Relatório

Artur Manuel Pascoal Ferreira

Eliminação de Variáveis em Redes Bayesianas

1 - Ler a rede usando um formato padrão

Como aconselhado pelo professor, foi escrito um parser para o formato BIF. Para o fazer, foi utilizado e modificado código do seguinte ficheiro:
https://github.com/eBay/bayesian-belief-networks/blob/master/bayesian/examples/bif/bif_parser.py.
A função parser recebe o nome do ficheiro e retorna uma lista de nós sendo cada um destes um objeto Node que contém o nome do nó, uma lista com os nós pais, uma lista com os nós filhos e a sua tabela de probabilidades.

2 - Implementar o algoritmo. O utilizador faz uma query e o programa deverá retornar uma lista de números

O algoritmo encontra-se no ficheiro variableElimination.py e a função chama-se elimination_algorithm. Esta função recebe como argumentos a lista de factores na rede, a lista de variáveis que aparecem na query, a lista de valores da evidência e o método de ordenação. Primeiro, ordena-se a lista de factores e determina-se quais são para eliminar. Para cada um destes factores, encontramos os relacionados inserindo-os numa lista e ordenamos esta. Tendo feito isto, podemos eliminar a variável usando a função eliminate que recebe como argumentos a variável a eliminar e a lista de fatores relacionados e retorna um novo factor. Aqui faz-se o join das tabelas, calculando as novas probabilidades e no final é feita a agregação dos valores da tabela por todas as variáveis tirando a que estamos a eliminar, somando as probabilidades. O resultado disto é o novo factor. Com isto, retiramos os factores usados anteriormente da lista de factores e seguimos para a próxima variável. Depois de remover todas as variáveis, ficamos com o fator final que após agregar os seus valores nos deixa com a resposta à query feita.

3 - Comparar estratégias de seleção diferentes

Foram implementadas três estratégias diferentes: número de variáveis no fator, tamanho da tabela do fator e aleatório. Para as testar, executamos três queries diferentes (sem evidência, com evidência de uma variável e com evidência de duas variáveis) em duas redes (ASIA e CANCER) cinco vezes e fizemos a média do tempo de execução. Os resultados mostram que como esperado, o aleatório é muito inconsistente isto é, por vezes é melhor que as outras duas estratégias mas também existem casos onde é pior. Relativamente às outras foi observado que, em geral, ordenar por tamanho da tabela é melhor que ordenar por número de variáveis o que faz sentido visto que a primeira tem a informação da segunda implícita.