Erläuterung zur Hausarbeit

Der Datensatz soll in 2 Teile geteilt werden, dem Lerndatensatz (9000 Beobachtungen) und dem Testdatensatz (1000 Beobachtungen). Der diskretisierte Lerndatensatz wird benutzt um die Parameter im kausalen Netz und beim Naive Bayes zu bestimmen. Zudem wird der Lerndatensatz genutzt um die Struktur und die Parameter eines 3. Netzes zu lernen. Nun hat man 3 verschiedene Modelle (kausales Netz, Naives Bayes Netz, gelerntes Netz), deren Vorhersagegenauigkeit geprüft werden soll, d.h. es soll geprüft werden, wie gut die Modelle den PGA hervorsagen können, wenn Magnitude, Distanz, Stress drop, ... gegeben sind. Dazu verwendet man den Testdatensatz.
Wir betrachten 2 Varianten die Vorhersagegenauigkeit zu testen.

1. Nutze den mittleren quadratischen Fehler eines Punktschätzers

Zu jeder Beobachtung im Testdatensatz wird die bedingte Wahrscheinlichkeitsverteilung von PGA gegeben den anderen Variablen $P(PGA|SD, Q_0, \kappa_0, Vs_{30}, M, R)$ bestimmt. Aus der Verteilung wird ein Punktschätzer für PGA bestimmt und anschließend der quadratische Abstand zum wirklichen PGA Wert berechnet.

Beispiel:
Angenommen wir untersuchen zuerst das Naive Bayessche Netz.

Die ersten Beobachtungen im Testdatensatz sind

SD	Q_0	κ_0	Vs_{30}	MAG	DIST	PGA
0.640	546.9	0.07815	1110	6.60	5.60	-2.85
36.856	4315.7	0.00190	1205	6.11	5.97	2.52
0.549	2326.1	0.00277	2798	7.49	107.25	-3.
76.075	255.1	0.01609	869	6.95	1.42	3.94
53.437	684.5	0.00858	2233	7.02	4.04	2.50
24.123	54.4	0.00837	1130	5.37	125.48	-6.38
29.956	1746.9	0.02081	1671	5.26	24.60	-1.64
343.352	1503.3	0.00835	723	5.86	16.45	2.41
22.192	688.5	0.01216	1246	6.75	1.32	2.88

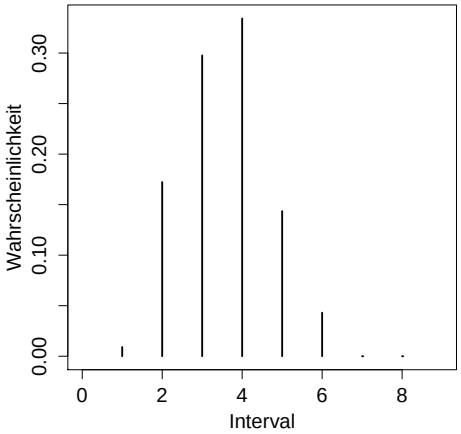
Bzw. wenn wir die diskreten Intervalle benutzen:

SD	Q_0	κ_0	Vs_{30}	MAG	DIST	PGA
1	2	3	1	2	2	3
4	2	1	1	1	2	7
1	2	1	2	2	4	3
5	1	2	1	2	1	8
4	2	1	2	2	1	7
4	1	1	1	1	4	1
4	2	2	1	1	3	4
5	2	1	1	1	3	7
4	2	2	1	2	1	7

Wir nehmen nun die erste Beobachtung und berechnen
$P(PGA|SD = 1, Q_0 = 2, \kappa_0 = 3, Vs_{30} = 1, M = 2, R = 2)$ mit dem Naive Bayes Modell.

In R:
```
P_PGA = cpdist(fit.naive_bayes,"PGA",
              evidence=(SD==tdat[1,1] & Q0==tdat[1,2] & kappa==tdat[1,3] &
                        Vs30==tdat[1,4] & MAG==tdat[1,5] & DIST==tdat[1,6]))
```

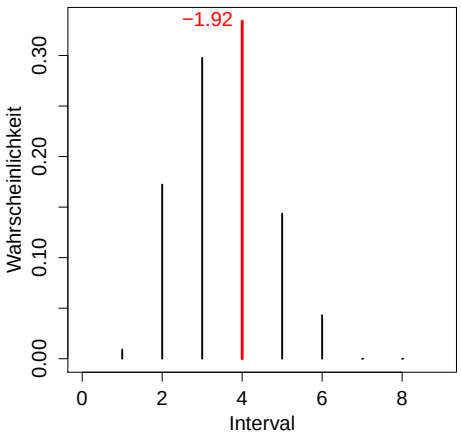
In P_PGA stehen nun die Wahrscheinlichkeiten für die 8 PGA-Intervalle.



