Universidade de São Paulo – USP
Instituto de Ciências Matemáticas e de Computação – ICMC
Departamento de Ciências de Computação – SCC

SCC-5949 – Inteligência Artificial II

Professor Gustavo Batista
gbatista@icmc.usp.br

Projeto – Redes Bayesianas

Data de entrega: 30/06

Este projeto é individual. A entrega deve ser realizada via TIDIA, fazendo *upload* na pasta "projeto" na ferramenta escaninho. Este projeto é baseado no projeto de disciplina de Peter Lucas, Institute for Computing and Information Sciences, Radboud Universiteit.

O objetivo deste projeto é implementar duas redes bayesianas calculando as probabilidades marginais e condicionais a partir de dois conjunto de dados. As redes bayesianas devem ser implementadas no programa disponível no site AI Space.

O domínio de aplicação deste projeto é o diagnóstico médico, mas especificamente de diagnóstico de câncer não Hodgkiniano, um tipo raro de câncer no estômago, e câncer de mama que afeta uma parte significativa da população mundial.

Diagnóstico de câncer não Hodgkiniano

O conjunto de dados da NHL (*non-hodgkin's lymphoma*) incorpora variáveis que são amplamente utilizadas pelos médicos na escolha de uma terapia adequada. A relevância da maioria dessas variáveis é apoiada pela literatura sobre fatores prognósticos de NHL gástrica.

Em primeiro lugar, a informação utilizada no manejo clínico da NHL gástrica primária foi subdividida em informações de pré-tratamento, ou seja, informações necessárias para seleção de tratamento, informações de tratamento, ou seja, as várias alternativas de tratamento e informações pós-tratamento, ou seja, efeitos colaterais e resultados de tratamento a longo prazo para a doença. As variáveis de pré-tratamento mais importantes são a variável "*clinical stage*", que expressa a gravidade da doença de acordo com uma classificação clínica padrão e "*histological classification*", que representa a avaliação por um patologista do tecido tumoral obtido a partir de uma biópsia.

Vários tratamentos são utilizados para NHL gástrica, como quimioterapia, radioterapia e uma combinação destes dois, que foi representada como a única variável "ct & rt-schedule" com possíveis valores: quimioterapia (*CT*), radioterapia (*RT*), quimioterapia seguida por radioterapia (*CT-next-RT*), e nem quimioterapia nem radioterapia (*none*). Além disso, a cirurgia é uma terapia que é modelada pela variável "*surgery*" com possíveis valores: "*curative*", "*palliative*" ou "*none*", na qual a cirurgia curativa significa ressecção total ou parcial do estômago com

remoção completa da massa tumoral. Finalmente, a prescrição de antibióticos também é possível.

As variáveis pós-tratamento mais importantes são a variável "*early result*", sendo o resultado endoscopicamente verificado do tratamento, seis a oito semanas após o tratamento (os possíveis resultados são: "*complete remission*" - ou seja, células tumorais não são mais detectáveis – "*partial remission*" – algumas células de tumor são detectáveis -, "*no change*" ou "*progressive disease*") e a variável "*5-year result*", que representa a sobrevivência do paciente ou não após cinco anos de tratamento. Uma rede bayesiana com distribuição de probabilidade a priori para NHL gástrica é mostrada na Figura 1.

