IA006 – Exercícios de Fixação de Conceitos
EFC 2 – 1s2019

Parte 1 – Teoria bayesiana de decisão

Considere que um conjunto de dados unidimensionais tenha duas classes subjacentes (C_1 e C_2). Os dados pertencentes à classe C_1 seguem uma densidade gaussiana de média nula e variância unitária, ou seja, $p(x|C_1)$ é uma densidade $N(0,1)$. Já $p(x|C_2)$ é uma gaussiana de média nula e variância igual a dois, ou seja, $N(0,2)$.

(a) Assuma que seja obtido um dado x'. Pelo critério de máxima verossimilhança, para quais valores de x' devemos considerar (decidir) que ele pertence à classe C_1? E à classe C_2?

(b) Considere agora que seja informado que a classe C_1 tem uma probabilidade a priori $P(C_1)$ que é duas vezes maior que a probabilidade a priori $P(C_2)$. Segundo o critério MAP, para quais valores de x' devemos considerar que o dado pertence à classe C_1? E à classe C_2?

(c) Compare os resultados dos itens (b) e (a) e teça um breve comentário.

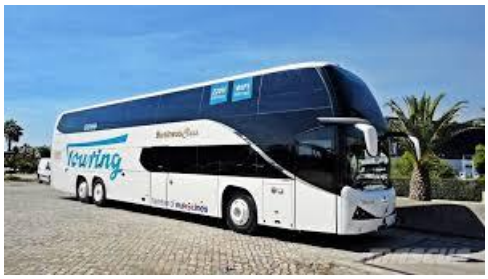
Parte 2 – Classificação binária

Considere o conjunto de dados rotulados disponível no arquivo two_moons.txt. A matriz $\mathbf{X} \in \mathbb{R}^{N\times 2}$ contém os atributos (*features*) dos N padrões existentes, enquanto o vetor $\mathbf{y} \in \mathbb{R}^{N\times 1}$ traz o rótulo da classe correspondente.

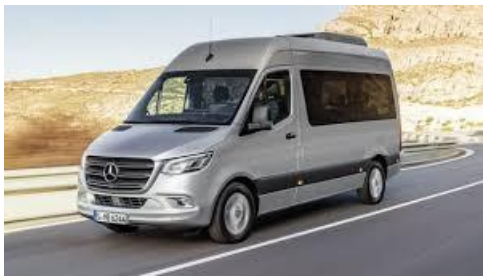
(a) Primeiramente, mostre a distribuição dos padrões no espaço original dos dados e discuta as características deste problema de classificação binária.
(b) Obtenha a direção ótima de projeção do discriminante linear de Fisher, mostrando-a junto aos dados disponíveis. Em seguida, realize a projeção e apresente os histogramas das classes projetadas, comentando o que pode ser observado.
(c) Após aplicar o discriminante de Fisher, obtenha a curva ROC variando o valor do *threshold* utilizado na etapa de decisão. Adicionalmente, exiba como a F_1-medida evolui conforme o valor do *threshold* é alterado. Discuta os resultados obtidos.
(d) Utilize, agora, o modelo de regressão logística para realizar a classificação dos padrões. Mostre a curva ROC correspondente e, semelhantemente ao item anterior, apresente a evolução da F_1-medida em função do valor do *threshold* de decisão. Comente os resultados obtidos, comparando-os com aqueles verificados para o discriminante de Fisher.

Parte 3 – Classificação multi-classe

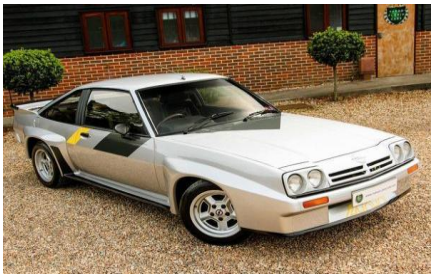
Agora, vamos considerar o problema de classificar exemplos de veículos tendo como base alguns atributos extraídos de imagens de suas silhuetas. Em particular, usaremos os dados disponíveis na base *Vehicle Silhouettes*.
https://archive.ics.uci.edu/ml/datasets/Statlog+(Vehicle+Silhouettes)
Cada padrão (veículo) é descrito por 18 atributos numéricos e existem quatro classes possíveis, conforme ilustrado na Figura abaixo.



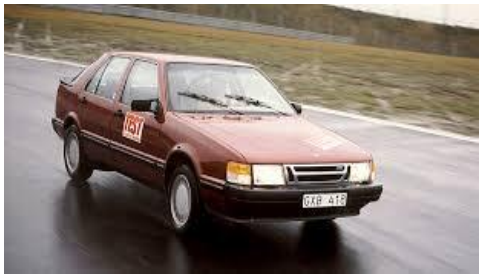
(a) *bus*



(b) van



(c) Opel Manta 400



(d) Saab 9000

Figura. Exemplos de veículos pertencentes a cada uma das classes possíveis.

Inicialmente, separe o conjunto de dados em duas partes, uma para treinamento, outra reservada para teste (*holdout*).

Dois métodos de classificação serão explorados nesta aplicação: regressão logística e *k-nearest neighbors*. No caso da regressão logística, vamos empregar uma abordagem do tipo um-contra-um (*one-vs-one*) para lidar com este cenário de classificação multi-classe.

(a) Faça o projeto de todos os classificadores binários com o modelo de regressão logística e implemente o mecanismo de desambiguação para a tomada de decisão final. Obtenha, então, a matriz de confusão para este classificador considerando os dados do conjunto de teste. Além disso, adote uma métrica global para a avaliação do desempenho (médio) deste classificador. (Consulte a referência: M. SOKOLOVA & G. LAPALME, A Systematic Analysis of Performance Measures for Classification Tasks. *Information Processing & Management*, vol. 45, no. 4, pp. 427-437, 2009). Discuta os resultados obtidos.
(b) Considere, agora, a técnica *k-nearest neighbors*. Varie o parâmetro *k* e analise as matrizes de confusão obtidas junto aos dados de teste e o desempenho

médio (computado com a mesma métrica adotada no item (a)). Comente os resultados obtidos, inclusive estabelecendo uma comparação com o desempenho da regressão logística.