Gabriela Sarti Kinker

Doutorando em Fisiologia Geral pelo Instituto de Biociências da Universidade de São Paulo. Atuo nas áreas de Fisiologia e Cronofarmacologia, com ênfase no estudo da implicação do sistema melatonérgico sobre a determinação do grau de agressividade de tumores cerebrais.

Exercícios

Introdução ao R - exercicios_aula_1.r

Análise exploratória - exercicios_aula_4.r

Criação e edição de gráficos - exercicios_aula_5.r

Regressão linear simples e análise de covariância - exercicios_aula_7.r

Regressão linear múltipla - exercicios_aula_7b.r

Reamostragem e simulação - exercicios_aula_8.r

Funções - exercicios_aula_9.r

Propostas do Trabalho Final

Plano A

Receiver Operating Characteristic (ROC), ou curva ROC, é uma representação gráfica que ilustra a performance de um sistema classificador binário de acordo com a variação de seu *cutoff* de descriminação. Cada ponto da curva ROC representa a relação, normalmente antagônica, entre as taxas de verdadeiro-positivos (sensibilidade) e falso-negativos (1 – especificidade) observadas utilizando-se um determinado valor de *cutoff*. Na clínica médica, o cálculo da área sob a curva ROC (AUC) é comumente utilizado para avaliar e comparar a sensibilidade e especificidade de testes diagnósticos, representando uma medida de precisão discriminativa. Além disso, curvas ROC são empregadas na tradução de uma variável diagnóstica quantitativa em um teste clínico dicotômico, por meio da identificação do valor de *cutoff* óptimo para a estratificação dos pacientes em subgrupos.

Dessa forma, a função que pretendo elaborar empregará curvas ROC tempo-dependente[1)], implementadas pelo pacote survivalROC, para a identificação do valor de *cutoff* de uma variável diagnóstica quantitativa que otimize a dicotomização dos pacientes com base em seus respectivos tempos de sobrevivência. Para tanto, utilizarei o método descrito por Adam et al., 2008[2)], que consiste em:

1. Avaliar a precisão da variável diagnóstica plotando-se as AUCs de curvas ROC calculadas para diferentes intervalos de tempo (ex: 10, 20, 30 meses, etc..).
2. Identificar o ponto do tempo em que a variável apresente maior precisão prognóstica (valor máximo de AUC).
3. Plotar a curva ROC para o ponto do tempo selecionado e identificar o valor óptimo de *cutoff*.
    1. Para AUC > 0.5, selecionar o ponto da curva com menor distância da coordenada (0, 1).
    2. Para AUC < 0.5, selecionar o ponto da curva com menor distância da coordenada (1, 0).
4. Construir uma curva Kaplan-Meire de sobrevivência e comparar, por meio do teste log-rank, os dois subgrupos de pacientes estratificados com base no *cutoff* óptimo calculado (pacientes abaixo do *cutoff* vs. pacientes acima do *cutoff*).

Entradas da função

Objetos:

- Vetor com os valores da variável prognóstica quantitativa (ex. expressão do gene *TP53* de 300 pacientes com câncer de pulmão).
- Vetor com o evento de cada paciente.
    - 1: ocorreu o evento de morte.
    - 0: paciente censurado (i. estudo foi finalizado e não ocorreu o evento de morte, ii. perda de seguimento ou iii. saída do estudo).
- Vetor com o tempo de morte/censura (em dias) de cada paciente.

Argumentos:

- Tipo da curva ROC tempo-dependente (KM ou NNE).
- Plotar a curva AUC vs. tempo?
- Plotar a curva ROC para o tempo selecionado?
- Plotar um histograma da distribuição dos dados, mostrando o valor de *cutoff* selecionado?
- Plotar a curva de Kaplan-Meire?

Saídas da função

- Os gráficos selecionados pelo usuário.
- Uma lista contendo:
    - O tempo de maior precisão (AUC máxima) da variável.
    - O valor da AUC máxima.
    - O valor de *cutoff* selecionado.
    - A taxa de verdadeiro-positivo (sensibilidade) e falso-negativo (1-sensibilidade) associado ao *cutoff* selecionado.
    - O valor de p e o risco relativo (+/- 95% de intervalo de confiança) calculado pelo teste log-rank comparando os dois subgrupos de paciente estratificados com base no *cutoff* selecionado (ex. pacientes com expressão de *TP53* < *cutoff* vs. pacientes com expressão de *TP53* > *cutoff*).
    - Média, mediana, e intervalo de 95% de confiança da variável prognóstica quantitativa para cada um dos dois subgrupos de pacientes (média, mediana, e intervalo de confiança da expressão de *TP53* dos pacientes com expressão acima vs. abaixo do *cutoff*).

*Pacotes utilizados:

- Curva ROC tempo-dependente - survivalROC

- Análises de sobrevivência - survival

Plano B

Na última década, a análise de expressão gênica global tornou-se um dos pilares da pesquisa na área de biologia molecular e genômica. O desafio não reside mais em obter os perfis de expressão gênica e sim em interpretar os resultados de maneira a gerar *insights* sobre o sistema biológico investigado. Nesse contexto, para demonstrarmos a relevância biológica de alterações nos níveis de expressão é necessário compreender como os produtos gênicos interagem entre si, formando complexos ou redes. Sendo assim a análise de conjuntos de genes que compartilhem aspectos funcionais, localização cromossômica ou vias de sinalização mostra-se mais informativa do que a análise de genes individuais.