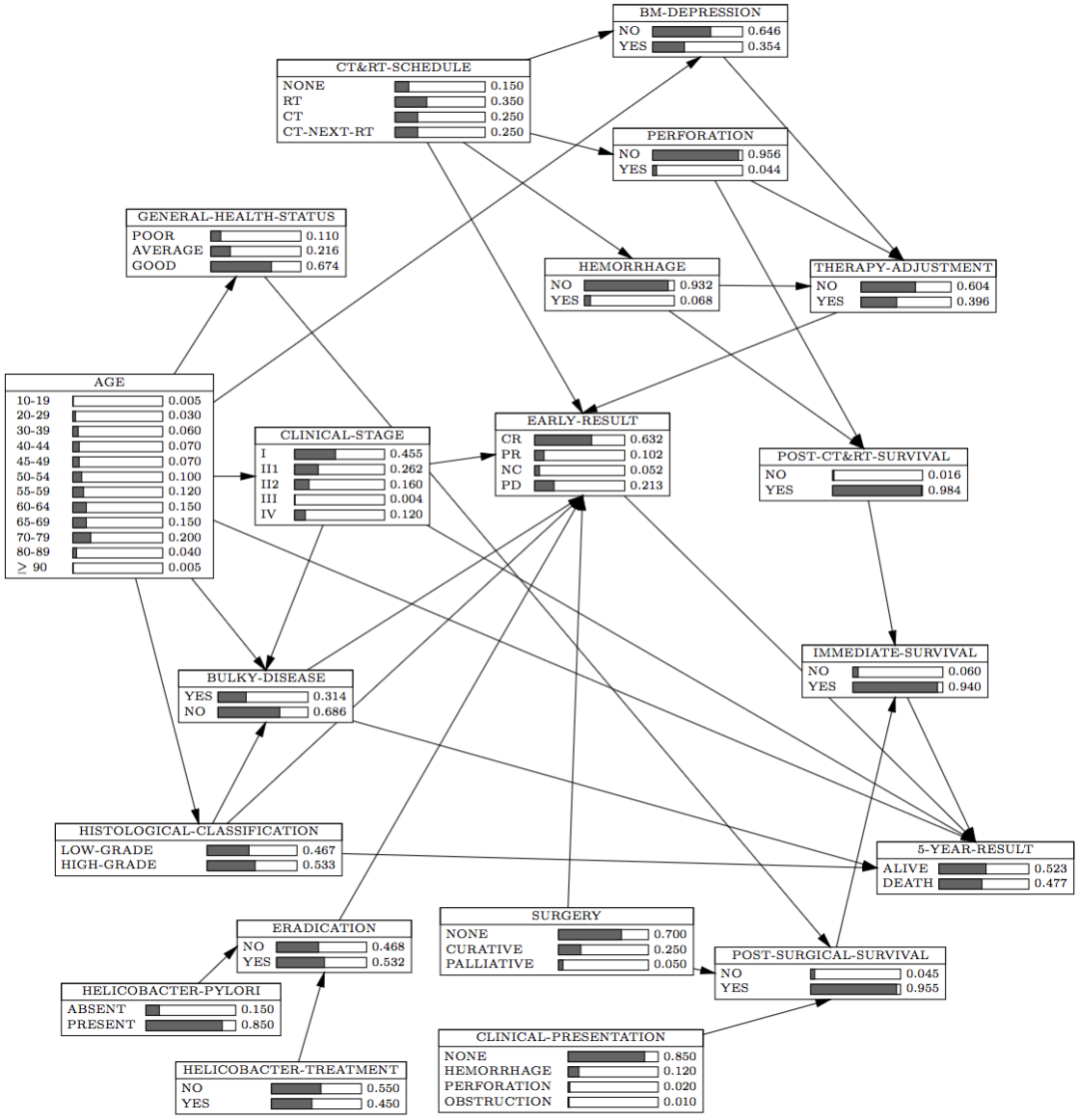

Figura 1: Rede bayesiana para o NHL gástrica.

Diagnóstico de câncer de mama

O câncer de mama é a forma mais comum de câncer e a segunda principal causa de morte por câncer em mulheres. Uma em cada 9 mulheres desenvolverá câncer de mama em seu tempo de vida. Embora não seja possível dizer o que causa exatamente o câncer de mama, alguns fatores podem aumentar ou alterar o risco de desenvolvimento de câncer de mama. Estes incluem idade, predisposição genética, histórico de câncer de mama, densidade mamária e fatores de estilo de vida. A idade, por exemplo, é o maior fator de risco para o câncer de mama não hereditário: as mulheres com idade igual ou superior a 50 anos têm maior chance de desenvolver câncer de mama do que as mulheres mais jovens. A presença de genes BRCA1 / 2 leva a um risco aumentado de desenvolver câncer de mama, independentemente de outros fatores de risco. Além disso, as características da mama, como a alta densidade da mama, são fatores determinantes para o câncer de mama.

