WPF Data Mining praktisch – DMC

Profs. Dr. Wolfgang Konen, Thomas Bartz-Beielstein **WS09, Beginn 13.10.2009**

Themenvorstellung im Detail

Gliederung / Termine Beginn: 12.15 Uhr

P: Projekt, S: Seminar (mit Demonstrationen), V: Vortrag Dozent, Ü: Übung alle

- Terminänderungen vorbehalten! -		Wieviele / Wer
13.10.09	Startup-Veranstaltung, evtl. Projekte verteilen	**Kg / Bb**
V/P 20.10.09	Einführung Computational Intelligence Design of Experiments I: Helicopter Design Experimente	**Bb**
V/P 27.10.09	Design of Experiments II: Helicopter Design Auswertung. Erste Analysen mit R (Skripte werden bereitgestellt)	**Bb**
	Projektarbeit: R-Einführungen lesen, z.B. [Klar06], S. 1-36, Übungszettel bearbeiten (vorher Kurzpräsentationen verteilen). Mit R vertraut machen. Helikopter Bericht schreiben	alle
V/Ü/S 10.11.09	R Tutorial / Übungen - Grundlagen, DataFrame-Manipulation, Function-Übung - Einleseübung - WalkThrough DM-Template „IRIS“, RF als Black Box	**Bb** (Kg) alle
V 17.11.09	DM: Einführung, CRISP-DM Datenvorverarbeitung: Aktivierung, Erste Schritte	**Kg** (Bb)
V/Ü 24.11.09	Modellierung: Taxonomie Entscheidungsbäume – das Prinzip Das empirische Lernproblem (Trainings- vs. Testfehler) Metalerner	**Kg**/Bb
	Projektarbeit: Einführung in Random Forest lesen, Aufgabe Random Forest bearbeiten (vorher Kurzpräsentationen verteilen)	alle
Ü/S 08.12.09	Random Forests – Ideen, Begriffe, Fragen	**Kg** (Bb) alle
V 15.12.09	Evaluierung: - Trainings- und Test-Set, Overfitting - Klassifikation: Fehlermaße, Kostenmatrix, Lift Chart, Precision, Recall - Regression Basics	**Bb** (Kg)
15.12.09	**Milestone 1**: Zwischenbericht abgeben	alle
	Projektarbeit: alle versuchen, Proj. DMC 2007 zu lösen.	alle
P 12.01.10	Projekt DMC 2007: Vorstellung der Ergebnisse. Durch Austausch/Kombination best-of-breed optimierte	alle

	Ergebnisse. Diskussion von Fragen/Problemen	
V/P 19.01.10	V FIWA / SOMA / Anwendungen	**Bb / Kg**
	Projektarbeit:Abschlussbericht	alle
26.01.10	**Milestone 2**: Abschlussbericht abgeben	alle
02.02.10	Mündl. Prüfungen, Abschlussveranstaltung, Feedback	Kg/Bb alle
Ergänzende Themen:		
V/P	Fehlende Daten: Strategien zur Beseitigung Anwendung: Aufgabe Missing Values - Ü1 Iris-Datensatz mit künstlich 20% Fehlstellen - Ü2 DMC2007-Datensatz mit Fehlstellen	**Kg** (Bb) alle
P	Vergleichende Umsetzung mit RapidMiner	
V/Ü	Naive Bayes als Klassifikator - in R als Übung implementieren	Kg/Bb alle

Der gelb unterlegten Blöcke kennzeichnen Projektarbeiten der Studierenden, die entweder als Work@home oder im Labor mit Fragemöglichkeit an die Dozenten gemacht werden
Ab dem grün unterlegten Block: Optionale Zusatzthemen, werden eingebunden, wenn es die Zeit erlaubt.

