Hausarbeit zur Vorlesung
Einführung in Bayessische Netze für Geowissenschaftler

Ein wesentliches Element in der seismischen Gefährdungsanalyse sind Bodenbewegungsmodelle, welche die zu erwartende Bodenbewegung abhängig von den Eigenschaften des Erdbebens, des Untergrundes und des Weges der seismischen Wellen beschreiben. Diese Modelle sind meist von Unsicherheiten geprägt. Werden diese Unsicherheiten nicht berücksichtigt, wird die seismische Gefährdung häufig unterschätzt. Probabilistische Modelle bieten den Vorteil, vorhandene Unsicherheiten in der Schätzung abbilden zu können. Das stochastische Modell von Boore [Boore, 2003] erfasst seismischen Eigenschaften und verbundene Unsicherheiten sehr gut. Es hat jedoch keine einfache analytische Form und erfordert hohe Rechenzeiten. In der probabilistischen seismischen Gefährdungsanalyse werden daher häufig vereinfachte Modelle (z.B. Regressionsmodelle) verwendet.

Der vorliegende Datensatz soll verwendet werden, um ein effizientes probabilistisches Bodenbewegungsmodell zu generieren. Der Datensatz wurden synthetisch erzeugt und enthält 10 000 Einträge. Dabei wurden die Variablen Magnitude (M), Spannungsabfall (SD), Entfernung (R), Dämpfung (Q_0 und κ_0) sowie Scherwellengeschwindigkeit (V_S30) unabhängig voneinander aus beschränkten Exponential- oder Gleichverteilungen gesampelt (siehe Tabelle 1). Zu jedem gesampelten "Ereignis" wurde der Bodenbewegungsparameter (PGA) mit dem stochastischen Modell von Boore bestimmt, welches die beschriebenen Einflussvariablen verwendet, um die resultierende Bodenbewegung zu schätzen.

Table 1: Der zu betrachtende Datensatz enthält die folgenden Variablen und wurde aus den entsprechenden Verteilungen gesampelt.

X_i	Beschreibung	Verteilung$_{[\text{range}]}$
	Einflussvariablen	
M	Momenten-Magnitude des Erdbebens	$\mathcal{U}_{[5,7.5]}$
R	Entfernung der Quelle zum Standort	$\text{Exp}_{[1\,\text{km},\,200\,\text{km}]}$
SD	Spannungsabfall während des Bebens (stress drop)	$\text{Exp}_{[0\,\text{bar},\,500\,\text{bar}]}$
Q_0	Dämpfung der Amplituden der seismischen Wellen in tiefen Schichten	$\text{Exp}_{[0\,\text{s}^{-1},\,5000\,\text{s}^{-1}]}$
κ_0	Oberflächennahe Dämpfung der Amplituden der seismischen Wellen	$\text{Exp}_{[0\,\text{s},\,0.1\,\text{s}]}$
V_S30	Durchschnittliche Scherwellengeschwindigkeit in den oberen 30 m	$\mathcal{U}_{[600\,\text{m}\,\text{s}^{-1},\,2800\,\text{m}\,\text{s}^{-1}]}$
	Bodenbewegungsparameter	
$\log PGA$	Logarithmus der größten horizontalen Bodenbeschleunigung	synthetisch erzeugt nach dem stochastischen Modell von Boore [Boore, 2003]

Beachten Sie, dass der Datensatz die Werte $\log PGA$, statt PGA enthält. Meist wird angenommen, dass $\log PGA$ einer Normalverteilung folgt. Ein kontinuierlichen Bayesschen Netzes, welches aus dem Datensatz mit den vorliegenden Wahrscheinlichkeitsverteilungen konstruiert wird, würde sehr lange Rechenzeiten für Inferenzen benötigen, da verschiedene, nicht kompatible Verteilungsfamilien miteinander kombiniert werden müssen. Statt dessen sollen diskrete Bayessche Netze gelernt werden. Diskritisieren Sie die Daten dazu in folgende Intervalle:

SD:	0,	0.8792,	5.438,	14.92,	58,	500			
Q_0:	0,	330,	5000						
κ_0:	0,	0.01053,	0.0345,	0.1					
V_S30:	600,	1704.5,	2800						
M:	5,	6.271,	7.5						
R:	1,	4.38,	15.4885,	55.84,	200				
$\log PGA$:	-Inf,	-5.135,	-3.722,	-2.627,	-1.20742,	0.145,	1.657,	3.175,	Inf

Die Intervalle wurden nach Kriterien gewählt, bei denen der Informationsverlust möglichst gering bleibt.

Teilen Sie den Datensatz in einen Lerndatensatz (bestehend aus 9000 Beobachtungen) und einen Testdatensatz (bestehend aus 1000 Beobachtungen).

