Laboratório Aprendizagem de Máquina

Lab1: Impactos da Representação

O script digits.py extrai a represenação mais simples possível de uma base de dados dígitos manuscritos. Para cada posição da imagem, verifica-se o valor de intensidade do pixel e se esse valor for > 128, a característica é igual a 1, caso contrário 0. A imagens tem tamanho variável e como os classificadores precisam de um vetor de tamanho fixo, as imagens são normalizadas utilizando as variáveis X e Y dentro da função rawpixel. Após a execução do programa, um arquivo chamado features.txt é criado no diretório corrente. Esse arquivo contem 2000 linhas no formato

0 0:0 1:0 2:1 3:1

O primeiro caractere indica o rótulo da classe. A sequencia i:v indica o índice da característica e o valor da mesma. Nesse caso, a características 0, 1, 2, e 3 tem valores 0, 0, 1 e 1, respectivamente.

Sua tarefa consiste em gerar diferentes vetores de características variando os valores de X e Y. Utilizando um kNN (k=3 e distância Euclidiana), encontre o conjunto de características que produziu os piores e melhores resultados de classificação. A base de dados deve ser dividida em 50% para treinamento e 50% para validaç!ao. Compare as matrizes de confusão nesses dois casos e reporte quais foram as confusões resolvidas pela melhor representação.

Para a sua melhor solução, verifique se é possível melhorar o resultados mudando os valores de k e métrica de distância.