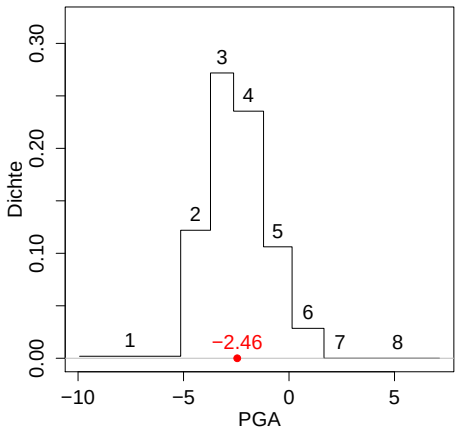
Das bedeutet, die max Bodenbeschleunigung eines Bebens mit den gegeben Eigenschaften (SD im 1. Interval, Q_0 im 2. Intervall, ...) liegt mit einer Wahrscheinlichkeit von ca. 0.33 im 4. Interval (PGA zwischen -2.63 und -1.21); mit einer Wahrscheinlichkeit von ca. 0.3 im 3. Interval (zwischen -3.72 und -2.63); etc.

Aus dieser Verteilung soll nun ein Punktschätzer gewonnen werden. Dies könnte bspw. der Mittelpunkt des wahrscheinlichsten Intervalls sein (siehe Abb.).

Tipp: Das Interval mit der größten Wahrscheinlichkeit erhält man in R mit:
`which(max(P_PGA)==P_PGA)`



Es könnte auch der Mittelwert der Wahrscheinlichkeitsverteilung sein:



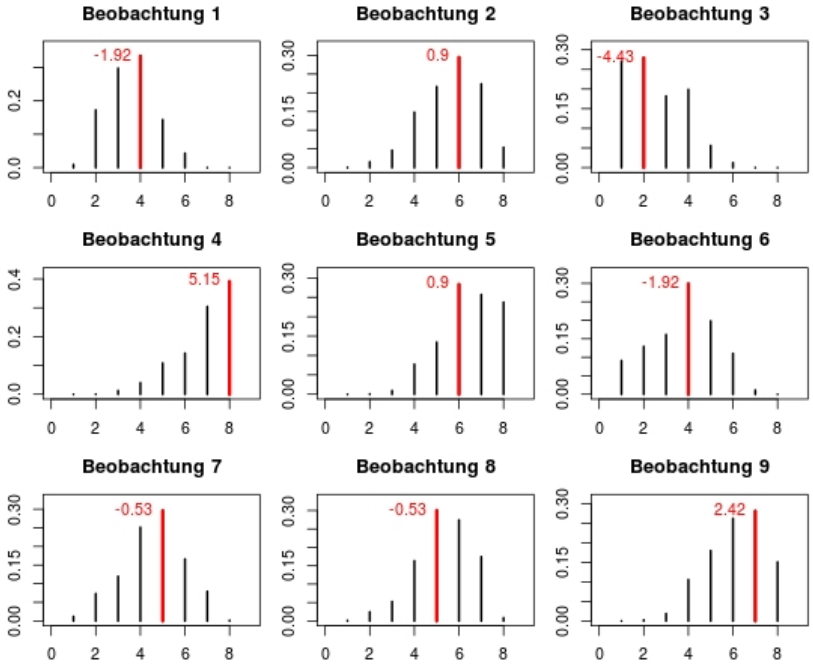
Die Höhe der Dichte entspricht der "Wahrscheinlichkeit des Intervals"/"Breite des Intervals" oder in anderen Worten: Dichte * Breite des Intervals = Wahrscheinlichkeit des Intervals.

Wie ihr den Punktschätzer wählt, ist euch überlassen. Ihr könnt auch einen anderen Wert nehmen. Wählt den Punktschätzer so, dass er euch sinnvoll erscheint und begründet eure Wahl evtl kurz.

Angenommen, ihr wählt den Mittelwert des wahrscheinlichsten Intervals als Punktschätzer, dann berechnet ihr nun den quadratischen Fehler, also den Abstand eures Punktschätzers (-1.92) zum

beobachteten PGA-Wert (PGA Wert der ersten Beobachtung im Testdatensatz: -2.85) zum Quadrat: $(-1.92 + 2.85)^2 = 0.865$

Dies wiederholt ihr für alle Beobachtungen des Testdatensatzes



und berechnet dann den Mittelwert aller quadratischen Fehler:
$\left((-1.92 + 2.85)^2 + (0.9 - 2.52)^2 + (-4.43 + 3.57)^2 + (5.15 - 3.94)^2 + (0.9 - 2.5)^2 + ...\right) /1000$

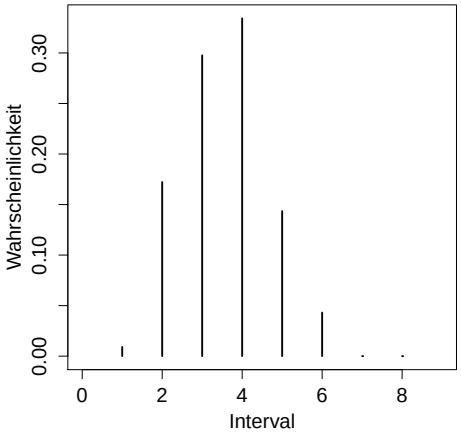
2. Welche Wahrscheinlichkeiten weisen die Modelle den beobachteten PGA-Intervallen zu

