Reconhecimento de Padrões 2015 – Prova P_1 – Prof Vêncio
Entrega: 13/outubro/2015 até 23:59h no e-mail oficial rvencio@usp.br

*Prova **individual**, com consulta a qualquer material não-vivo. Valor das questões é uniforme. As resoluções devem ser acompanhadas de explicação e justificativa detalhadas, bem como os código-fonte utilizados para a solução, inclusive com output original obtido. Qualquer linguagem de programação é permitida, desde de que a lógica do código não esteja demasiado obscura para que um não-fluente possa entender.*

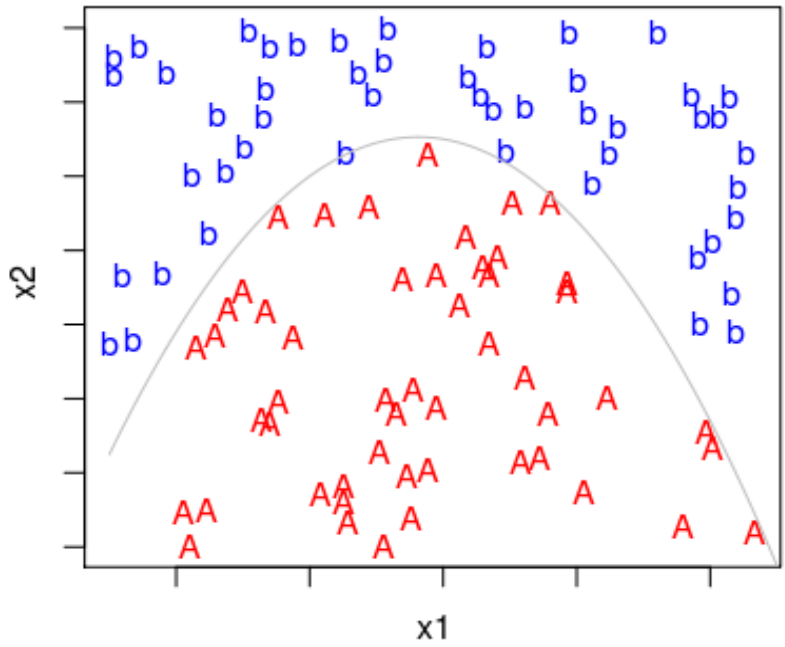
Questão 1
Durante a dedução do método de PCA a partir de seus primeiros princípios foi dito que a variância da primeira componente principal Var(ξ_1) era igual a $\mathbf{r}_1^T \Sigma \mathbf{r}_1$ (video http://goo.gl/D8dvWn *t*=6:01 até 9:12), onde $\mathbf{r}_1 = [r_{1,1}, r_{2,1}, \ldots, r_{n,1}]^T$ da matriz de rotação R, e Σ é a matriz de covariância. Mostre que isso é verdade.

Questão 2
No artigo "*BayBoots: a model-free Bayesian tool to identify class markers from gene expression data*" (Vêncio *et al* Genetics and Molecular Research 2006) os autores tentam convencer o leitor de que o método que desenvolveram é bom para identificar manifestações cardíacas (classe C: cepas B115, B147 e B13) ou assintomáticas (classe A: cepas FAMEMA e Berenice) do agente etiológico da Doença de Chagas, o *Trypanosoma cruzi*. Uma crítica válida ao artigo seria a de que autores buscaram por biomarcadores individuais que classificam os parasitas nesses dois grupos mas não analisaram os transcritomas como um todo separando as principais fontes de variância. Seria possível, utilizando os mesmos dados (ID GSE1828 no banco de dados público NCBI-GEO), separar essas duas classes apenas com as duas primeiras componentes principais por PCA? Qual é a porcentagem da variância "armazenada" nessas duas primeiras componentes principais?

Questão 3
No artigo "*Processing of voided urine for prostate cancer RNA biomarker analysis.*" (Quek *et al* Prostate 2015) os autores tentam convencer o leitor de que o RNA amplificado diretamente da urina de pacientes pode servir de fonte para análise de biomarcadores evitando assim diversos inconvenientes de coleta de urina, como por exemplo "o toque", etc, etc. A motivação é descobrir se é tecnicamente factível buscar biomarcadores moleculares a partir da amostragem mais simples possível. Uma crítica válida ao artigo seria a qualidade da parte referente a Reconhecimento de Padrões. Cite duas rasões pelas quais a figura 4 do artigo é, basicamente, inútil. Quais seus principais defeitos do ponto de vista técnico? A crítica anterior procede ou não?

Questão 4

Considere o problema de classificação bidimensional representado pela figura abaixo.



Percebendo uma tendência parabólica na fronteira de decisão, podemos propor um classificador quadrático (portanto não-linear): $a\ x_2 + b\ x_1 + c\ x_1^2 + d = 0$. Se criarmos um espaço 3D alternativo estendido (x_2, x_1, x_1^2) a partir das "dimensões" (*features*) 2D originais, o hyplerplano obtido com técnicas propostas para problemas lineares pode resolver um problema não-linear. Pegue os dados exemplo conforme seu número USP em https://goo.gl/DDpdeC. Use a técnica de minimização de erro (via integral "máscara" ou via simulação, a escolha é livre) para obter um classificador linear e transforme-o de volta numa fronteira de decisão quadrática. Quão longe ficou o resultado obtido com este "truque" da fronteira parabólica original que você estipulou? O "truque" funciona?