Identifizierung repräsentativer Risikoszenarien mittels Data Analytics

Zusammenfassung der Bachelorarbeit an der Universität Ulm

Laurin Schwartze

Diese Arbeit beschäftigt sich mit einer Methode, mit der Risikoszenarien mittels eines Clusterings klassifiziert werden können. Die für das Verfahren notwendigen Algorithmen werden in Python umgesetzt. Getestet wird das Vorgehen anhand von 1.000 Pfaden einer eigens erzeugten Simulation eines Aktienkurses. Diese Arbeit verwendet Methoden aus der Wahrscheinlichkeitstheorie, der Optimierung, der Statistik und der Informatik. Die Ergebnisse sind für die Aktuarwissenschaften von Bedeutung, da eine große Menge an Risikoszenarien auf eine Teilmenge von repräsentativen Vertretern beschränkt werden kann. Diese Vertreter ermöglichen in natürlicher Weise eine Einteilung der Risikoszenarien in verschiedene Kategorien. Die daraus gewonnenen Informationen erleichtern die Einschätzung realer Situationen bezüglich ihres Risikos und bilden dadurch eine gute Grundlage für das weitere Vorgehen.

Erzeugung des Datensatzes

Im Laufe der Arbeit wird immer wieder auf denselben Datensatz des Grundbeispiels zurückgegriffen. Die benötigten Python Programme werden in Serie geschaltet. Damit wird sichergestellt, dass die Ergebnisse des vorherigen Programmes unverändert weiter verwendet werden. Zunächst werden daher 100.000 Pfade eines Aktienkurses über eine Laufzeit von fünf Jahren simuliert. Das geschieht mit einer geometrischen Brown'schen Bewegung $S_t = S_0 \cdot exp\left(\mu - \frac{\sigma^2}{2}\right)t + \sigma W_t$. Die Zeitintervalle werden äquidistant mit $\Delta t = 1$ gewählt. Als Startwert wird $S_0 = 100$, als Volatilität $\sigma = 0{,}2$ und als Drift $\mu = 0{,}04$ festgelegt. Anschließend werden die Pfade nach dem Wert der Aktie zum Zeitpunkt $t = 5$ sortiert. Für das Verfahren werden nur noch jene 1.000 Pfade mit den niedrigsten Werten für S_5 weiterverwendet. Durch diese Einschränkung auf das unterste Perzentil werden also die Szenarien mit dem höchsten Risiko betrachtet. Man bemerke, dass der Datensatz

aufgrund der geometrischen Brown'schen Bewegung log-normalverteilt ist. Die Pfade können nun als zeitdiskrete Funktion oder als sechs-dimensionale Vektoren aufgefasst werden. Die erste Interpretation wird gelegentlich verwendet, um spezifische Risikoszenarien darzustellen. Das Verfahren selber verwendet die Interpretation als Vektoren. Zunächst werden die Pfade aber zu Renditen transformiert. Dabei ist die Rendite im t –ten Jahr definiert als $R_t = \frac{S_t}{S_{t-1}} - 1$. Hier ist zu bemerken, dass die Renditevektoren nur noch fünf-dimensional sind und dass diese Transformation aufgrund der Definition $S_0 = 100$ invertierbar ist. Es geht also keine Information verloren. Das Verfahren wird anschließend auf die Vektoren der Rendite angewandt.

Wahl der Repräsentanten

Die grundlegende Idee des Verfahrens ist es, den Datensatz geeignet zu clustern. Dadurch wird für jede Gruppe ein Vertreter definiert, der als Repräsentant dieses Clusters verwendet werden kann. Im Grundbeispiel wird das k-means Clusteringverfahren verwendet. Vorteil dieses Verfahrens ist, dass der Vertreter jedes Clusters ein Vektor aus dem ursprünglichen Datensatz, der sogenannte Medoid, sein muss. Der Medoid eines Datensatzes ist definiert als jener Datenwert, der die durchschnittliche Distanz zu den anderen Werten dieses Datensatzes minimiert. Das bedeutet, der Medoid ist eine Verallgemeinerung des Medians auf höher dimensionale euklidische Räume. Die Distanz ist hierbei von der gewählten Metrik abhängig. In dieser Arbeit wird durchgehend die euklidische Metrik verwendet. Für die Anzahl der Cluster wird $k = 10$ verwendet. Das Clustering selber kann mit verschiedenen Algorithmen erzeugt werden. Hier wird der Algorithmus „Partitioning Around Medoids" (PAM) verwendet. Dieser läuft in zwei Phasen ab. Zuerst werden $k = 10$ Medoide gewählt, sodass die Abstände zu Vektoren, die noch keinem anderen Cluster zugeordnet sind, minimiert werden. In der zweiten Phase wird überprüft, ob sich das Clustering durch Vertauschungen zwischen den einzelnen Gruppen verbessern lässt. Wird der Tausch letztendlich vorgenommen, so kann der Medoid dadurch verschoben werden. Die zweite Phase wird wiederholt, bis das Clustering optimal ist oder bis die programmseitig vorgegebene maximale Anzahl an Iterationen erreicht ist.
Die Qualität eines Clusterings wird hierbei an der summierten quadratischen Distanz aller Paare von Vektoren gemessen. Anhand dieses Wertes wird begründet, dass das Clustering gute Gruppen gefunden hat. Die Vektoren innerhalb eines Clusters weisen nun also

eine geringe euklidische Distanz zueinander und damit einen ähnlichen Verlauf auf.

