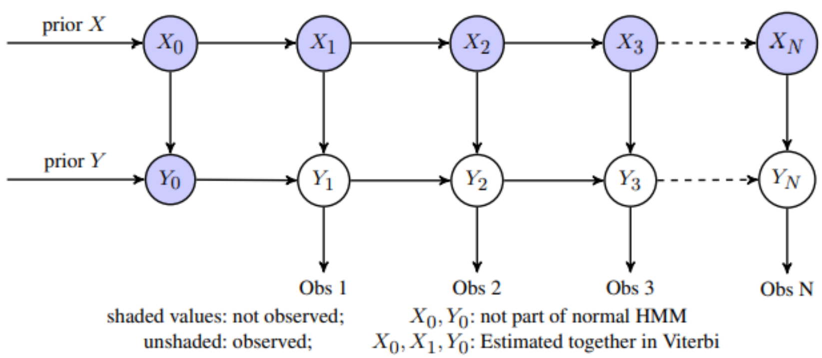
Expectation Maximization Algorithm of the Markov Observation Model for Predicting Healthcare Insurance Fraud

In order to create a better prediction tool for healthcare insurance fraud, we introduce a novel statistical approach, called Markov Observation Model (MOM), developed by Professor Kouritzin at the University of Alberta, to model the simulated health insurance claims in a sequential set up using counting to capture fraudulent claims whenever they appear. It is a Hidden Markov Model which is expanded to allow for Markov chain observations that would estimate the most likely sequence of hidden states given the sequence of observations with high accuracy. In MOM, the hidden state is a homogeneous Markov chain (i.e constant transition probability) , and the observations are also assumed to be a Markov chain whose one-step transition probabilities depend on the hidden Markov chain. This model allows us to estimate transition probabilities for both the hidden state and the observations, as well as the initial joint hidden-state-observation distribution. MOM aims to narrow the gap between the limited Hidden Markov Model (HMM) and more complex models.



The figure is a good overview of the model. The next section will deal with a mathematical derivation of the Expectation Maximization (EM) Algorithm for the Markov Observation Model, which is extremely important in estimating the transition probabilities of the hidden Markov Chains and to establish a relationship between the hidden layer and the observations.

Expectation Maximization (EM) Algorithm for the MOM

NOTES :

- We use the shorthand notation $P(Y_1, \ldots, Y_n)$ for $P(Y_n = y_n \,|\, y_1 = Y_1, \ldots, y_{n-1} = Y_{n-1})$.

- $\alpha_k^n(x) = P_k(X_n = x|Y_1, \ldots, Y_n)$ and $\beta_k^n(x) = P_k(Y_{n+1}, \ldots, Y_N|X_n = x)$ are probabilities computed using the current forward, backward and initial estimates ($p_k^{x\to x_0}$, $q_k^{y\to y_0}(x)$, and $\mu_k(x,y)$) of the transition and initial probabilities. These are key variables in the forward and backward propagation steps of our EM algorithm.

- The filter $\pi_k^n(x) = P(X_n = x\,|\,Y_1, \ldots, Y_n)$ and $\chi_k^n(x) = \beta_k^n(x)\cdot\frac{P(Y_1,\ldots,Y_n)}{P(Y_1,\ldots,Y_{N-1})}$ are used in our EM algorithm to replace α_k^n and β_k^n of the algorithm. While $\alpha_k^n(x)\beta_k^n(\xi)$ can be the product of two unequally sized factors, π_k^n and χ_k^n are scaled to manageable factors, ensuring efficiency and avoiding the small number problem (i.e avoiding incorrect representation of data due to small samples).

In this section, we develop a recursive expectation-maximum (EM) algorithm for estimating the transition and initial probabilities of our Markov Observation Models (MOM). The main goal is to find estimates for forward probability $p_{x\to x_0}$ for all $x, x_0 \in E$, backward probability $q_{y\to y_0}(x)$ for all $y, y_0 \in O$, and initial joint probability estimates $\mu(x,y)$ for all $x \in E$, $y \in O$. Each time step is considered a transition in a discrete-time Markov chain.

Ideally, the transition probabilities can be set as follows:

$$p_{x\to x_0} = \frac{\text{transitions } x \text{ to } x_0}{\text{occurrences of } x} \tag{1}$$

$$q_{y\to y_0}(x) = \frac{\text{transitions } y \text{ to } y_0 \text{ when } x \text{ is true}}{\text{occurrences of } y \text{ when } x \text{ is true}} \tag{2}$$

However, since we cannot directly observe x or x_0 in MOM, we estimate them using the expected transitions and occurrences:

$$p_{x\to x_0} = \frac{\text{Expected transitions } x \text{ to } x_0}{\text{Expected occurrences of } x} \tag{3}$$

$$q_{y\to y_0}(x) = \frac{\text{Expected transitions } y \text{ to } y_0 \text{ when } x \text{ is true}}{\text{Expected occurrences of } y \text{ when } x \text{ is true}} \tag{4}$$

To compute these estimates, we need to compute $P(Y_0 = y, X_1 = x, Y_1, \ldots, Y_N)$, $P(X_n = x, Y_1, \ldots, Y_N)$ for all $0 \le n \le N$, and $P(X_{n-1} = x, X_n = x_0, Y_1, \ldots, Y_N)$ for all $1 \le n \le N$.

