MAE5776 – Análise Multivariada 1º Sem/2020

Questões - Consolidando o Aprendizado

1. Em análise Multivariada, por que é importante entender a estrutura dos dados?

2. Justifique a afirmação: "Muitas das análises de dados multivariados envolvem direta ou indiretamente a matriz de distâncias entre observações".

3. Como a teoria de Espaços Duais pode ser útil na redução de dimensionalidade de dados sob a estrutura *Big_p*? Como essa redução de dimensionalidade pode ser formulada no contexto de ajustes de regressão regularizados e penalizados?

4. O *Biplot* é uma ferramenta útil na visualização de dados multivariados. Explique por que e como esse gráfico é construído.

5. Justifique a afirmação: "As técnicas matriciais de Decomposição em Valores Singulares, bem como de Decomposição Espectral, são a base de muitas das análises de redução de dimensionalidade".

6. Como a Análise de Correspondência pode ser formulada a partir da Análise de Coordenadas Principais?

7. Na Análise de Fatores (Análise Fatorial), como estão definidos os fatores comuns e os específicos? Você pode usar a solução via Componentes Principais para responder.

8. Na Análise de Fatores (Análise Fatorial), por que e como os vetores de cargas (*loadings*) são rotacionados? Você pode usar a solução via Componentes Principais para responder.

9. Na análise de redução de dimensionalidade de uma matriz de dados Ynxp, os Componentes Principais satisfazem quais propriedades? O que garante que dois componentes reduzem bem os dados?

10. Na análise de redução de dimensionalidade de uma matriz de dados Ynxp, os Eixos Discriminantes da Solução Linear de Fisher satisfazem quais propriedades? O que garante que dois eixos discriminantes reduzem bem os dados?

11. Na análise de redução de dimensionalidade de uma matriz de dados Ynxp, com $p=p_1+p_2$, os Eixos Canônicos da Correlação Canônica satisfazem quais propriedades? O que garante que o primeiro par desses eixos reduzem bem os dados?

12. Considere a redução de dimensionalidade de uma matriz de dados Ynxp. Se $n<<p$, quais são os problemas na realização da Análise de Componentes Principais "Clássica"? Que alternativas de análise podem ser usadas?

13. Considere a redução de dimensionalidade de uma matriz de dados Ynxp, com $n=n_1+n_2+...+n_G$. Se $n<<p$, quais são os problemas na realização da Análise Discriminante "Clássica" de Fisher? Que alternativas de análise podem ser usadas?

14. Considere a redução de dimensionalidade de uma matriz de dados Ynxp, com $p=p_1+p_2$. Se $n<<p$, quais são os problemas na realização da Análise de Correlação Canônica "Clássica"? Que alternativas de análise podem ser usadas?

15. Na MANOVA qual é a importância da equação envolvendo o seguinte determinante: $|S_B - \lambda S_W| = 0$, em que S_B e S_W são matrizes quadradas (pxp) conhecidas de soma de quadrados e produtos cruzados dos efeitos Entre e Dentro de grupos, respectivamente, e $\lambda \in \Re^+$ é tal que $(S_W^{-1} S_B)V = \lambda V$, com $V \in \Re^p$?

16. Na análise de uma matriz de dados multivariados Ynxp, como a distância de Mahalanobis pode ser usada para definir Regiões de Concentração dos dados e Regiões de Confiança para o centroide? Como pode ser proposto um critério de diagnóstico de observação atípica (*oulier*)?

17. Na análise de uma matriz de dados multivariados Ynxp ~ $(\mu;\Sigma)$, com p=2, ilustre, em um gráfico de dispersão, possíveis diferenças entre os Intervalos de Confiança Univariados Clássicos, Intervalos de Confiança Univariados Simultâneos, Intervalos de Confiança Univariados com Correção de Bonferroni e Regiões de Confiança para o vetor μ. Compare essas quatro abordagens no contexto de correções para múltiplos testes.

18. Considere o banco de dados ***Exp*** do Projeto Final da disciplina que consiste da resposta de intensidade de expressão de muitos genes avaliados em 1000 pacientes (Caso e Controle) com câncer de ovário. Considere ainda que na pré-análise desses dados o primeiro componente principal foi obtido, o qual explicou 10% da variância total dos dados. Na sequência da análise a resposta normalizada, ***NorExp***, foi obtida com base na seguinte transformação dos dados:

$$y_{ij} - \nu_{1j} z_{1i},$$

em que, y_{ij} é a intensidade de expressão do j-ésimo gene avaliada no i-ésimo paciente, z_{1i} é o escore desse paciente no primeiro componente principal e ν_{1j} é o j-ésimo coeficiente do correspondente autovetor. Finalmente, nos dados normalizados, **NorExp**, suponha que testes "t" foram aplicados aos dados de cada gene, com o objetivo de identificar os genes significantes.