P: Projekt, S: Seminar (mit Demonstrationen), V: Vortrag Dozent, Ü: Übung

Startup-Veranstaltung

Typ: Vortrag und Demonstration

1. Einleitung, Beispiele vorstellen
2. Vorstellen der Unterrichtsformen
    1. Vorlesung
    2. Übungen (die alle bearbeiten, z.B. R-Code erstellen/modifizieren, Austesten an konkreten Daten)
    3. Projektarbeiten der Studierenden (jeder versucht zu lösen, Zusammenbringen der Lösungen im Team): Vorstellung Ergebnisse
3. Literatur:
    1. [WittenFrank01 bzw 05] bestellt für Abt.-Bibl. GM,
    2. ddmobook.pdf (Skript im ILIAS)
4. R, Tinn-R: OpenSource, auch installiert auf Rechnern in Labor.
5. Bewertungskriterien
    1. Zwischenbericht zur Übungs- und Projektarbeit (30%)
    2. Abschlussbericht zur Übungs- und Projektarbeit (30%), mit Vortrag
    3. Glossareinträge: jeder stellt **2 Glossarbegriffe** in ILIAS (s.u.) einbauen. Die Begriffe mit Namen oder Matrikelnummer identifizierbar machen (10%)
    4. Mündliche Prüfung (30%)
6. Bewertet wird (zu 1.-3.)
    (a) Verständlichkeit der Darstellung (für "Aussenstehenden") und Form der schriftlichen Berichte / der Glossarbegriffe
    (b) mündl. Beteiligung, Kurzvorträge

(c) Extrapunkte, wenn auch weiterführende Untersuchungen durchgeführt und dokumentiert sind (hier lassen wir uns gerne von Ihrer Kreativität überraschen)

7. Organisatiorisches:
   1. Jeder, der teilnehmen will, bitte **bei ILIAS anmelden** http://ilias.fh-koeln.de/start.php (Fakultät 10, WPF beitreten **UND** darin der **Gruppe WS09** beitreten) >> Mailingliste
   2. Bereitstellen der Ausarbeitung im ILIAS durch Upload unter **Gruppe WS09**.
   3. (Übrigens: Löschen der selbst hochgeladenen Dateien geht in ILIAS über "Administrationsmenüs AN")

Die Berichte/Zwischenberichte bestehen aus schriftlicher Ausarbeitung, die **zum Milestone fertig sein muss** und (b1) in Papierform abgegeben wird (einmal für uns) und (b2) elektronisch hinterlegt wird (ILIAS-Server). Dies ist eine wichtige Voraussetzung.

8. Vorstellen der Kapitel und Themen
   1. vorläufiger Terminplan
   2. Kommentiertes Literaturverzeichnis

Aufgabe Helicopter Design

Eine ausführliche Beschreibung finden Sie im Skript ddmobook.pdf (Kapitel 2), das Sie im ILIAS zur Verfügung gestellt bekommen. Ziel dieser Aufgabe ist **die systematische Planung und Durchführung von Experimenten und den zugehörigen Datenanalysen**. Diese Aufgabe wird in der Fachliteratur als "Design of Experiments" bezeichnet und findet in der Praxis bei der Bewältigung komplexer Probleme Anwendung, Wir werfen auch einen Blick auf Techniken, die zum Tunen von Algorithmen dienen, so dass wir möglichst wenige Läufe durchführen müssen.
Diese Aufgabe weist eine Besonderheit auf: Die Daten für diese Aufgabe werden von Ihnen selbst erzeugt, indem Sie die Flugdauer eines Helikopters optimieren.

Aufgabe Random Forests (RF)

Im Vorlesungsteil wird das Grundprinzip der Entscheidungsbäume und das Grundproblem des empirischen Lernens (Trainings- vs. Testmengenfehler) vorgestellt.
Bearbeiten Sie mit der Random-Forest-Beschreibung unter http://stat-www.berkeley.edu/users/breiman/RandomForests/cc_home.htm (sowie fallweise auch mit eigener Recherche zu Random Forests) in Ihrer Home Session folgende Übungs- / Seminaraufgaben:

1. Installieren Sie auf Ihrem Rechner das R-Paket ***randomForest*** und machen Sie sich über das Help-Manual mit seiner Nutzung vertraut.[1]
2. Erläutern Sie das **Grundprinzip** des Random Forest! Wie schafft es dieses Verfahren, das Übertrainieren zu vermeiden? Erläutern Sie den Begriff **Bootstrap**! Erläutern Sie die Begriffe **OOB-Prediction** und **OOB-Error**! Finden Sie heraus, wie man in der R-Implementierung diese Größen für einen Datensatz (z.B. IRIS) ermitteln kann.
3. Erklären Sie den Begriff **Importance** im RF-Kontext. Ermitteln Sie (in R) die Importance der Variablen x_2 … x_5, wenn Sie künstlich einen Datensatz

