SMU – PROJECT 2

Author: Michal Najman, Username: *najmami2*

The data given in the task record relations between traffic accidents and various possible influencers of such undesirable events one meets regularly on roads. The recorded variables count for instance straight-forward initiators such as the danger level on roads but also seemingly distanced influencers such as season or country.

BASELINE NETWORK

The Baseline Network was thoroughly built upon personal experience and intuition. See Figure 1 for the model's structure.

Consensually, season strongly determines weather which in turn influences condition of roads, e.g. snow is much probable in winter and worsens dramatically the road condition. Therefore, these variables were linearly linked in the presented order. As motorists often stay home or go on vocation in direct manners on the weekends, the relative frequency of journeys decreases – thus again the linear connection.

Elaborating on an average speed on roads, if there is less moto-journeys taken, causing emptier roads, one may expect average speed increases. Moreover, I assume drivers in different countries tend to drive at different speeds as culture standards and national laws may influence that. Lastly, in theory if the road condition worsens, automobiles slow down – all three variables stated above are linked to the average speed node in a converging connection.

I propose, a danger level on road is influenced by increased average speed, road condition and a police activity which may help to reduce dangerous events on the roads. And finally, if roads are dangerous, one may expect more accidents and, consecutively, if accidents occur more often then, sadly, there is an increased chance that those might happen to be fatal.

CONDITIONAL INDEPENDENCE

In the table below, three statements involving conditional independence are examined. The graphical interpretation of the findings is shown in Figure 2.

	Statement	*Validity*	*Intuitive explanation*
(1)	Season ⊥ NoFatalities | NoJourneys	False	If one observes NoJourneys, the frequency of fatal accidents still depends on Season.
(2)	Weather ⊥ NoAccidents | RoadCond	True	If a road condition is bad, we assume that now there is no link between weather and accidents.
(3)	Season ⊥ Weekend | NoAccidents	False	If NoAccidents is observed, it opens a link between Season and Weekend as those two are distanced influencers of NoAccidents.

Missing Data Problem

The dataset was split into three parts: an initial training set, a test set and a missing-values set. The last stated was completed using the EM algorithm and the MLE method iteratively. Due to relatively long duration of model inference, the number of iterations was limited to 10.

CPTs Analysis

CPTs for the Baseline Model are derived from a dataset merging the initial training set and now-completed missing-values set (together called *the full training set*). The conditional probabilities table of the node AvgSpeed is saved in *cpt-avgspeed.txt*.

Analyzing the learnt values, one can see clearly that the value of RoadCond substantially influences the distribution of the average speed. Concretely, if the road condition is bad, the probability of the average speed being low is roughly 76% in every realization of the other variables. Conversely, if the road condition is rather good, the probability of low speeds decreases to 38% – again in all realizations of remaining variables.

Yet, Country and NoJourneys seem to have little or zero effect on the distribution of the average speed. Concretely, the probability of the average speed being low oscillates around same percentages no matter the value of Country or NoJourneys – 76% if RoadCond is bad and 38% otherwise. The largest offsets are seen in samples recorded in Europe, however, still having low significance. This suggests Country, or NoJourneys, or most likely both were linked to AvgSpeed mistakenly when forming the Baseline Model.

Final Network

The final model was derived from the Baseline Network using Hill Climbing Search on the full training set. The Baseline Network achieves -8851 K2 score and the extended model with optimized structure gets -8739 K2 score, both on the test set.

The learnt structure exposes extra and missing edges in the Baseline Network. For example, the link between NoJourneys and AvgSpeed was removed by the algorithm – this follows the prior findings in the CPTs analyzis. Conversely, the search added a link from NoJourneys to NoAccidents as this intuitive relation had to be remained after removing the previously mentioned edge.

The Final Network was largely derived from the optimized structure. Nevertheless, few changes needed to be made. First, some edges were reversed to capture the intuitive causal influence. Second, an edge from Weekend to NoJourneys was added as it intuitively makes sense for it to be there. Also note that the final structure contains a node that is not connected to any nodes – Country. It seems from the data this variable is independent on the others. The structure of the Final Network is depicted in Figure 1.

Last but not least, the samples generated by the Final Network and the completed dataset were compared. Their distributions differ slightly, Jenses-Shannon divergence is 0.108 and total variation distance is 0.00566.

