Prof. Dr. Katharina Morik,
JProf. Dr. Uwe Ligges
Dipl.-Inform. Hendrik Blom,
M. Sc. Nadja Bauer,
Daniel Horn

Dortmund, 10.05.13
Abgabe: bis Fr, 17.05., 23:59 Uhr an
hendrik.blom@tu-dortmund.de

Übungen zur Vorlesung

Wissensentdeckung in Datenbanken

Sommersemester 2013

Blatt 4

Aufgabe 4.1 (4 Punkte)

1. Welche Fehlerarten kennen sie? Schreiben sie die Namen und Formeln auf und eine Eigenschaft.
2. Was ist das Optimierungsproblem bei linearen Modellen? Erstellen sie ein Beispiel und bilden Sie die partielle Ableitung zur Herleitung von $\hat{\vec{\beta}}$. Leiten sie $\hat{\vec{\beta}}$ für ihr Beispiel her.

Aufgabe 4.2 (6 Punkte)
Die HeidelburgCement Ag ist um die Qualität ihres Zements besorgt. Damit eine gleichbleibende Qualität des Zements gewährleistet werden kann, werden sie gebeten ein Modell über die "ConcreteCompressiveStrength" des bisher hergestellten Zements zu erstellen.

Die Daten finden sie unter(Original:http://archive.ics.uci.edu/ml/):

`http://www-ai.cs.uni-dortmund.de/LEHRE/VORLESUNGEN/KDD/SS13/Concrete_Data.csv`

1. Laden sie die CSV-Datei herunter und importieren sie die Daten in ihr *RapidMiner* Repository!
2. Laden sie die Daten in ein *RapidMiner* Experiment und starten sie das Experiment. Betrachten sie die Daten in der Ergebnisansicht von *RapidMiner* und überlegen sie sich, welche Attribute am besten für die Modellierung geeignet sind.
3. Kürzen sie die Attributnamen geeignet ab.
4. Die Daten enthalten fehlende Werte und nicht sinnvolle Daten (<0). Führen sie eine Vorverarbeitung der Daten durch, indem sie fehlende Werte geeignet behandeln und eine Plausibilitätsprüfung durchführen. Begründen sie bitte ihr Vorgehen.

5. Führen sie die Modellierung mit der linearen Regression und einem anderen Operatoren für die Regression mit einer 10-fachen Kreuzvalidierung durch und erklären sie kurz die Unterschiede der Algorithmen.

6. Bewerten die die Güte des Modells (Performance) mit 3 Gütemaßen und erklären sie diese kurz. Stellen sie die Ergebnisse geeignet dar.

7. Wenden sie das gelernte und kreuzvalidierte Modell auf alle Daten an und erstellen sie einen Plot (Wahrer Wert zu Prognose).

Hinweis: Das sehr gute Buch "The Elements of Statistical Learning" ist Online als PDF unter `http://www-stat.stanford.edu/~tibs/ElemStatLearn/` verfügbar.
Viele Originalpaper sind als PDF über `http://scholar.google.de/` verfügbar. Dazu müssen sie sich aber im Netz der Universität befinden. Für Informationen zum VPN schauen sie bitte unter `http://www.itmc.tu-dortmund.de/de/dienste/netz-und-server-dienste/netzwerk-zugaenge/vpn.html`.