(a) Comente sobre quais possíveis fontes de variação devem estar sendo explicadas pelo primeiro componente principal calculado.

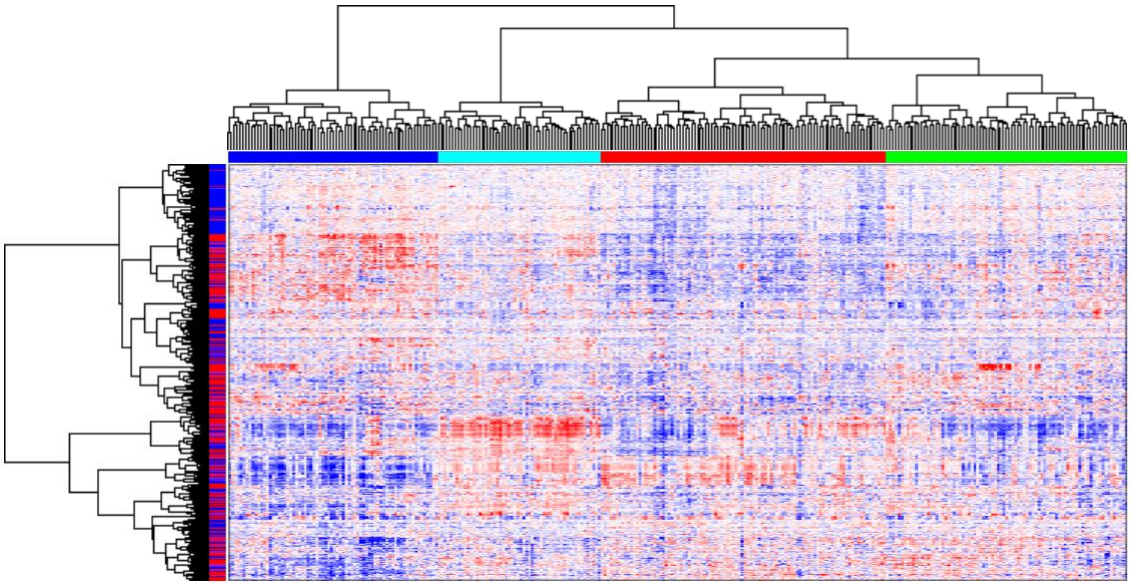
(b) Justifique por que essa transformação dos dados deve ter sido usada.

(c) Apresente críticas a essa estratégia de análise de normalização e significância em um estudo multivariado.

(d) Proponha alternativas a essas análises.

19. Considere os dados reais de intensidade de expressão de vários "genes" avaliados em 309 pacientes com câncer de ovário do Projeto TCGA, dos quais, parâmetros foram extraídos para geração dos dados simulados disponibilizados no Projeto Final da disciplina. Para visualização dos dados reais o *heatmap* foi construído e está apresentado a seguir. Nas linhas estão indicados os "genes" e nas colunas os pacientes. Os agrupamentos foram formados segundo o método de ligação completa (vizinho mais distante). As cores no interior do gráfico indicam valores da resposta padronizada (de azul para menos expresso, a vermelho para mais expresso). A barra vertical de cores indica genes alvo para câncer de ovário (em azul) e genes altamente

variáveis (em vermelho). A barra horizontal de cores indica os quatro grupos de pacientes formados.



(a) Que informações podem ser extraídas deste gráfico?

(b) Se a formação dos grupos de pacientes, bem como dos grupos de variáveis ("genes"), fosse feita por meio de outro método de agrupamento hierárquico, seria esperado obter o mesmo padrão de similaridade no heatmap? Justifique.

(c) Caso os pacientes estivessem estratificados segundo seu tempo de sobrevivência (Caso: pacientes com tempo de sobrevida menor que 3 anos; Controle: pacientes com sobrevida maior que 3 anos), como essa informação poderia ser acrescentada no heatmap?

20. Considere os cinco bancos de dados do projeto Final da disciplina (CNV, Exp, NorExp, Methy, Protein) contendo informação de 1000 pacientes com câncer de ovário estratificados de acordo com o tempo de sobrevida. Com o objetivo de integrar os dados de expressão gênica (**NorExp**) e proteica (**Protein**), diferentes alternativas de análise foram usadas. Avalie as propriedades dos escores (dos pacientes) e das cargas (das variáveis) obtidas em cada caso a seguir:

(a) Foram obtidos os Top100 Componentes Principais de **NorExp**, os Top100 Componentes Principais de **Protein**, e então realizada a análise de Correlação Canônica destes dois conjuntos de Top100 componentes.

(b) Usando o gráfico Vulção construído com resultados de uma ANOVA, foram obtidas as Top100 variáveis de **NorExp** mais significantes. O mesmo procedimento foi usado para obter as Top100 variáveis de **Protein**. Então, foi realizada a análise de Correlação Canônica destes dois conjuntos de Top100 componentes.

