Project: Hidden Markov models and Value of Information

Faiga Alawad, Martin Outzen Berild, Håkon Gryvill

November 2019

The aim of this exercise is to predict the chance of a snow avalanche on a railroad track. The track is discretized into 50 regions in increasing altitude, where each region is assigned a probability being a high-risk region. The probability of area i being a high-risk region is $p(x_i = 1)$, while the probability of the being a low-risk region is $p(x_i = 0)$.

We also assume the following transition probabilities:

$$
\begin{aligned}
p(x_{i+1} = 0 | x_i = 0) &= 0.95 = r \\
p(x_{i+1} = 1 | x_i = 0) &= 0.05 \\
p(x_{i+1} = 0 | x_i = 1) &= 0 \\
p(x_{i+1} = 1 | x_i = 1) &= 1
\end{aligned}
$$

That is, if a given region is a high-risk region, then all the following regions are high-risk as well. It is impossible to go from a high-risk region to a low-risk region. There is a 95 % chance of a region being low-risk if the previous region also was low-risk.

a)

The probability of state i being high-risk can be calculated as follows: We know that: $p(x_1 = 0) = 0.99 = p_1$. Thus,

$$p(x_2 = 0) = p(x_2 = 0 | x_1 = 0) p(x_1 = 0) = p_1 \cdot r \tag{1}$$

Generally, we have that

$$p(x_i = 0) = p_1 r^{i-1} \tag{2}$$

Hence, $p(x_i = 1) = 1 - p(x_i = 0) = 1 - p_1 r^{i-1}$. The probabilities are visualized in Figure 1. We see that the probability of being a high-risk region increases monotonously. Since state 1 is absorbing (i.e. we cannot leave state 1 once we entered it), this is what we would expect.

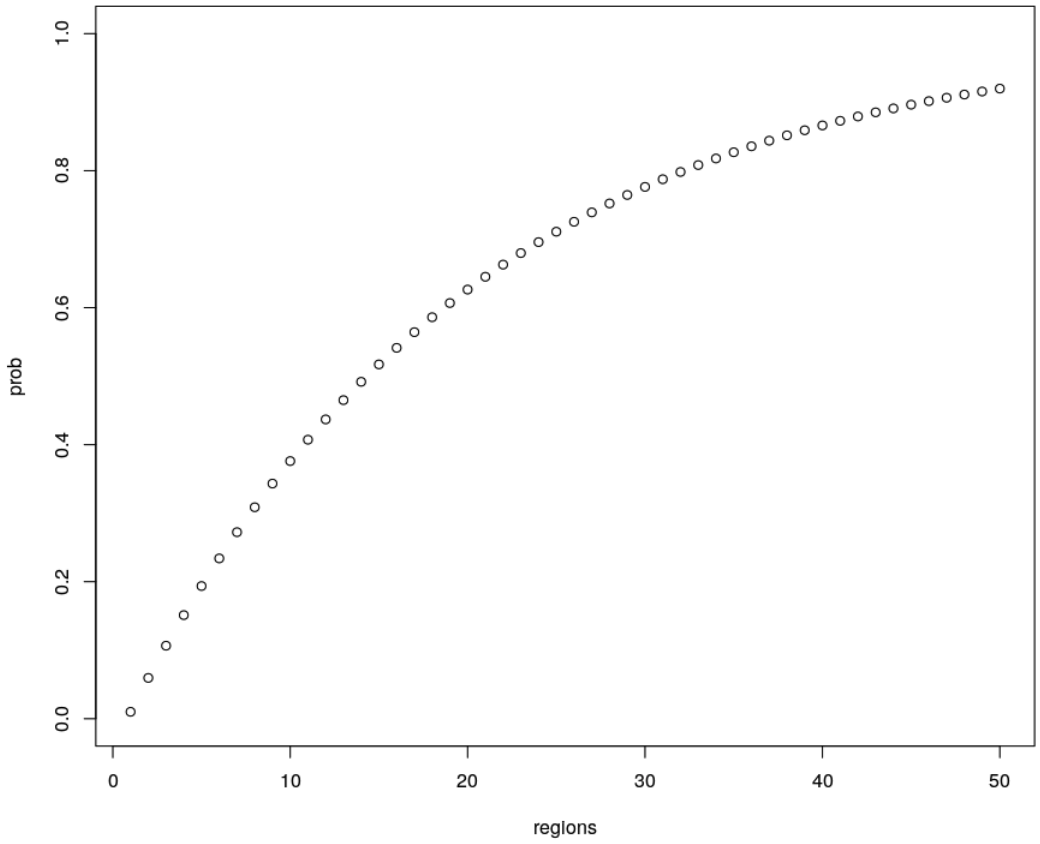

Figure 1: The probability of a region being a high-risk region $p(x_i = 1)$, for each region i.

