Übung 16: Weka

1. Kopieren Sie die Datei `/home/public/vorkurs_ss19/playoutside.arff` in Ihr Vorkurs-Verzeichnis. Der Datensatz enthält verschiedene Wetterbedingungen zusammen mit der Angabe, ob man draußen spielen kann oder nicht.

2. Schauen Sie sich die Datei an. Die Datei enthält die Wetterdaten Vorhersage (outlook), aktuelle Temperatur (temperature, in Fahrenheit), Luftfeuchtigkeit (humidity) und Windverhältnisse (windy, TRUE oder FALSE, also ein boolescher Wert). Als letztes enthält die Datei außerdem das Datum "play", das angibt ob man spielen kann oder nicht. Es handelt sich also um *a*nnotierte Daten.

   Möchte man einen Classifier produktiv einsetzen (um echte Daten zu klassifizieren) braucht man natürlich auch unannotierte Daten, die man klassifizieren möchte. Wir werden später solche Daten erstellen. Wichtig ist, dass die annotierten und unannotierten Datensätze die gleichen Daten enthalten – abgesehen vom letzten Datum, das die Klasse beschreibt.

3. Wir werden nun zunächst einen geeigneten Algorithmus suchen. Dazu verwenden wir ausschließlich die annotierten Daten, da wir damit auch gleich evaluieren können, wie gut der Algorithmus funktioniert hat. Wichtig ist generell, dass man nicht die gleichen Daten zum trainieren und testen verwendet (Weka sorgt dafür, dass uns das nicht passiert).

   Aktivieren und starten Sie Weka 3.7.7 (zu finden im Verzeichnis `/resources/stat_ml/weka-3.7.7`). Als erstes öffnet sich der GUI Chooser. Wählen Sie den Explorer, öffnen Sie die Datei. Weka zeigt Ihnen zunächst eine genauere Analyse der Daten an (welche Werte kommen wie oft vor etc.). Schauen Sie sich die Daten genau an.

4. Wählen Sie dann den Reiter "Classify" und probieren Sie einige Classifier aus. Einige können nicht mit numerischen Werten umgehen, andere nicht mit nominalen. Entsprechende Fehlermeldungen weisen Sie darauf hin[1]. Lassen Sie die Test options auf 10 fold cross-validation eingestellt. Merken Sie sich den Classifier, der die besten Ergebnisse liefert.

5. Erstellen Sie nun eine Datei in Ihrem Vorkurs-Verzeichnis, die z.B. `Daten.arff` heißt. Darin sollen neue annotierte Daten stehen, die Sie sich selbst ausdenken dürfen. Es reicht, wenn ein bis zwei Datensätze drinstehen. Wichtig: Sie müssen den Header (alles vor der `@data`-Markierung) in die Datei schreiben (oder kopieren).

6. Wechseln Sie wieder in das Weka-Fenster und den Reiter "Classify". Wählen Sie nun unter test options "Supplied test set", klicken Sie auf set und sorgen Sie dafür, dass die eben erstelle Arff-Datei zum testen verwendet wird. Klassifizieren Sie mit dem gleichen Classifier wie davor.

[1] Kaum ein Classifier kann mit String-Werten umgehen.

7. Unter “More Options” gibt es weitere Optionen zur Klassifizierung. Aktivieren Sie dort “Output predictions”. Damit werden die tatsächlichen Vorhersagen für einzelne Datensätze ausgegeben. Klassifizieren Sie erneut und schauen Sie sich die einzelnen predictions an.

8. Weka lässt sich auch in der Kommandozeile verwenden. Wechseln Sie wieder in Ihr Terminalfenster. Verwenden Sie den kompletten (qualifizierten) Namen des Classifiers als Java-main-class (z.B. `:~$ java weka.classifiers.trees.J48`). Wenn Sie keine Optionen angeben, bekommen Sie eine Hilfeseite angezeigt, die die möglichen Optionen listet. Versuchen Sie die die gleiche Ausgabe wie auch in der GUI zu bekommen.