A principal técnica utilizada atualmente para a detecção de câncer de mama é a mamografia, uma imagem de raios-X da mama. Baseia-se na absorção diferencial de raios-X entre os vários componentes tecidos da mama, como gordura, tecido conjuntivo, tecido tumoral e calcificações. Em uma mamografia, os radiologistas podem reconhecer o câncer de mama pela presença de uma massa focal, distorção arquitetônica ou microcalcificações. As massas são achados localizados, geralmente assimétricos em relação à outra mama, distintos dos tecidos circundantes. As massas em uma mamografia são caracterizadas por uma série de características, que ajudam a distinguir entre massas malignas e benignas (não cancerosas), como tamanho, margem, forma. Por exemplo, uma massa com forma irregular e margem mal definida é altamente suspeita para o câncer, enquanto que uma massa com forma redonda e margem bem definida provavelmente será benigna. A distorção arquitetônica é a ruptura focal do padrão normal de tecido mamário, que aparece em uma mamografia como uma distorção em que os tecidos mamários circundantes parecem ser "puxados para dentro" para um ponto focal, levando muitas vezes a espiculação (estruturas estelares). As microcalcificações são pequenos pedaços de cálcio, que podem aparecer em aglomerações, ou em padrões (como círculos ou linhas) e estão associados à atividade celular extra no tecido mamário. Eles também podem ser benignos ou malignos. Também se sabe que a maioria dos cancros estão localizados no quadrante superior da mama. Finalmente, o câncer de mama é caracterizado por uma série de sintomas físicos: secreção mamilar, retração da pele, nódulo palpável.

O câncer de mama se desenvolve por etapas. O estágio inicial é referido como in situ ("no lugar"), o que significa que o câncer permanece confinado ao seu local original. Quando invadiu o tecido adiposo circundante e, possivelmente, se espalhou para outros órgãos ou linfaplasma, a chamada metástase, é referido como câncer invasivo. Sabe-se que a detecção precoce de câncer de mama pode ajudar a melhorar as taxas de sobrevivência. As técnicas computadorizadas parecem ajudar especialistas médicos a esse respeito. As redes bayesianas são especialmente úteis, dada a incerteza e complexidade na análise mamográfica. A Figura 2 apresenta um modelo causal para o diagnóstico de câncer de mama com base no conhecimento apresentado acima.