b)

We are now presented with the following dilemma: clean all tracks for a sum of 100000 kroner, or all clean all high-risk tracks for 5000 kroner each. We want to choose to cheapest outcome, i.e. the alternative with the minimum expected cost. This decision situation is formulated as follows:

$$\text{PV} = \max\{-100000, -5000 \sum_{i=1}^{50} p(x_i = 1)\}. \tag{3}$$

PV denotes the prior value, i.e. the value given no further information about the risk in each region. By summing over the the probabilities $p(x_i = 1)$ found in a), we find that PV $= -100000$ kroner. That is, we should clean all tracks in advance.

c)

We are now given some additional information about the railroad tracks. At region i we observe y_i, which is assumed to be Gaussian distributed around x_i with a noise term:

$$p(y_i|x_i) = N(x_i, \tau^2), \tag{4}$$

where $\tau = 0.3$. At regions 20 and 30 we observe the values $y_{20} = 0.2$ and $y_{30} = 0.7$, respectively. We now want to compute the posterior probabilities, i.e. the probability of each region being high-risk, given the new data:

$$p(x_i = 1|y_{20} = 0.2, y_{30} = 0.7). \tag{5}$$

The posterior probabilities are visualized in Figure 2. We notice that these two data points change probabilities drastically, when compared to 1. Note that if these values were observed with much higher variance, say $\tau = 30$, we would obtain approximately the same probabilities. This is due to the fact that our observations would be given less weight, as they would be deemed imprecise.

d)

We are now given the opportunity to place a sensor at a region of our choice, at a cost of 10000 kroner. If a sensor is placed at region k, this sensor measures the value y_k, which is distributed according to (4). We want to find the optimal location for a sensor, that is, the location that maximizes the posterior value:

$$\text{PoV}(k) = \int \max\{-100000, -5000 \sum_{i=1}^{50} p(x_i = 1|y_k)\} p(y_k) dy_k. \tag{6}$$

We can then compute the value of information, VOI(k) in each region as follows:

