documents/PhD/Routput/Alteck/R")
# setwd("D:/Documents/these_pablo/Alteckendorf2016/00_TransparencyFolder")

getwd()
```

```
## [1] "D:/Documents/these_pablo/Alteckendorf2016/HydrologicalMonitoring"
```

Models to evalaute

CART - Regression Trees (RTs)

- In regression trees the output of each leaf is a real number (dependent variable), namely the average of the value at that leaf
- Unlike linear regression models, RTs can capture non-linearities
- Compare fitness based on “cp” parameter
- Apply cross-validation