Bei der 2. Methode schauen wir uns, welche Wahrscheinlichkeiten die Intervalle mit den beobachteten PGA Werten bekommen. Dazu gehen wie folgt vor:

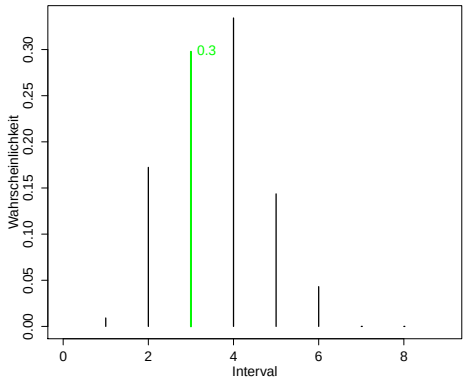
Angenommen wir untersuchen wieder zuerst das Naive Bayessche Netz. Wir gehen wieder jede Beobachtung einzeln durch und berechnen zunächst wieder die bedingten Wahrscheinlichkeitsverteilungen $P(PGA|SD, Q_0, \kappa_0, Vs_{30}, M, R)$ (bzw. kennen wir da ja schon von der 1. Variante). Zur Erinnerung, die erste Beobachtung im Testdatensatz (diskretisiert) ist:

SD	Q_0	κ_0	Vs_{30}	MAG	DIST	PGA
1	2	3	1	2	2	3

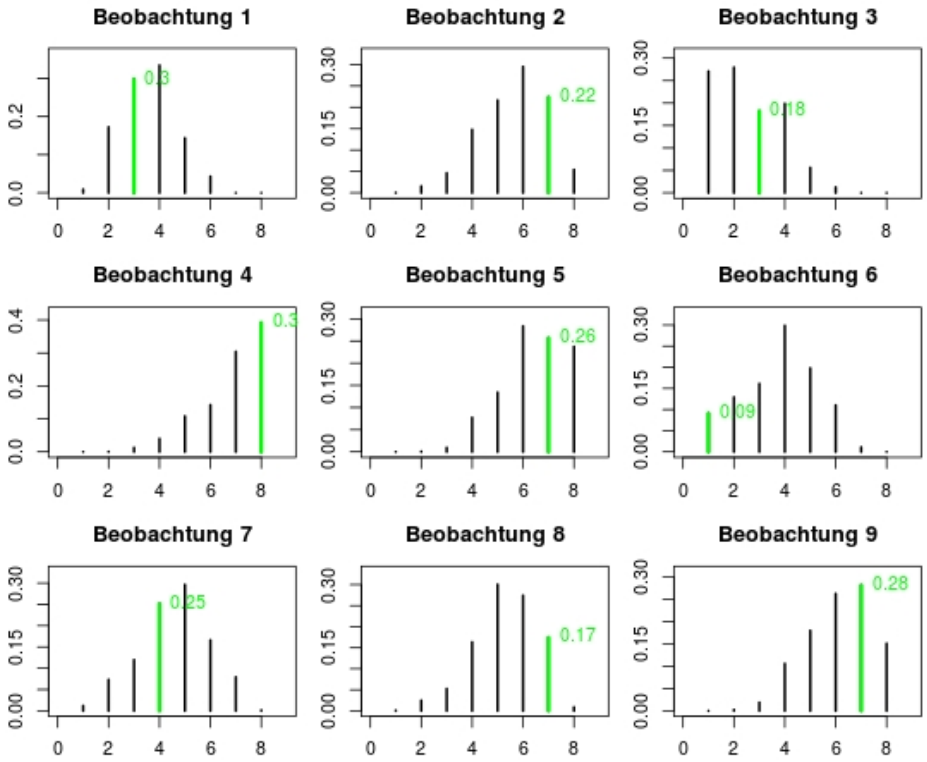
und die zugehörige bedingte Wahrscheinlichkeitsverteilung von PGA ist:



Nun suchen wir das Intervall raus, indem der beobachtete PGA-Wert aus dem Testdatensatz liegt. Das ist das 3. Intervall. Dieses Intervall hat eine Wahrscheinlichkeit von 0.3.



Das wiederholen wir jetzt auch wieder für jede Beobachtung und multiplizieren die Wahrscheinlichkeiten miteinander:



$0.3 \cdot 0.22 \cdot 0.18 \cdot 0.3 \cdot 0.26 \cdot 0.09 \cdot 0.25 \cdot 0.17 \cdot 0.28 \cdot ...$
Da dieses Produkt sehr klein wird können wir alternativ auch die Summer der logarithmischen Werte berechnen: $\log 0.3 + \log 0.22 + \log 0.18 + \log 0.3 + ...$

Erklärung: Wenn ein Modell den richtigen PGA gut vorhersagen kann, wird es dem entsprechendem Intervall eine hohe Wahrscheinlichkeit zuweisen und dies möglichst für alle Beobachtungen aus dem Testdatensatz. Das Modell ist also besser, je größer der berechnete Wert ist.