UNIVERSIDADE FEDERAL DE MINAS GERAIS

ELT135 - Introdução ao Reconhecimento de Padrões

Exercício 6 - 28/08/2019

Classificador Bayesiano aplicado a um problema multivalorado e utilizando validação cruzada

Hernane Braga Pereira - 2014112627

1. Introdução

Este relatório tem como objetivo exemplificar o uso do classificador bayesiano em um problema multivalorado e de classificação binária, utilizando a técnica de validação cruzada para determinar a de acurácia do modelo encontrado.

2. Validação cruzada em um Classificador Bayesiano

Para este exercício foi utilizada a base de dados *spambase*, um modelo que possui 57 variáveis de entrada e uma saída binária, que representa se o email é um spam, ou não. O classificador utilizado foi o bayesiano para n variáveis de entrada e o modelo foi treinado dividindo-se aleatoriamente a base de dados em 10 grupos, onde 9 dos grupos eram usados para treino e o último para validação. Após uma iteração, os grupos se alternam, até que todos os dados da base tenham sido usados tanto para treino, quanto para teste. A acurácia final é encontrada através da média de acurácia de cada iteração.

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1n}\sigma_1\sigma_n \\ \rho_{21}\sigma_2\sigma_1 & \sigma_2^2 & \cdots & \rho_{2n}\sigma_2\sigma_n \\ \vdots & \vdots & \cdots & \vdots \\ \rho_{n1}\sigma_n\sigma_1 & \rho_{n2}\sigma_n\sigma_2 & \cdots & \sigma_n^2 \end{bmatrix}$$

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}$$

Figura 1. PDF multivariada utilizada no classificador

Nº Teste	Taxa de acurácia para a Classe 1	Taxa de acurácia para a Classe 2	Taxa de acurácia Total
1	96,15%	3,85%	50,00%
2	92,86%	7,14%	50,00%
3	91,76%	8,24%	50,00%
4	99,45%	0,55%	50,00%
5	98,90%	1,10%	50,00%
6	96,13%	3,87%	50,00%
7	95,58%	4,42%	50,00%
8	97,79%	2,21%	50,00%
9	94,48%	5,52%	50,00%
10	92,82%	7,18%	50,00%
Média Geral	**95,59%**	**4,40%**	**50%**

Quadro 1. Comparação de acurácia para diferentes tamanhos de amostras de testes

Ao analisar os resultados, conclui-se que a técnica k-fold para validação cruzada proporciona uma estimativa razoável da acurácia de um modelo. Neste exemplo houveram problemas na classificação das classes, isto ocorreu devido a matriz de covariância ser singular em alguns casos, não existindo sua inversa e impossibilitando o cálculo da PDF. Um melhor estudo sobre a base de dados e uma adequação das variáveis de entrada pode solucionar este problema e assim melhorar a acurácia do modelo.

3. Referências

[1] Classificador Bayesiano:, Notas de aula, agosto de 2019.