We define:

$$\alpha_0(x,y) = P(Y_0 = y, X_0 = x) \tag{5}$$

$$\alpha_n(x) = P(Y_1, \ldots, Y_n, X_n = x), \quad 1 \le n \le N \tag{6}$$

$$\beta_0(x_1,y) = P(Y_1, \ldots, Y_N, X_1 = x_1, Y_0 = y) \tag{7}$$

$$\beta_n(x_{n+1}) = P(Y_{n+1}, \ldots, Y_N, X_{n+1} = x_{n+1}, Y_n), \quad 0 < n < N-1 \tag{8}$$

$$\beta_{N-1}(x_N) = P(Y_N, X_N = x_N, Y_{N-1}) = q_{Y_{N-1}\to Y_N}(x_N) \tag{9}$$

The joint conditional probability can be constructed recursively as:

$$P(X_{n-1} = x, X_n = x_0, Y_1, \ldots, Y_N) = \alpha_{n-1}(x)\beta_{n-1}(x_0)p_{x\to x_0}\frac{\alpha_N(\xi)}{\alpha_N(\xi)} \tag{10}$$

By the Markov property, we have:

$$P(X_n = x, Y_1, \dots, Y_N) = \alpha_n(x) \sum_{x_{n+1}} \beta_n(x_{n+1}) p_{x \to x_{n+1}} \frac{\alpha_N(\xi)}{\alpha_N(\xi)} \tag{11}$$

$$P(X_0 = x, X_1 = x_0, Y_1, \dots, Y_N) = \sum_y \alpha_0(x, y) p_{x \to x_0} \beta_0(x_0, y) \frac{\alpha_N(\xi)}{\alpha_N(\xi)} \tag{12}$$

These equations allow us to recursively compute α_n and β_n using prior estimates of $p_{x \to x_0}$, $q_{y \to y_0}(x)$, and μ.

For the initial distribution, we estimate $\mu(x, y)$ using Bayes' rule:

$$\mu(x, y) = \frac{P(X_0 = x, Y_0 = y, Y_1, \dots, Y_N)}{P(Y_1, \dots, Y_N)} \tag{13}$$

To iteratively refine the estimates for $p_{x \to x_0}$, $q_{y \to y_0}(x)$, and $\mu(x, y)$, we use the following update equations:

$$p^{(t+1)}_{x \to x_0} = \frac{\sum_y \alpha^{(t)}_0(x, y) p^{(t)}_{x \to x_0} \beta^{(t)}_0(x_0, y) + \sum_{n=1}^{N-1} \sum_{x_{n+1}} p^{(t)}_{x \to x_{n+1}} \alpha^{(t)}_n(x) \beta^{(t)}_n(x_{n+1})}{\sum_y \alpha^{(t)}_0(x, y) \beta^{(t)}_0(x, y) + \sum_{n=1}^{N-1} \sum_{x_{n+1}} \alpha^{(t)}_n(x) \beta^{(t)}_n(x_{n+1})} \tag{14}$$

$$q^{(t+1)}_{y \to y_0}(x) = \frac{\sum_y 1_{Y_1 = y_0} \beta^{(t)}_0(x, y) + \sum_{n=1}^{N-1} 1_{Y_n = y} \beta^{(t)}_n(x)}{\sum_y p_{\xi \to x} \alpha^{(t)}_0(\xi, y) + \sum_{n=1}^{N-1} \alpha^{(t)}_n(x) p_{\xi \to x}} \tag{15}$$

$$\mu^{(t+1)}(x, y) = \frac{\sum_{x_1} \beta^{(t)}_0(x_1, y) p_{x_1 \to x} \mu(x, y)}{\sum_\xi \alpha^{(t)}_N(\xi) \mu(\xi, y)} \tag{16}$$

where t denotes the iteration number. These update equations are iteratively applied until convergence is achieved. The below figure shows a code output that provides a snapshot of how the algorithm works.

```
Iteration No:  19

Matrix p:

[[0.311  0.689 ]
 [0.0117 0.9883]]

Matrix q:

[[9.000e-04 0.000e+00 9.991e-01]
 [3.861e-01 2.575e-01 3.563e-01]]
New forward probability:  1.3931545687632912e-05
Difference in forward probability:  1.0995004989085946e-07

Iteration No:  20

Matrix p:

[[0.312  0.688 ]
 [0.0088 0.9912]]

Matrix q:

[[3.000e-04 0.000e+00 9.997e-01]
 [3.850e-01 2.567e-01 3.582e-01]]
New forward probability:  1.40112936024903e-05
Difference in forward probability:  7.974791485738815e-08
```

Conclusion

The converged metrics of p, q and μ are the final canonical probability estimates and are fundamental to characterize the Markov models and to make probabilistic inferences based on observed sequences. These probabilities dictate the likelihood of transitioning from one hidden state to another, the likelihood of observing a particular symbol given the current hidden state and the likelihood of starting the sequence in a particular hidden state, respectively. They are crucial to finding the most likely sequence of hidden states, given the observed sequence since they create a canonical representation of all the care providers that may or may not be defrauding insurance providers. The algorithm is a modified form of the Baum-Welch algorithm, and it has seen uses in Speech Recognition, Crypanalysis and many more areas.