(c) Usando o gráfico Vulção construído com resultados de uma ANOVA, foram obtidas as Top100 variáveis de **NorExp** mais significantes. O mesmo procedimento foi usado para obter as Top100 variáveis de **Protein**. Então, foi realizada a análise PLS destes dois conjuntos de Top100 componentes.

21. Considere que na análise do banco de dados de proteína (**Protein**) do Projeto Final da disciplina foi extraído um subconjunto de 2 proteínas alvo do estudo, as quais são descritas na literatura como sendo influenciadas por um CNV localizado no cromossomo 8. Os 1000 pacientes foram então estratificados em dois grupos de acordo com a categoria desse CNV (Alterado={-2,-1,1,2} e Normal={0}). A matriz de dados resultante (1000x2) foi analisada por meio de uma MANOVA com as seguintes fontes de variação: Grupo (Caso e Controle), CNV (Alterado e Normal), Interação Grupo*CNV, além do Resíduo.

(a) Apresente a tabela de MANOVA com os correspondentes números de graus de liberdade e expressões das somas de quadrados e produtos cruzados das fontes de variação analisadas (SS_Grupo, SS_CNV, SS_Interação e SS_Resíduo). Quais são as hipóteses sob teste nesse tipo de análise? Que suposições são feitas?

(b) Foram então obtidos os vetores reducionistas associados a cada uma dessas fontes de variação, isto é, aqueles obtidos da decomposição espectral de SS_Resíduo, $(SS\_Resíduo)^{-1}SS\_Grupo$, $(SS\_Resíduo)^{-1}SS\_CNV$, $(SS\_Resíduo)^{-1}SS\_Interação$. Ilustre em um gráfico de dispersão as possíveis direções ótimas dessas razões sinal-ruído. Essas reduções de dimensionalidade atendem a quais objetivos?

22. Considere a matriz de dados Ynxp e as correspondentes formas quadráticas YY' e Y'Y.

(a) Se λ é um autovalor de Y'Y com autovetor v. Mostre que λ é um autovalor de YY' com autovetor Yv (equivalentemente, com autovetor padronizado $Yv\lambda^{-1\backslash 2}$).

(b) Como esse resultado pode ser usado para relacionar Componentes Principais e Coordenadas Principais?

23. Considere a matriz de dados Ynxp e as correspondentes formas quadráticas YY' e Y'Y.

(a) Estabeleça relações entre os autovalores e autovetores da decomposição em valores singulares da matriz retangular Y e das correspondentes decomposições espectrais das formas quadráticas.

(b) Como esse resultado pode ser usado em *big-data*?

24. O seguinte texto foi extraído do livro *Eigenproblems in Pattern Recognition*: *"While Principal Component Analysis (PCA) deals with only one data space X where it identifies directions of high variance, Canonical Correlation Analysis (CCA) proposes a way for dimensionality reduction by taking into account relations between samples coming from two spaces X and Y. The assumption is that the data points coming from these two spaces contain some joint information that is reflected in correlations between them. Directions along which this correlation is high are thus assumed to be relevant directions when these relations are to be captured"*. Justifique esse resultado e indique como as direções da CCA podem ser obtidas a partir da direção de PCA.

25. O seguinte texto foi extraído do livro *Eigenproblems in Pattern Recognition*: *"Partial least squares (PLS) can be interpreted as a covariance maximizer instead of a correlation maximizer, as is the case with Canonical Correlation Analysis (CCA)"*. Justifique esse resultado e indique diferenças entre essas análises por meio de uma aplicação.

26. Use diagramas de caminhos para representar a estrutura latente de Componentes Principais, Análise Discriminante, Análise de Correlação Canônica e Mínimos

Quadrados Parciais (PLS). Compare essas técnicas relativamente aos objetivos de cada uma.

27. O objetivo da Análise Discriminante Linear de Fisher (AD) é identificar direções no espaço das variáveis que melhor separam grupos. A estrutura latente obtida dessa análise atribui maior peso para variáveis que variam pouco dentro do grupo e muito entre os grupos. No caso de dados de base familiar, isto é, contendo informação de indivíduos e seus familiares, as famílias podem ser modeladas como grupos e a estrutura de dependência pode ser capturada pelos componentes de covariância. Estabeleça equivalências entre os Componentes Principais em dados agrupados em famílias e a AD em observações independentes agrupadas.

28. Validação cruzada (*cross validation*) é um método bastante usado em algoritmos de aprendizado. A partir dos bancos de dados do Projeto Final da disciplina, defina um problema de Análise Discriminante e explique como o método de validação cruzada fold=K, para algum K, pode ser implementado.

29. Estabeleça diferenças entre os métodos de Agrupamento Ligação Completa (ou do Vizinho mais Distante) e o K-Médias. No K-Médias, como K pode ser escolhido? Como agrupar observações em dados Big-p?

30. Em análise multivariada, qual é a utilidade da matriz Σ^{-1}?