

# Intelligente Datenanalyse

Prüfung: Feuer im Naturpark

Datum: 13. Juni 2016

Gerrit Gruben  
Prof. Tobias Scheffer

Die Projektaufgabe ist Teil der Prüfung *Intelligente Datenanalyse*. Jede Aufgabe soll durch einen Studenten selbstständig bearbeitet und die Lösung innerhalb der mündlichen Prüfung vorgestellt werden. Ein Ausdruck des Python-Programmcodes und der Ergebnisse in Form eines Diagramms, Tabelle o.ä. werden vorausgesetzt; die Art der Präsentation der Ergebnisse ist dem Studenten freigestellt.

## Problemstellung

Die Verwaltung des Naturparks *Montesinho* im Nord-Osten Portugals möchte basierend auf Wetterdaten und des Feuer-Wetter-Indexes (FWI) eine Vorhersage über mögliche Waldbrände treffen. Das Ziel ist es die Fläche und damit die Intensität eines bevorstehenden Waldbrandes möglichst frühzeitig zu erkennen um die Gefahr welche von dem Feuer ausgeht einschätzen zu können. Zu diesem Zweck wurden die Daten von 517 Bränden gesammelt. Die Attribute sind

- X (X-Koordinate des Brandes im Park: 1 bis 9)
- Y (Y-Koordinate des Brandes im Park: 2 bis 9)
- month (Kalendermonat: „jan“ bis „dec“)
- day (Wochentag: „mon“ bis „sun“)
- FFMC (FFMC-Index des FWI-Systems: 18,7 bis 96,2)
- DMC (DMC-Index des FWI-Systems: 1,1 bis 291,3)
- DC (DC-Index des FWI-Systems: 7,9 bis 860,6)
- ISI (ISI-Index des FWI-Systems: 0,0 bis 56,1)
- temp (Temperatur in Grad Celsius: 2,2 bis 33,3)
- RH (Relative Luftfeuchtigkeit in %: 15 bis 100)
- wind (Windgeschwindigkeit in  $km/h$ : 0,4 bis 9,4)
- rain (Niederschlagsmenge in  $mm/m^2$ : 0,0 bis 6,4)
- area (Verbrannte Waldfläche in Hektar: 0,09 bis 1090,84)

Sie wurden damit beauftragt aus den gegebenen Daten ein Vorhersagemodell zu entwickeln welches möglichst genau die verbrannte Waldfläche angibt.

## Aufgabe

Lesen sie die Daten in Python ein und führen sie eine Datenvorverarbeitung – insbesondere eine geeignete Normierung der Attribute – durch. Beispielsweise ist das Ziellabel *area* sehr ungleichmäßig verteilt, sodass z.B. die Transformation  $area' = \log(1+area)$  angebracht ist. Identifizieren und implementieren sie ein geeignetes Lernverfahren in Python. Bestimmen sie die Modellparameter; trainieren und evaluieren sie das Modell. Überlegen sie wie eine triviale Baseline aussehen könnte. Geben sie den RMSE (*root mean square error*) ihres Modells und der Baseline für 10-Fold Cross-Validation an. Begründen und dokumentieren sie kurz alle durchgeführten Schritte.