Konstruieren Sie ein kausales Netz (causal network), welches die kausalen Zusammenhänge der Variablen im vorliegenden Datensatz darstellt. Bestimmen Sie die zugehörigen Parameter aus dem Lerndatensatz.

Konstruieren Sie ein "Naive Bayes"-Netz und bestimmen Sie die zugehörigen Parameter aus dem Lerndatensatz. Vergleichen Sie die Komplexität beider Netze und schätzen Sie die Fähigkeiten beider Netze ein, die zu erwartende Bodenbewegung (PGA) bei gegebenen Einflussvariablen vorherzusagen.

Lernen Sie Struktur und Parameter eines Bayesschen Netzes aus dem Lerndatensatz. Vergleichen Sie die Struktur des Netzes mit der Struktur des kausalen Netzes und erläutern Sie die Ursachen für eventuelle Unterschiede.

Vergleichen Sie die Vorhersagegenauigkeit aller 3 Netze! Nutzen Sie dazu den Testdatensatz bestehend aus 1000 Beobachtungen und prüfen Sie wie genau die Modelle die Bodenbewegung (PGA) aus den gegebenen Einslussvariablen vorhersagen können. Überlegen Sie dazu, wie Sie aus der berechneten diskreten bedingten Wahrscheinlichkeitsverteilung einen Punktschätzer ableiten können. Begründen Sie ihre Wahl. Bestimmen Sie den mittleren quadratischen Fehler des Punktschätzes zu den PGA-Werte aus dem Testdatensatz. Vergleichen und erläutern Sie die Ergebnisse.

Eine andere Möglichkeit die Güte der Vorhersage zu beurteilen, besteht darin die Wahrscheinlichkeiten zu berechnen, die die Modelle den PGA-Werten im Testdatensatz zuordenen. Wir bezeichnen mit $I[PGA^{(k)}]$ das Interval, in dem der PGA-Wert der k-ten Beobachtung des Testdatensatzes liegt. Liegt der PGA-Wert der ersten Beobachtung im Testdatensatz bspw. im dritten Interval ($I[PGA^{(1)}] = 3$), wird ein 'gutes' Model dem dritten Interval, für die gegebenen Werte der Einflussvariablen, vermutlich eine hohe Wahrscheinlichkeit zuweisen. Entsprechend sollte ein 'gutes' Model allen PGA-Intervallen des Testdatensatzes eine relativ hohe bedingte Wahrscheinlichkeit, gegeben den beobachten Einflussvariablen, zuordnen. Die Gesamtwahrscheinlichkeit die das Modell den beobachten Intervallen zuweist ergibt sich aus dem Produkt der Einzelwahrscheinlichkeiten

$$\mathrm{P}((I[PGA^{(1)}], I[PGA^{(2)}], \ldots, I[PGA^{(1000)}]) \mid \mathbf{d}) =$$

$$\prod_{k=1}^{1000} \mathrm{P}(I[PGA^{(k)}] \mid M^{(k)}, SD^{(k)}, R^{(k)}, V_S30^{(k)}, Q_0^{(k)}, \kappa_0^{(k)})$$

Da diese Berechnung zu einem Produkt sehr kleiner Zahlen führt, rechnet man häufig statt dessen mit dem Logarithmus der Wahrscheinlichkeiten und erhält dadurch eine Summenberechnung:

$$\log \prod_{k=1}^{1000} \mathrm{P}(I[PGA^{(k)}] \mid M^{(k)}, SD^{(k)}, R^{(k)}, V_S30^{(k)}, Q_0^{(k)}, \kappa_0^{(k)}) =$$

$$\sum_{k=1}^{1000} \log \mathrm{P}(I[PGA^{(k)}] \mid M^{(k)}, SD^{(k)}, R^{(k)}, V_S30^{(k)}, Q_0^{(k)}, \kappa_0^{(k)})$$

Berechnen Sie zu den 3 Bayesschen Netzen die bedingten Wahrscheinlichkeiten der beobachteten PGA-Intervalle des Testdatensatzes. Vergleichen Sie diese Wahrscheinlichkeiten untereinander und vergleichen Sie ihr Ergebnis mit den Ergebnissen des mittleren quadratischen Fehlers des Punktschätzers. Welches Modell würden Sie für eine probabilistische Gefährdungsanalyse wählen?

Dokumentieren Sie zu allen Fragestellungen, welche Methoden/Verfahren Sie verwendet haben und warum.

References

[Boore, 2003] Boore, D. (2003). Simulation of ground motion using the stochastic method. *Pure and Applied Geophysics*, 160(3-4):635–676.