[1] Man installiere *randomForest* von CRAN-Download Mirror: R – Pakete – Installiere Pakete…, aber wähle dabei den richtigen **Download Mirror Austria**, denn nur der hat eine sehr vollständige Liste aller Pakete, insbes. auch *randomForest*.

$$y \sim 0.1\,\varepsilon + 0.2\,x_2 + 0.3\,x_3 + 0.4\,x_4 + 0.5\,x_5,$$

worin x_2 … x_5 standardnormalverteilte (unabhängige) Input-Größen sind und ε ein standardnormalverteilter Rauschterm ist, erstellen. Entspricht das Ergebnis Ihren (vorab notierten) Erwartungen?

4. Machen Sie mit ***randomForest*** eine Vorhersage für die CPU-Daten [Witten&Frank01, Tab. 1.5]. Stellen Sie nun 33% der Daten als Testmenge an die Seite und trainieren Sie mit den verbleibenden 66%. Macht ***randomForest*** eine korrekte Vorhersage über den Testset-Fehler RMSE[2]? (evtl. *mehrfach* mit verschiedenen Testmengen wiederholen, evtl. große Output-Werte (Outlier) ausschließen. Wieso?). Welche Methode der Fehlerberechnung würde eine viel zu optimistische Schätzung liefern (im Vgl. zum realen Testset-Fehler)?

Jeder (einzeln oder in 2er-Gruppen) bearbeitet jeden der Punkte 1.-4. und bereitet dazu einen kurzen Vortrag (je 5-10 min, max. 3 Folien je Punkt) vor. Im WPF wird dann zu jedem der Punkte ein Studierender (eine Gruppe) nach vorne gebeten. Danach können die anderen jeweils noch fehlende Details einfügen, Fragen diskutieren, usw.

Projekt DMC 2007

- Besorgen Sie sich die Aufgabenstellung, die Daten und die Ergebnisse des DMC Challenges 2007 (entweder von www.data-mining-cup.de oder über ILIAS)
- Machen Sie sich mit den Daten vertraut: Typus, Wertebereich, Visualisierung, fehlende Werte, Statistiken, Ausreisser oder Fehler?
- Kontrollieren Sie die eingelesenen Daten auf Vollständigkeit und Plausibilität
- Überlegen Sie, ob Sie sinnvolle abgeleitete Inputgrößen definieren können.
- Schreiben Sie ein R-Skript, das die Aufgabe löst. (Sie können sich dabei an dem IRIS-Template orientieren)
- Wieviele Punkte erreichen Sie?
- Beachten Sie: Die reale Klassifikation der Testdaten darf nur ganz am Schluss bei der Evaluation verwendet werden, zu nichts anderem.

Projekt DMC 2006

- Gleiche Anleitung wie Projekt DMC 2007.

- Worin unterscheidet sich die Aufgabenstellung, die Datenlage vom vorherigen Projekt?

Aufgabe Missing Values

1. Nehmen Sie den Datensatz IRIS oder CPU. Löschen Sie in einer Spalte 20% der Daten (zufällige Auswahl). Wieviel schlechter wird die Klassifikation?
2. Testen Sie verschiedene Strategien zur Ersetzung der fehlenden Werte. Welche funktioniert am besten? Sind die Ergebnisse von der Zufallsauswahl und von der Wahl der Spalte abhängig?
3. Löschen Sie nun zufällig 20% der Daten in jeder Spalte. Wieviele Datensätze sind noch vollständig? Welche Strategie zur Ersetzung funktioniert jetzt am besten?
4. (Machen Sie eine ähnliche Untersuchung auf dem Datensatz DMC 2007)

Vergleichende Umsetzung mit Rapid Miner

[2] RMSE = „root mean squared error“ = Wurzel des MSE, schwankt weniger stark als MSE