Übungsblatt 2

Machine Learning (WS 16/17) 3. November 2016

Stefan Edelkamp

Sämtliche Aufgaben sind von der Gruppe selbständig zu lösen. Die Verwendung von Hilfsmitteln und Quellen außerhalb der Vorlesungsmaterialien gilt es in expliziter Weise zu dokumentieren.
Abgabe ist am Donnerstag, den 17.11.2016 im Tutorium.
Der Source-Code muss dokumentiert in Java vorliegen und ist am Abgabedatum an edelkamp@tzi.de zu schicken.

1 Begriffsdefinitionen

1. Beschreiben Sie in eigenen Worten den Unterschied zwischen Klassifikation und Clustering. (2 P)
2. Beschreiben Sie in eigenen Worten die Begriffe *Kernobjekt*, *Dichte-Erreichbar* und *Dichte-Verbunden*. (3 P)
3. Was ist Lazy Learning? Fällt das Konzept in den Bereich Klassifikation und / oder Clustering? (2 P)

2 DBSCAN

Noch letzte Woche waren Sie der Chef eines Fahrradkurierunternehmens. Leider mussten Sie Ihren Chefposten jedoch abgeben, da die Kurierfahrer aufgrund einiger sehr weiter Strecken ihre Arbeit eingestellt haben. Der neue Chef hatte ein Herz und stellte Sie als Fahrradkurier an. Da Sie sich wieder bei ihren Kollegen beliebt machen wollen, versuchen Sie ihren Chef zu überzeugen, dass er sich von einigen Kunden trennen sollte um weite Wege zu vermeiden. Da ihr Chef ebenfalls sehr bewandert im Bereich *Maschinelles Lernen* ist, kann er überzeugt werden, ein dichtebasiertes Clustering durchzuführen und auf dessen Grundlage eine Entscheidung zu treffen.

1. Implementieren Sie einen Algorithmus zur Berechnung eines *k-Distanzdiagramms*. Verwenden Sie die Formeln aus der Vorlesung um ein geeignetes initiales k und *MinPts* zu finden. Berechnen Sie ausgehend davon mindestens fünf weitere *k-Distanzdiagramme* mit unterschiedlichen k. Bestimmen Sie auf Grundlage der Diagramme geeignete Werte für ϵ und $MinPts$. Markieren Sie das entsprechende Objekt im Diagramm. Begründen Sie Ihre Entscheidung. Verwenden Sie die Fahrradkurierdaten des ersten Ubungszettels (10 P)
2. Implementieren Sie den *DBSCAN*-Algorithmus. Wenden Sie diesen auf die Fahrradkurierdaten des Übungszettels 1 mit den zuvor gefundenen Parametern an. Geben Sie für die Ergebnisse die Anzahl der Cluster, die Kosten für den Chef (basierend auf der Kostenfunktion des letzten Übungszettels) und die Anzahl der Noise Daten an. (15 P)
3. Vergleichen Sie Ihre Lösung durch *DBSCAN* mit der Lösung basierend auf *k-Means* von dem letzten Übungszettel. Geben Sie jeweils mit einer kurzen Begründung an, für wen die jeweilige Lösung von Vorteil bzw. von Nachteil ist (Fahrer, Chef, Kunden). Geben Sie außerdem die Kosten TD und den Silouettenkoeffizienten für Ihrer Lösungen an. (8 P)

3 Decision Tree Learning

Der ID3 Algorithmus mit Information Gain soll verwendet werden um einen Entscheidungsbaum zu trainieren. Dieser soll verwendet werden um ein Personenratespiel (ähnlich dem Akinator, http://de.akinator.com/) zu erstellen. Es soll ein Entscheidungsbaum trainiert werden, der Anhand von 3 Fragen feststellt (nach Geschlecht, Geburtsort und Haarfarbe), um welche Person es sich handelt.

Geschlecht	Geburtsort	Haarfarbe	Person (Klasse)
Frau	Deutsch	Hell	Angela Merkel
Mann	USA	Hell	Bon Jovi
Mann	Belgien	Dunkel	Jean-Claude Van Damme
Mann	USA	Dunkel	Martin Luther King

1. Trainieren Sie einen Entscheidungsbaum mit dem ID3 Algorithmus und verwenden Sie dabei die Information Gain Heuristik. Beschreiben Sie jeden Trainingsschritt mit den zugehörigen Berechnungen (Entropie, Information Gain). (20 P)
2. Welche Personengruppe wird anhand des Entscheidungsbaumes als Angela Merkel klassifiziert? (5 P)
3. Beschreiben Sie unter Verwendung eines neuen Beispiels (eine beliebige Person), wie mit dem trainierten Entscheidungsbaum klassifiziert wird! (5 P)
4. Welche Probleme treten auf, wenn Sie Arnold Scharzenegger oder Pipi Langstrumpf als Testbeispiel verwenden? Wo liegt der Fehler und wie kann er behoben werden? (5 P)

4 Regression

Gegeben sei eine Menge $P = \{(x_1, t_1), (x_2, t_2), (x_3, t_3), \ldots, (x_n, t_n)\}$ von Attribut-Werten $x_i \in \mathbf{R}$, die zu den Zeitpunkten $t_i \in \mathbf{R}$ aufgenommen wurden. Berechnen Sie eine lineare Regressionsfunktion, die die Werte approximiert und dabei den quadratischen Fehler minimiert. Bildlich gesprochen wird einen Gerade $x_i = m \cdot t_i + b + \epsilon_i$ mit zu minimierenden Fehler ϵ_i durch die Punktwolke der Messung gelegt.

1. Finden Sie eine Formel für m und b, in dem Sie $\arg\min_{(m,b)\in R^2} \sum_{i=1}^{n} (x_i - (b + m \cdot t_i))^2$ bestimmen. (5 P)
2. Diskutieren Sie die Grenzen und Möglichkeiten der linearen Regression für den folgenden Datensatz: (5 P)

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1996	138.2	137.5	138.4	148.2	151.8	146.4	141.4	135.4	130.2	123.7	121.7	122.6
1997	123.4	124.3	125.3	126.9	125.4	123.2	121.7	123.2	120.9	119.3	120.1	120.8
1998	120.2	118.3	117.9	115.3	114.7	113.8	111.4	111.1	107.6	105.6	108.5	108.1
1999	108.1	108.3	109.5	112.2	113.7	115.5	112.5	113.1	110.8	109.7	110.6	111.8
2000	112.8	114.0	115.5	117.8	117.2	116.2	116.8	117.1	116.6	116.6	116.5	116.6
2001	117.5	117.5	116.6	115.8	116.9	117.9	117.6	116.1	114.4	113.5	115.1	115.4
2002	115.4	115.2	114.5	113.7	111.7	109.0	107.9	110.9	110.9	111.6	110.9	110.0
2003	109.3	108.8	107.3	107.4	106.1	106.8	109.6	112.2	119.1	121.9	130.4	131.8
2004	136.2	134.0	132.3	132.4	122.4	119.7	115.2	108.3	104.7	103.6	103.9	99.4