

1 Objetivo

Este trabalho tem como objetivo fazer o aluno ter alguma experiencia em explorar problemas de classificação e agrupamento.

O arquivo digitos.zip tem imagens de números. Cada imagem do conjunto possui 64 x 64 pixels no formato PGM onde cada pixel tem um valor 0 ou 1. Cada imagem tem um nome no formato X_yyy.BMP.inv.pgm onde X é o dígito representado na imagem.

O projeto tem que ser feito em R <http://cran.r-project.org/>. O R já tem implementado (talvez em pacotes externos que precisam ser instalados) todos os algoritmos que usaremos neste projeto.

2 Classificação

Usando apenas as imagens para os dígitos 1 e 7, e validação cruzada com 5-folds, teste os algoritmos k-NN e SVM. Lembre-se que com validação cruzada de 5-folds, voce fará o mesmo experimento 5 vezes, cada vez usando 4 folds para aprendizado e 1 para teste. Para cada experimento use um dos folds como conjunto de validação, e os outros 3 folds para aprendizado. Usar acurácia (taxa de acerto) como métrica para escolher os parâmetros.

2.1 K-vizinhos

Faça um classificador baseado em K vizinhos (knn em R). Teste para $k=1,3,5,11,21,31,51,71,e$ 101.

Reporte o valor de K para cada um dos experimentos, e a acurácia média em todos os experimentos.

2.2 SVM

Faça um classificador baseado em SVM com kernel rbf (ou gaussiano). Teste para $C=10e-3$ ate $10e3$, em múltiplos de 10. Use gama de $10e-3$ a $10e3$ em múltiplos de 10.

Reporte o valor de C e gama para cada experimento, e a acurácia média.

3 Agrupamento

Usando os centroides iniciais como um dado aleatório de cada dígito, rode um k-means ($K=10$). Para cada grupo, vamos dizer que o grupo “corresponde” ao dígito 1 se as imagens deste dígito são as mais comuns para o grupo. Se houver

empate, escolha qualquer um dos dígitos mais comuns como o sendo o correspondente do grupo. Usando esta ideia de que classe corresponde a cada grupo, qual é a taxa de erro ao final?

Bole um jeito de imprimir cada um dos 10 centroides finais.

Inicie o K-means com pontos aleatórios no espaço. Qual é a taxa de acerto no final?

4 Entrega

Enviar via email o relatório em PDF com no máximo 6 páginas. O relatório deve ser acompanhado de um arquivo R.

Na parte de classificação, para cada experimento do k-NN, reportar o k. Reportar a acurácia média. Para cada experimento do SVM reportar o C e o gama. Reportar a acurácia média.

Para o agrupamento, reportar a taxa de erro final, e imprimir os centroides para a inicialização como sendo um dado aleatório de cada dígito.

Para a inicialização aleatória reporta a taxa de erro final.