Übungsblatt 1

Machine Learning (WS 16/17) 20. Oktober 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 3.11.2016 im Tutorium.
Der Source-Code muss dokumentiert in Java vorliegen und ist am Abgabedatum an edelkamp@tzi.de zu schicken. Nicht lauffähige Programme werden nicht bewertet.

1 Begriffsdefinitionen

1. Beschreiben Sie die Begriffe Precision, Recall und False Positive in eigenen Worten. (3 P)
2. Grenzen Sie die Begriffe Partitionierendes und Hierarchisches Clustering voneinander ab. (4 P)
3. Erklären Sie den Ähnlichkeitsbegriff beim Clustering und geben Sie mindestens jeweils ein Beispiel für eine Distanz, die numerische bzw. kategorische Attribute verwendet. (4 P)

2 k-Means Clustering

Die Firma WIDM (WoIstDerMarsroboter) hat verschiedene Roboter zum Mars geschickt. Sie weiSS allerdings nicht genau, wie viele dort angekommen sind und auch nicht genau wo sich diese auf dem Mars befinden. Ihre Aufgabe ist es die zerstreuten Roboter mit Hilfe von einzelnen, empfangenen Funksignalen zu lokalisieren. Die Firma WIDM stellt Ihnen die Positionen der Funksignale zur Verfügung, diese sind in `ml/code` hinterlegt worden.

1. Implementieren Sie den k-Means Algorithmus in Java. Verwenden Sie hierfür das *Clustering*-Interface aus `ml/code` und dokumentieren Sie den Source-Code an relevanten Stellen. Der Source-Code des Algorithmus muss sowohl im Übungsblatt abgedruckt sein als auch in der elektronischen Abgabe als lauffähiger Source-Code vorliegen. (15 P)
2. Die Firma WIDM ist sich nicht sicher, ob Sie zwei oder drei Roboter auf dem Mars hat und weiSS auch nicht genau wo diese sind. Von Zeit zu Zeit können aber Funksignale empfangen werden die zur Ortung verwendet werden können (vgl. Daten aus ml/code). Verwenden Sie den k-Means Algorithmus mit $k = 2$ und mit $k = 3$ um eine vermutete Region der zwei bzw. drei Roboter zu berechnen. Verwenden Sie die Centroide $c_1 = (10/10)$, $c_2 = (50/50)$ und $c_3 = (80/50)$. Gehen Sie davon aus, dass es keine initiale Clusterzuordnung gibt sondern beginnen Sie einfach mit k-Means um Punktweise eine Clusterzuordnung zu berechnen. Zeichnen Sie sowohl die Referenzdaten mit Clusterzugehörigkeit als auch die ermittelten Centroiden in ein Koordinatensystem ein. Erklären Sie das Ergebnis anhand dieser Zeichnung. (10 P)
3. Verwenden Sie die initialen Centroide $c_1 = (10/10)$, $c_2 = (20/20)$ und $c_3 = (30/30)$ mit $k = 3$. Das Ergebnis scheint keine gute Schätzung für die Position der Roboter zu sein. Erklären Sie, was das Problem ist und wie man dieses Problem umgehen kann. (5 P)
4. Ordnen Sie die Reihenfolge der Referenzdaten so an, dass mit den initialen Centroiden $c_1 = (40/50)$ und $c_2 = (100/50)$ und $k = 2$ nach einem Durchlauf ein Centroid im Bereich $x > 100$ gefunden wird. (5 P)

3 Fahrradkurier

Stellen Sie sich vor, Sie sind der Chef eines Fahrradkurierunternehmens mit sechs Mitarbeitern. Sie besitzen die Aufzeichnung der Kurierdienst-Aufträge des letzten Jahres mit Absender- und Empfängeradresse. Nun sollen Sie ihre Mitarbeiter für das kommende Jahr in feste Gebiete einteilen. Dazu wandeln Sie zunächst die Adressdaten in Koordinaten um. Diese Koordinaten sind in `ml/code` als CSV (Comma Seperated Value) Datei hinterlegt. Jede Koordinate setzt sich aus den Werten der Spalte zwei und drei zusammen, wobei die erste Spalte eine laufende Nummer ist und ignoriert werden kann.

1. Implementieren Sie das *Clustering durch Varianzminimierung* um ein geeignetes Gebiet für jeden Fahrer zu finden. Verwenden Sie auch hierfür das *Clustering*-Interface aus `ml/code`. GroSSteile dieser Aufgabe können Sie aus Ihrer bisherigen Implementierung (vgl. Aufgabe 2) ableiten. Visualisieren Sie die Centroide und die Klassenzugehörigkeit in einem Koordinatensystem. Beschreiben Sie Auffälligkeiten in den Daten/im Clustering. (5 P)

2. Berechnen Sie die Kosten TD (nicht TD^2) des Clusterings. (5 P)

3. Sie konnten mit den Mitarbeitern vereinbaren, dass ihr derzeitiges Gehalt G anhand der Kosten des Clusterings TD bestimmt wird. Sie vereinbaren $G = \frac{TD}{12}$ (Gummimünzen) als fixes einheitliches Grundgehalt für jeden Mitarbeiter (einmalig, basierend auf den Daten des letzten Jahres, nicht basierend auf den noch folgenden Daten). Darüber hinaus bekommt jeder Mitarbeiter noch eine Vergütung für die zurückgelegte Strecke in Höhe der halben Kosten seines Clusters. Als Chef möchten Sie nun ihr Unternehmen optimieren und feststellen, ob Sie lieber weitere Mitarbeiter einstellen oder ggf. sogar welche entlassen sollten. Als Kostenfunktion verwenden Sie dazu $K(k) = G \cdot k + \frac{TD(k)}{2}$, wobei k die Anzahl der Mitarbeiter ist. $G \cdot k$ entspricht dem Grundgehalt der k Mitarbeiter wobei $\frac{TD(k)}{2}$ die gesamte Prämie für die Mitarbeiter repräsentiert. Implementieren Sie das Finden eines optimalen k zur Minimierung der Kostenfunktion $K(k)$ mit $k \in [2, 20]$. (5 P)

4. Berechnen Sie den Silhouetten Koeffizienten für die Clusterings mit den unterschiedlichen k. Gibt es einen Zusammenhang zwischen den Silhouetten Koeffizienten und den Ergebnissen der Kostenfunktion aus der letzten Frage? (5 P)