Project : Hidden Markov models and Value of Information

Parts of a railroad track are at risk of snow avalanche. The avalanche risks are discretized in low (0) or high (1) classes. Low risk is more likely for low altitude, while high risk is more likely for higher altitude. The railroad track is split in 50 different sections, of increasing altitude, and the risk class $x_i \in \{0, 1\}$, $i = 1, \dots, 50$ is modeled by a Markov chain as follows: Initial probability of low risk is $p(x_1 = 0) = 0.99$ and the transition probabilities are defined by $p(x_{i+1} = 0|x_i = 0) = 0.95$, $p(x_{i+1} = 1|x_i = 1) = 1$.

a)

Compute and plot the marginal probabilities $p(x_i = 1)$ for $i = 1, \dots, 50$.

b)

A particularly warm day in springtime snow avalanches will surely occur on the railroad line. One is worried about the cost associated with these avalanches. Decision alternative $a = 0$ is to clean the tracks at all locations ahead of the daily operation, at the cost of $v(\boldsymbol{x}, a = 0) = -100000$ kroner, no matter the outcome of $\boldsymbol{x} = (x_1, \dots, x_{50})$. Decision alternative $a = 1$ means to take the uncertain costs connected with high/low risks: Assume that there are no costs at low risk locations, while each high risk location will cost 5000 kroner to clean and this gives value function $v(\boldsymbol{x}, a = 1) = -5000 \sum_{i=1}^{50} I(x_i = 1)$. The railroad company chooses the alternative with minimum expected cost, and this decision situation gives prior value (PV) equal to

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

What is the optimal decision?

c)

By installing one or more sensors one gets information about high or low risks, and this allows informed decisions about cleaning the tracks up front or not. When data is gathered at location i we get observation y_i. These data are modeled by density function $p(y_i|x_i) = N(x_i, \tau^2)$, and conditionally independent given x_i, $i = 1, \dots, n$. We set $\tau = 0.3$. The data will represent a hidden Markov model (HMM). For a sampling design $D = \{D_1, \dots, D_{|D|}\}$, observations are denoted $\boldsymbol{y}_D = (y_{D_1}, \dots, y_{D_{|D|}})$, where the subscripts indicate sensor location(s) which are a subset of the 50 sections.

Assume data $\boldsymbol{y}_D = (0.2, 0.7)$ are observed for design $D = \{20, 30\}$. Use the forward-backward algorithm for HMMs to compute the posterior probabilities $p(x_i = 1|\boldsymbol{y}_D)$, $i = 1, \dots, 50$. Plot these posterior probabilities.

(The simplest way to implement this on the computer might be to generate data at all sections - and when doing the forward-backward calculations on these simulated data the likelihood noise term is set much higher, say $\tau = 100\tau$, at locations that are not part of the design.)

d)

Before placing sensor(s), a goal is to find the best design using value of information (VOI) analysis. The expected posterior value (PoV) with information, when a single sensor is placed at location $k = 1, \dots, n$ is:

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

The VOI is then

$$\mathrm{VOI}(k) = \mathrm{PoV}(k) - \mathrm{PV}.$$

Use Monte Carlo sampling of data to approximate the PoV(k) and VOI(k) for different k, using, say 1000 or 10000 Monte Carlo samples. For each Monte Carlo sample of data, the forward-backward algorithm is run to calculate the posterior marginal probabilities and the integrand required for the PoV. Plot the VOI as a function of single sensor locations $k = 1, \dots, n$. What is the best location for the single sensor? The total rental and installation cost of one sensor is 10000 kroner. Will sensor information be worth the price? (It is also possible, and more computationally efficient, to approximate the VOI by numerical integration in this case, using a discretized version of $p(y_k)$.)

e)

The VOI of multiple sensor data is

$$\mathrm{VOI}(D) = \mathrm{PoV}(D) - \mathrm{PV},$$

$$\mathrm{PoV}(D) = \int \max\{-100000, -5000 \sum_{i=1}^{50} P(x_i = 1 | \boldsymbol{y}_D)\} p(\boldsymbol{y}_D) d\boldsymbol{y}_D.$$

What is the VOI at the design $D = \{20, 30\}$ considered above?

Find a near-optimal design of size $|D| = 2$? (You could here start by design $D = \{20, 30\}$, and then iteratively suggest random changes to one or both of the elements and accept the change if it has larger VOI. A more nuanced version of this approach is called the exchange algorithm.)

The price of having two sensors is 15000 kroner. Would two-sensor data be worth it?

f)

There is another option for information gathering from processing satellite data, at a cost. This data will be available in all 50 segments, but they are of relatively poor quality. The model for the data is $p(y_i|x_i) = N(x_i, 1^2)$, with conditional independence given x_i, $i = 1, \dots, n$.

Approximate the VOI of this data using Monte Carlo sampling.