$$\text{VOI}(k) = \text{PoV}(k) - \text{PV}. \tag{7}$$

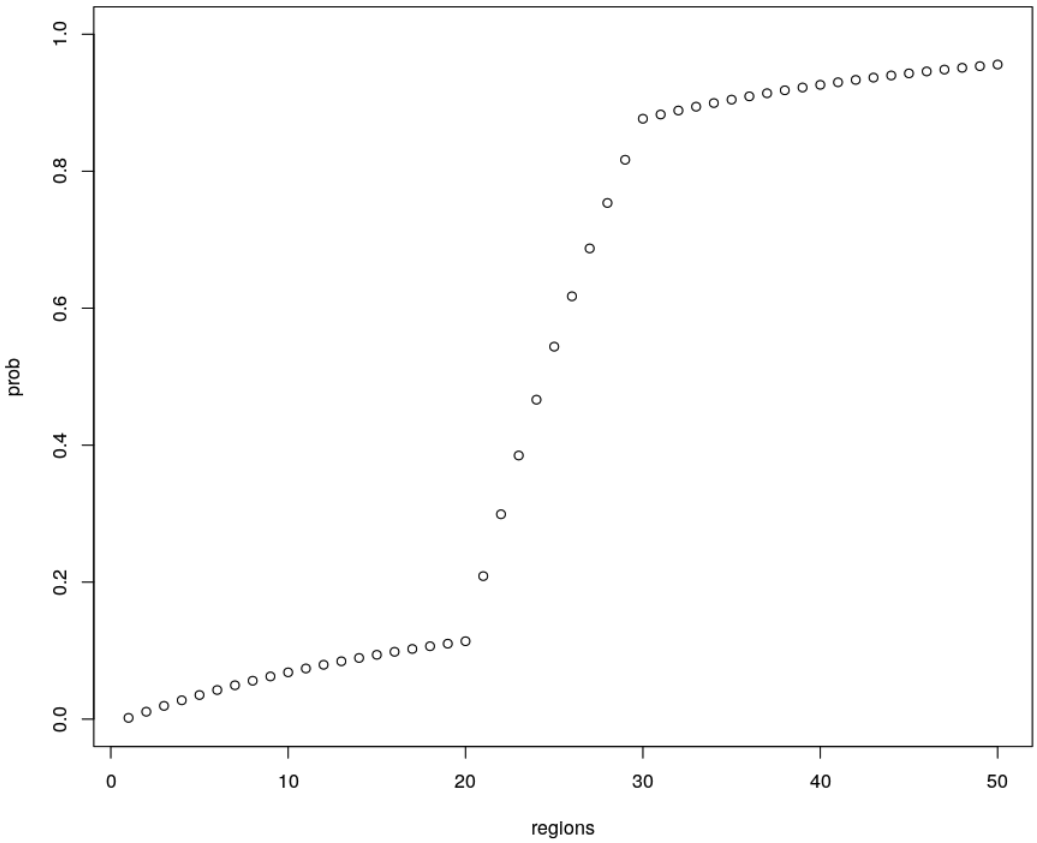

Figure 2: Posterior probabilities $p(x_i = 1|y_{20} = 0.2, y_{30} = 0.7)$.

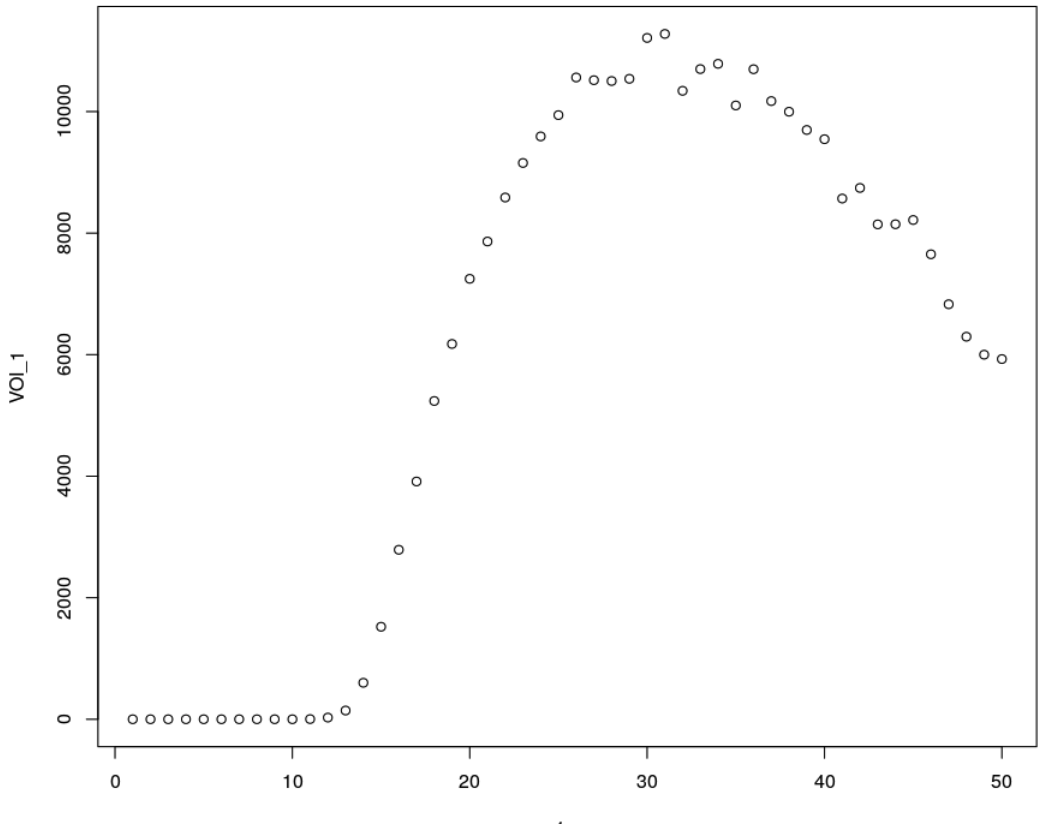

Figure 3: The approximate value of information, VOI, in each region $k = 1, \ldots, 50$.