Appendix

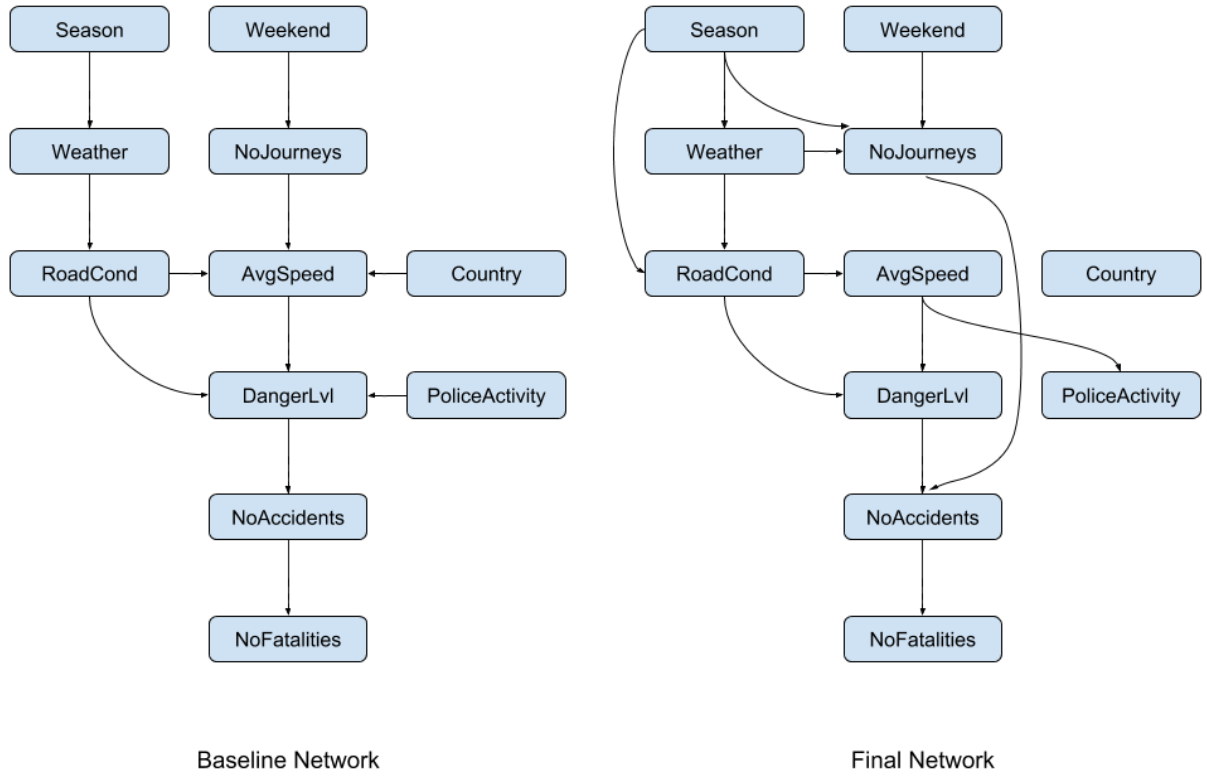

Figure 1 – Baseline Network and Final Network structures

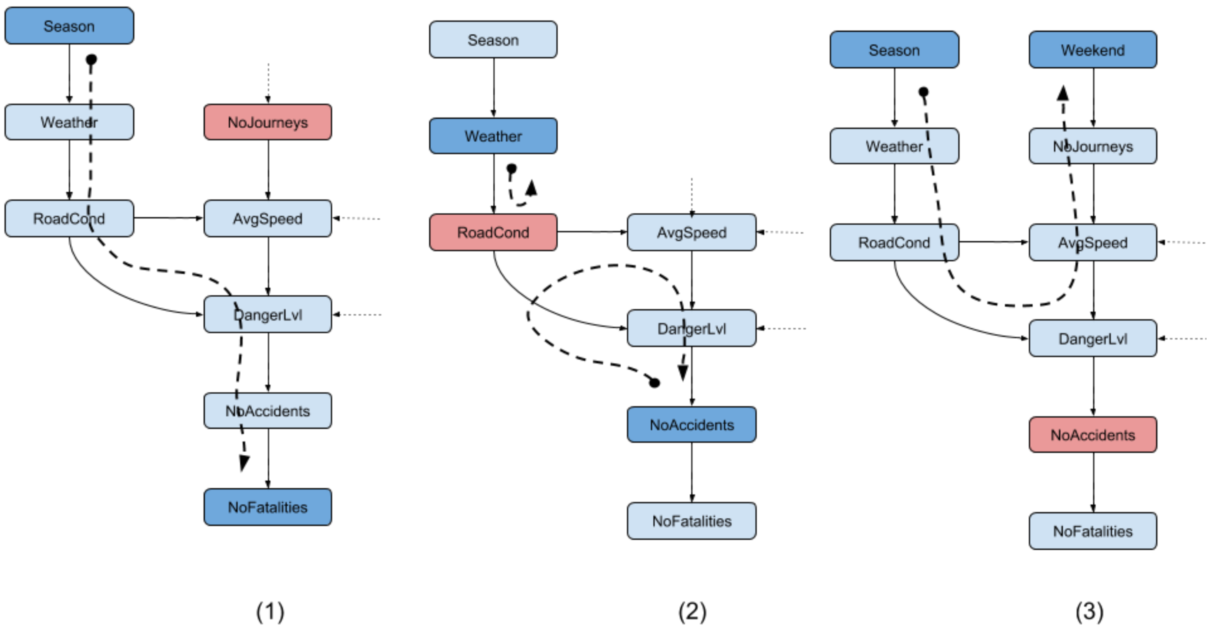

Figure 2 - Information Flow Diagram for given instances