Weitere Analysen

Zuletzt wird das Verfahren allgemein analysiert, mit anderen Verfahren verglichen und ergänzende Analysen werden vorgeschlagen. Als erstes wird der Medoid jedes Clusters mit dem Mittelwert verglichen. Der Mittelwert liefert dabei ein besseres Zentralmaß, das heißt, die durchschnittliche euklidische Distanz der Vektoren im Cluster zum Mittelwert ist geringer als zum Medoiden. Dafür ist der Medoid ein Wert aus dem ursprünglichen Datensatz und damit für das Grundbeispiel besser geeignet. In der Tat wird für das Grundbeispiel berechnet, dass die Distanz zwischen Medoid und Mittelwert im Cluster immer sehr gering ist. Das bedeutet, der Medoid ist nur marginal schlechter als der Mittelwert, bringt aber als Vertreter des Datensatzes mehr Information. Auf eine ähnliche Weise wird das Verfahren mit k-medoids mit dem k-means Clustering verglichen. Diese Art von Clustering funktioniert sehr ähnlich wie k-medoids, aber hier wird um das arithmetische Mittel der Vektoren geclustert. Dadurch könnten ganz andere Clusterings entstehen und da die Wahl des Vertreters nicht so restriktiv ist, kann wieder ein besseres Clustering erreicht werden. K-means wird in dieser Arbeit auf zwei verschiedene Weisen umgesetzt, die sich in der Wahl der initialen Zentroide unterscheiden. Diese werden einmal zufällig bestimmt und bei der zweiten Version werden als Zentroide die vom PAM Algorithmus gefundenen Medoide verwendet. Wie zu erwarten, liefern beide Alternativen einen besseren Wert für die summierte quadratische Distanz als k-Medoids und damit werden bessere Clusterings gefunden. Wie beim Vergleich zum Mittelwert fällt auch hier auf, dass die Verbesserung durch die weniger restriktive Wahl der Vertreter nur sehr gering ist. Das bedeutet, das vorgestellte Verfahren liefert Repräsentanten, die von den Alternativen kaum übertroffen werden. Außerdem stellt sich die Frage, ob sich die gefundenen Vektoren stark voneinander unterscheiden. Auch hier stellt sich im Laufe der Arbeit heraus, dass die Ergebnisse der einzelnen Verfahren kaum Unterschiede aufweisen. Deshalb ist die Verwendung der Methode mit k-medoids Clustering anstelle der anderen Varianten gerechtfertigt. Die bisherigen Ergebnisse sind immer nur für das Grundbeispiel betrachtet worden und enthalten darum einen randomisierten Anteil aus der geometrischen Brown’schen Bewegung. Daher stellt sich die Frage, ob eine erneute Generation eines Datensatzes ähnliche Ergebnisse liefern würde. Es soll daher überprüft werden, ob die Medoide eines neuen Datensatzes mit den bisher gefundenen verglichen werden können.

Dafür wird ein Matching zwischen den Mengen der Repräsentanten der beiden Generationen erstellt. Grundlage des Matchings ist bei der ersten Analyse die euklidische Distanz zweier Medoide. Die zweite Analyse hingegen matcht Medoide aufgrund ihres Verlaufes als Aktienkurs. Diese Ergebnisse sind interessanter, denn durch den Verlauf können die Risikoszenarien klassifiziert werden. In der Tat liefert das Verfahren angewandt auf verschiedene Generationen der geometrischen Brown'schen Bewegung Vertreter, die die verschiedenen Datensätze in ähnliche Kategorien aufteilen. Das ist ein zentrales Resultat dieser Arbeit. Daraus kann geschlossen werden, dass die Medoide als Vertreter des gesamten Datensatzes eine Charakterisierung der typischen Risikoszenarien erlauben. Diese ist unabhängig von der spezifischen Ausprägung der Simulation und liefert stabile Resultate.

Fazit

Für andere Anwendungen der Datenreduktion müsste das Verfahren allerdings weiter erprobt werden. Die Aktienkurse stellen eine Situation dar, bei der große Datensätze die Norm sind. In genau solchen Anwendungen kann von einer reduzierten Grundmenge profitiert werden. Im Grundbeispiel sind die Ergebnisse sehr aussagekräftig. Hier ist aber anzumerken, dass die Definition von $k = 10$ nie mathematisch begründet worden ist. Für weitere Verbesserungen des Verfahrens könnte hier angesetzt werden. Außerdem werden an verschiedenen Stellen dieser Arbeit der Einfachheit halber suboptimale Verfahren verwendet, die durch exakte Alternativen ersetzt werden könnten. Aber wie das Grundbeispiel zeigt, findet das in dieser Arbeit beschriebene Verfahren bereits sehr gute Vertreter, die als typische Risikoszenarien herangezogen werden können.