UNIVERSIDADE FEDERAL DE MINAS GERAIS
Graduação em Engenharia de Sistemas
Introdução ao Reconhecimento de Padrões - Aula 16
Nikolas Dias Magalhães Fantoni - 2018019400

Análise de Componentes Principais (PCA) e Máquina de Vetores de Suporte (SVM)

Exercício

No exercício proposto, foram obtidos dados de câncer de mama, chamados de *BreastCancer*, a partir da biblioteca do R *mlbench*. Após a retirada dos valores nulos e da coluna relativa ao *id* das amostras, obtivemos uma base de dados de 683 amostras, constituída por 9 variáveis, classificadas como tumores benignos (considerados como da classe $C_1 = -1$) e tumores malignos (considerados pertencentes à classe $C_2 = 1$).

Assim, técnica PCA foi implementada para definir as variáveis que possuíam maior variância nos dados. Assim, mesmo que com uma certa perda de valores, foi possível analisar os dados com um número menor de variáveis.

Para a técnica PCA, primeiro foram calculados os autovalores e os autovetores utilizando funções próprias do R. Na figura 1 são mostrados os maiores autovalores, em ordem decrescente, das variáveis. Só a primeira classe já corresponde com 69% da variância total do problema. Porém, para o classificador, procurou-se encontrar o número de variáveis que correspondesse a pelo menos 80% da variância do problema. Assim, foi definido que o número de variáveis considerada no PCA seria de 3.

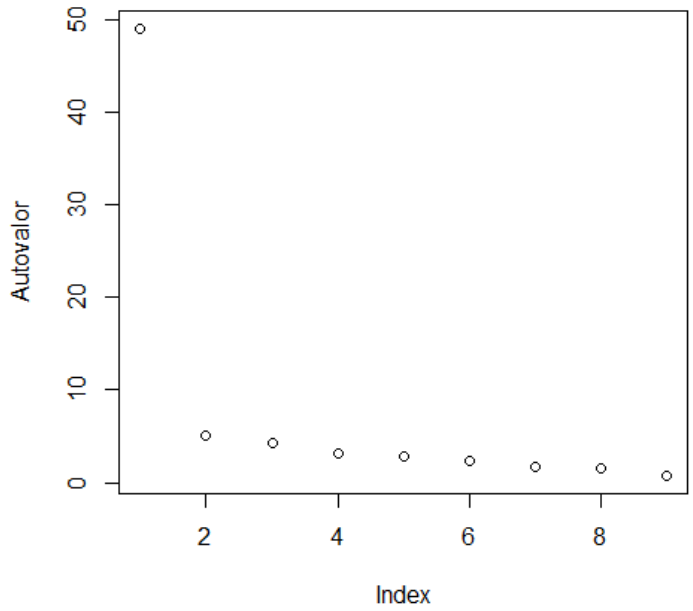

Figura 1: Autovalores dominantes dos dados de entrada.

Foram feitas 10 simulações, dividindo-se os dados em 10 *folds*, sendo que a cada iteração um *fold* seria usado como dado de teste e o restante seria utilizado para o treinamento de um classificador SVM, com

parâmetros $\sigma = 0.8$ e $C = 20$, para diminuir os erros de classificação. Para classificar os dados, foi utilizada a função *ksvm*, contida na biblioteca *kernlab* do R. As acurácias obtidas em cada iteração são mostradas na tabela 1. A média dos testes foi de 95,75%, enquanto o desvio padrão percentual foi de 1,47%.

Iteração	**Acurácia**
1	95,59%
2	97,10%
3	94,03%
4	98,53%
5	94,20%
6	95,59%
7	94,12%
8	97,10%
9	95,65%
10	95,59%
Média	**95,75%**

Tabela 1: Acurácias obtidas no classificador SVM utilizando 3 variáveis de maior variância no PCA.

A acurácia média do problema mostrou-se excelente com apenas 3 das 9 variáveis utilizadas para o classificador, correspondendo a 80% da variância total do problema. É curioso notar que se após a técnica do PCA fosse analisado para o classificador apenas uma variável, a de maior autovalor, a acurácia média seria de $97,21\%$ e se fosse utilizado as 9 variáveis, a acurácia média seria de $95,31\%$, de acordo com os testes realizados. Com isso, nota-se que um número alto de variáveis para determinados problemas pode tornar um classificador mais impreciso, ao contrário do que pode se imaginar *a priori*, devido à inconsistência das variáveis adicionais, ao seus tipos (inteira, binária, continua) ou até devido às suas altas covariâncias em relação às outras variáveis.

Por fim, implementar uma técnica PCA mostrou-se ser de grande ajuda, ao definir um novo espaço amostral para o problema, com menos interdependência entre as variáveis e um custo computacional menor, uma vez que o classificador não é bombardeado com uma sequência de informações que são dispensáveis.