

*Prof. Dr. Vladimir Shikhman*  
*Professur für Wirtschaftsmathematik*  
*Technische Universität Chemnitz*

*Übungsleiter: David Müller*  
*david.mueller@mathematik.tu-chemnitz.de*

**Mathematische Grundlagen von Big Data Analytics (SS 2018)**  
**Übung 9: Clustering I**

1) Gegeben seien die folgenden Punkte:

$$X = \{(1, 1), (2, 1), (4, 3), (5, 4)\}$$

- a) Führen Sie den k-Means- Algorithmus durch. Wählen Sie hierfür als Dissimilarity Maß die euklidische Distanz und  $k=2$ . Zur Initialisierung der Clusterzentren nehmen Sie die Punkte mit der größten Distanz zueinander.
- b) Zusätzlich zu den Daten aus a) steht Ihnen ein neuer Trainingspunkt (0,6) zur Verfügung. Berechnen Sie die zwei Cluster neu.
- c) Ist das Ergebnis zufriedenstellend?

2) Anstelle der euklidischen Distanz, verwenden Sie für das Trainings Set aus Aufgabe 1b) jetzt die sogenannte "Manhattan-Distanz" (auch "Cityblock- oder Mannheimer-Distanz"):

$$d(x, y) = \sum_{i=1}^p |x_i - y_i|$$

Überlegen Sie sich eine geeignete Wahl für die Zentren und gehen dann analog zu Aufgabe 1 vor. Vergleichen Sie das Resultat des Algorithmus mit dem aus Aufgabe 1b).

3) Zeigen Sie, dass für eine spaltenstochastische Matrix  $P$  gilt:

$$\text{Eigenwerte } |\lambda_i| \leq 1 \quad \forall i.$$