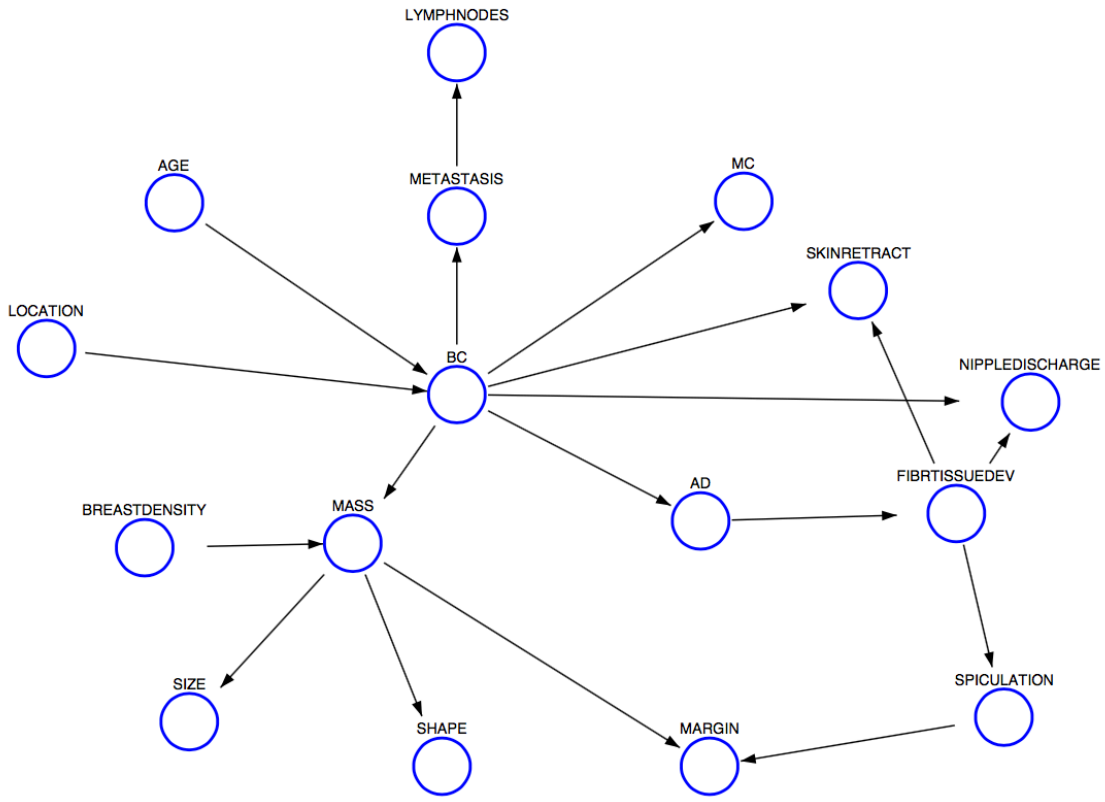

Figura 2: Rede bayesiana para o câncer de mama.

Este projeto possui três partes descritas a seguir:

Parte I – Calculo das probabilidades a partir de dados e das redes bayesianas fornecidas (30% da nota final)

Faça um programa que estime as probabilidades marginais e condicionais para as duas redes bayesianas de diagnóstico de câncer. Os conjuntos de dados de câncer não Hodgkiniano possui valores desconhecidos identificados por "NA" (not available). Os dados de câncer de mama foram gerados por amostragem a partir da rede bayesiana e, portanto, não possuem valores desconhecidos.

Existem diversas abordagens para tratamento de valores desconhecidos. Uma possibilidade simples é descartar os registros com valores desconhecidos, o que pode ser um problema para conjuntos de dados pequenos. Outra possibilidade é a imputação, ou seja, a substituição de valores desconhecidos por valores estimados. A média do atributo pode ser utilizada como valor estimado, mas isso pode distorcer a variância do atributo. Médias condicionais tentem a fornecer melhores estimativas.

Parte II – Construção da rede bayesiana (20% da nota final)

A partir das tabelas de probabilidades calculadas no item anterior, construa as redes bayesianas apresentadas as Figuras 1 e 2 utilizando o utilitário disponível no site AI Space (http://www.aispace.org).

Utilize a rede bayesiana para identificar fatores mais importantes que levam a morte dos pacientes ou a metástases. Utilize a rede bayesiana para realizar consultas do tipo “e se” contrastando fatores positivos (por exemplo, paciente jovem) e negativos (por exemplo, uma condição severa) e como esses fatores influenciam na gravidade ou sobrevivência ao câncer.

Parte II – Relatório (50% da nota final)

Escreva um relatório de até 10 páginas explicando as suas decisões de projeto. Em particular, descreva como você realizou o tratamento dos valores desconhecidos. Discuta os resultados obtidos nas suas análises e as possíveis limitações dos dados (como o número restrito de exemplos da base NHL), presença de valores desconhecidos, dados sintéticos da base de câncer de mama, etc.