Übungsblatt 3

Machine Learning (WS 16/17) 17. November 2016

Stefan Edelkamp

Sämtliche Aufgaben sind von der Gruppe selbstständig zu lösen. Die Verwendung von Hilfsmitteln und Quellen außerhalb der Vorlesungsunterlagen gilt es in expliziter Weise zu dokumentieren.
Abgabe ist am Donnerstag, den 1.12.2016 in der Übung.

1 Begriffsdefinitionen

1. Was ist der Bias bei Hidden Markov Modellen. Gegen Sie Beispiele an, wann dieser Bias zutrifft und wann nicht. (5 P)
2. Warum ist *The inductive learning hypothesis* (s.u.) aus der Vorlesung falsch? Geben Sie ein Gegenbeispiel! (10 P)

> **The inductive learning hypothesis** Any hypothesis found to approximate the target function well over a sufficiently large set of training examples will also approximate the target function well over other unobserved examples.

2 Concept Learning

Leider sollen Sie in Ihrem neuen Job auch Lieferungen ausliefern. Dazu benötigen Sie ein Fortbewegungsmittel. Um ein Konzept zu erstellen, welches Fortbewegungsmittel am geeignetesten ist, führen Sie einige davon Ihren Kollegen vor. Diese beantworten Ihre Fragen eine Weile klaglos, danach sind Sie aber nicht mehr bereit weitere Exemplare zu erklären.

Jetzt ist es an Ihnen das allgemeine Konzept eines für einen Fahrradkurier geeigneten Fortbewegungsmittels zu finden. Konzepte können in einer Notation angegeben werden, die bestimmt, welche Werte für das Konzept gültige Belegungen sind (wie z.B. $\{?, ?, ?\}$). Wir untersuchen die Attribute Räder, Antrieb, Farbe und Aussehen. Die möglichen Belegungen der einzelnen Attribute ist in der folgenden Tabelle angegeben:

Variable	Werte
Räder	$\{Eins, Zwei, Drei, Vier\}$
Antrieb	$\{Pedal, Pferd, Motor\}$
Farbe	$\{Rot, Grün, Blau\}$
Aussehen	$\{Schön, hässlich, wunderschön\}$

1. Beschreiben Sie in eigenen Worten, was das Konzept $\{Zwei, Pedal, ?, ?\}$ bedeutet. Um welchen Fahrzeugtyp handelt es sich dabei wahrscheinlich? (3 P)
2. Geben Sie ein Konzept in der beschriebenen Notation für ein Trike an und erklären Sie was es bedeutet! (Bitte beachten Sie, dass ein Trike sowohl motorisiert als auch nicht motorisiert sein kann.) (2 P)
3. Wieviele unterschiedliche Referenzdaten (Beispiele) sind durch die angegebenen Wertebereiche möglich (Bitte beschreiben Sie das Ergebnis kurz! Nicht einfach eine Zahl hinschreiben!)? (5 P)
4. Wieviele Hypothesen sind durch die angegebenen Wertebereiche möglich (wir gehen davon aus, dass der Hypothesenraum mit den aus der Vorlesung bekannten Constraints, biased ist)? Beschreiben Sie das Ergebnis! (5 P)

5. Wie groSS wird der Hypothesenraum ohne Bias (Bitte beschreiben Sie das Ergebnis!)? (5 P)

Räder	Antrieb	Farbe	Aussehen	Geeignet
Vier	Pferd	Rot	schön	Nein
Zwei	Pedal	Grün	wunderschön	Ja
Vier	Motor	Blau	hässlich	Nein
Eins	Motor	Grün	schön	Nein
Zwei	Pedal	Blau	wunderschön	Ja
Zwei	Motor	Rot	wunderschön	Ja

Tabelle 1: Trainingsdaten

6. Für die Suche nach einem korrekten Konzept kann eine "Genereller-als"($>_g$) Beziehung zwischen den Hypothesen ausgenutzt werden. Ordnen Sie folgende Hypothesen anhand der $>_g$ Relation: $\{Eins, ?, ?, ?\}$, $\{?, ?, ?, ?\}$, $\{?, ?, , ?\}$, $\{?, Pedal, Blau, ?\}$! (5 P)

7. Verwenden Sie den $FIND-S$ Algorithmus um die speziellste Hypothese aus den Daten (Tab. 1) zu generieren. Geben Sie dabei jeden Schritt an und beginnen Sie mit der speziellsten Hypothese $\{, , , \}$! (10 P)

8. Verwenden Sie den $CANDIDATE-ELIMINATION$ Algorithmus auf den selben Daten (Tab. 1) um den Version-Space aufzubauen. Protokollieren Sie jeden Ihrer Schritte! Welche Fahrzeugtypen werden demnach eigentlich von den Kurier-Radfahrern bevorzugt? (20 P)

3 Naive Bayes

Für einen Kommilitonen der es leider in seinem Studium verpasst hat, die Machine Learning Veranstaltung zu besuchen, muss ein **Klassifikator Klassifikator** erstellt werden. Dieser Klassifikator Klassifikator soll klassifizieren, welchen Klassifikator man bei einer gegebenen Problemstellung einsetzen sollte. Da diese Entscheidung natürlich nicht pauschal getroffen werden kann, soll ein probabilistisches Modell verwendet werden.

1. Lernen Sie einen Naive Bayes **Klassifikator Klassifikator**. Dieser soll klassifizieren, welcher Klassifikator für eine durch die Attribute Wertebereich, Trainingsdatenumfang und Rauschen beschriebene Problemstellung häufig verwendet wird (best practice). Die Referenzdaten werden in `ml/code` unter Dateien als CSV (Comma Seperated Value) Datei hinterlegt. Sie müssen aus dieser Datei die Attribute bestimmen, die Wertebereiche der Attribute und die Wertebereiche der Klasse bestimmen. Es wird angenommen, dass die letzte Spalte die Klasse enthält. Nachfolgend müssen die relevanten Wahrscheinlichkeiten aus den Referenzdaten bestimmt werden. Eine weitere CSV enthält Anfragedaten (ebenfalls in `ml/code`), die klassifiziert werden müssen. Dazu soll bei jeder Klassifikation die Wahrscheinlichkeitsverteilung über die Klassen angegeben werden. (25 P)