Como plano B, pretendo elaborar uma função que permita automatização da análise comparativa entre dois grupos de amostras de microarray/RNA-seq quanto à expressão de genes alvo de NFκB, um dos principais fatores de transcrição envolvidos no desenvolvimento e progressão tumoral. Dentre os mais de 200 genes alvo de NFκB, selecionarei um grupo de cerca de 50 genes, os quais participam da regulação de processos como proliferação e migração celular, apoptose e angiogênese. Os objetivos gerais da função consistem em:

1. Comparar a expressão dos genes alvo de NFκB entre dois grupos de amostra e gerar uma representação gráfica.
2. Comparar as redes de co-expressão dos genes alvo de NFκB entre dois grupos de amostra e gerar uma representação gráfica.

Entradas da função

Objetos:

- Dataframe contendo os dados de expressão gênica global pré-processados. Amostras nas linhas e genes nas colunas (ex. RNA-seq de 300 amostras de câncer de pulmão).
- Vetor de classificação das amostras (ex. tumores resistentes ou não à quimioterapia).

Argumentos:

- Filtrar os genes alvo de NFκB por processo biológico (proliferação, migração, apoptose, angiogênese ou todos).
- Tipo de dados (microarray ou RNAseq) - determina a escolha do algoritmo para a análise de expressão diferencial.
- Método para a inferência da rede de co-expressão (Pearson, Spearman, ou Kendall).
- Tipo de rede de co-expressão (*weigthed* ou *unweigthed*).
- Número de permutações para o cálculo do p-valor da comparação de distribuição espectral dos networks dos dois grupos de amostra.

Saída da função

- Uma lista contendo:
  - P-valor e *fold change* da análise de expressão diferencial dos genes alvo de NFκB selecionados.
  - P-valor da comparação das redes de co-expressão dos genes alvo de NFκB selecionados.
- *Heatmap* comparando os dois grupos de amostra quanto à expressão dos genes alvo de NFκB

selecionados.
- Para cada grupo de amostra, uma rede de co-expressão dos genes alvo de NFκB selecionados.

*Pacotes utilizados:

- Comparação da expressão gênica: edgeR para RNAseq e limma para microarray
- Comparação de redes de co-expressão: psych e CoGa
- Representações gráficas: gplots e igraph

Comentários — *Alexandre Adalardo de Oliveira 2016/04/28 09:06* Gostei de ambas as propostas. Elas estão claramente descritas e parecem interessantes. A primeira acho mais geral e interessantes. Minha sugestão é que tente fugir dos pacotes quando for possível. Por exemplo, a estimativa de sobrevivência de Kaplan-Meier é simples de implementar e pode ser construída na sua função e não precisa ser chamada no pacote `survival`. Veja o capítulo dois do livro no link abaixo, mais especificamente a tabela 2.3: Applied Survival Analysis

Questões menores:

- não entendi no tópico 4 da proposta o que são os dois subgrupos de paciente, deixe isso claro. Não há na entrada da função nada falando sobre uma variável que classifica grupos.
- além do teste, retorne uma statistical descritiva para dar uma ideia do tamanho do efeito: p.ex. mediana +- 95% de intervalo de confiança
- duvida: "censurada" é o dado `right censored` no seu caso. Não conheço a tradução dos termos de análise de sobrevivência. De qq forma, como não é algo trivial, deixe claro que são os dados (indivíduos) onde o evento (morte) não ocorreu durante a duração do experimento.

BOM TRABALHO!

Olá, Professor. Obrigada pelos comentários. Alguns esclarecimentos:

- No tópico 4 os subgrupos referem-se a pacientes estratificados com base no *cutoff* óptimo calculado no item 3. Portanto, nesse tópico comparamos a taxa de sobrevivência de, por exemplo, pacientes com expressão de *TP53* < cutoff vs. pacientes com

expressão de *TP53 > cutoff*.
- “Censurado” é de fato o “Right Censored” - adicionei uma pequena explicação na proposta para deixar mais claro o significado da censura.
- Adicionei as estatísticas descritivas na saída da função.

Trabalho Final

O código da função, o arquivo da função, e a página de ajuda estão disponíveis no link abaixo:

exec

1)
Heagerty, P. J., Lumley, T. and Pepe, M. S. (2000), Time-Dependent ROC Curves for Censored Survival Data and a Diagnostic Marker. Biometrics, 56: 337–344. doi: 10.1111/j.0006-341X.2000.00337.x
2)
Adams, H., Tzankov, A., Lugli, A., & Zlobec, I. (2009). New time-dependent approach to analyse the prognostic significance of immunohistochemical biomarkers in colon cancer and diffuse large B-cell lymphoma. Journal of clinical pathology, 62(11), 986-997.