The integral is approximated with Monte Carlo-sampling with $B = 10000$ samples for each $k = 1, \ldots, 50$. The approximate value of information in each region is presented in Figure 3. From the plot, we find that the optimal location for a sensor is $k = 31$. This sensor yields a value of information of $\text{VOI}(31) = 11278$ kroner. Since the value of information is larger than the cost of the sensor, we would like to install the sensor.

e)

We are now given the opportunity to place two sensor at regions of our choice, for a price of 15000 kroner. We follow the same procedure as in task d), except that (6) has to be replaced by an integral that takes into account that two

sensors will be placed simultaneously:

$$\text{PoV}(k) = \int \max\{-100000, -5000 \sum_{i=1}^{50} p(x_i = 1|\boldsymbol{y_D})\} p(\boldsymbol{y_D}) d\boldsymbol{y_D}, \qquad (8)$$

where D denotes the locations of the two sensors, and y_D represents the observed values in the two locations. The value of information is calculated in the same way as in d). We approximate the integral with Monte Carlo sampling for each combination of locations. Figure 4 illustrates the value of information for each combination. We used $B = 1000$ samples when the two locations where different, i.e. $d_1 \neq d_2$ when $D = [d_1, d_2]$, and $B = 10000$ samples when $d_1 = d_2$.

The optimal locations were $D = [21, 33]$, which gives a value of information of VOI = 14811. Since the value of information is less than the cost of placing the sensors, we would not prefer to place sensors.

f)

Lastly, we are given the opportunity to obtain information about all of the locations. This time, however, the quality of the information is poorer than previous: $\tau = 1$. We approximate (8) with $D = [1, \ldots, 50]$ and $B = 50000$ Monte Carlo samples. The approximated value of information is VOI = 12830. If the cost of this information is smaller than this value, the information would be worth the price.

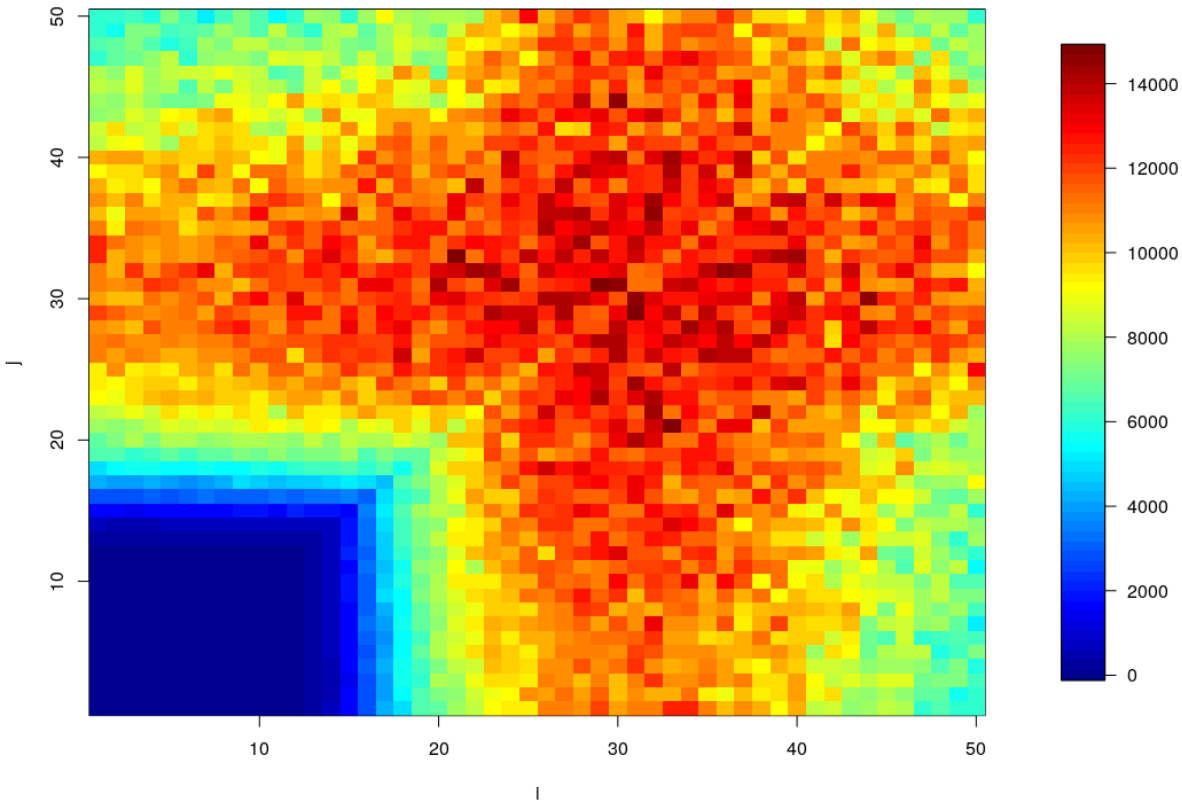

Figure 4: The approximate value of information, VOI, when placing two sensors. For example, we see that placing two sensors in the area [0, 18] is worthless.