Pedro Martins Moreira Neto

Relatório do Laboratório 1 de Machine Learning
Tarefa:

1. 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
2. Compare as matrizes de confusão nesses dois casos e reporte quais foram as confusões resolvidas pela melhor representação.
3. Verificar se é possível melhorar os resultados mudando os valores de k e métrica de distância.

Resultados:
Os vetores de características foram gerados com doze tamanhos diferentes com as imagens sendo binarizadas usando um *threshold* de 250 afim de reforçar os contornos e remover os ruídos. Junto a estes dados foram adicionadas informações de média, variância e desvio padrão dos pixels como característica da imagem.

Para cada tamanho foi aplicado o algoritmo KNN com k=3 e distância euclidiana. Os resultados são demonstrados na tabela 1 e na figura 1 abaixo:

Dimensões da Imagem	Precisão
5x5	0.812
10x10	0.893
15x15	0.913
20x20	0.908
25x25	**0.920**
30x30	0.912
35x35	0.914
40x40	0.910
45x45	0.913
50x50	0.909
55x55	0.909
60x60	0.907

Tabela 1 – Tamanho da imagem vs Precisão

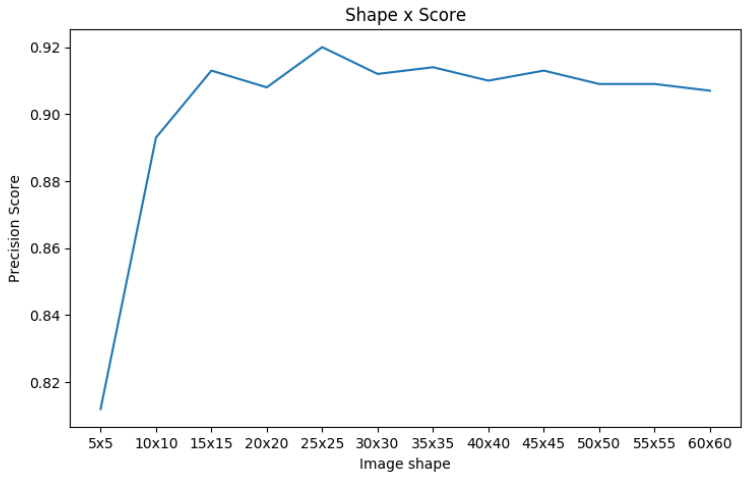

Figura 1 – Tamanho da imagem vs Precisão

Como podemos observar na tabela 1 e na figura 1, os valores que maximizaram a precisão foram X=25 e Y=25 que alcançou uma taxa de 92% de acerto, imagens com maiores dimensões não apresentaram ganho na taxa de precisão, estabilizando em na média de 91%. Logo, o conjunto mínimo de características que melhor descrevem os dados são os extraídos de imagens com dimensões de 25x25.

Em contrapartida os valores de X=5 e Y=5 foram os que tiveram a pior performance na tarefa de classificar os dígitos. Ao definir X e Y com valor 5 estamos reduzindo cada imagem cerca de 10x do seu tamanho original resultando em perda sumária de informação dos pixels o que reflete diretamente na precisão da classificação que foi de 81.2%.

Ao analisar as matrizes de confusão dos casos destacados acima, observa-se que no pior caso (X=5, Y=5) o algoritmo KNN apresenta maior taxa de erro ao classificar os dígitos 2, 7, 8 e 9 com precisão menor que 60%. Como discutido acima, este comportamento está atribuído ao fato de que reduzir a imagem a um fator de 10x seu tamanho original implica diretamente na perda de informação, principalmente de contorno, dos dígitos. Quando comparada com a matriz de confusão do caso ótimo, observa-se que os dígitos antes classificados incorretamente agora começam a ser corretamente classificados com precisão acima de 85%. Apesar da melhora nos dígitos 7, 8 e 9 ainda observamos uma fraqueza na classificação do dígito 2 e 4.

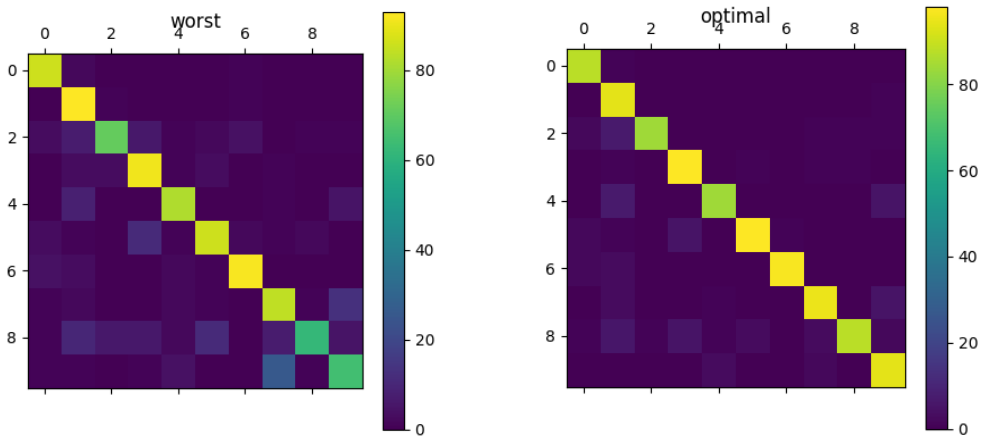

Figura 2 – Matrizes de confusão do pior caso e do caso ótimo.

Em posse da representação com maior taxa de acerto para esse problema, foi realizado testes modificando as métricas de distância e número de vizinhos do algoritmo KNN. Foram testadas as distâncias *euclidean* e *manhattan*, cada métrica foi avaliada calculando o número de vizinhos entre 3 e 10.

Com base nos resultados apresentados nos gráficos abaixo, conclui-se que ambas as métricas apresentam resultados bem similares, onde o número de vizinhos igual a 3 é o que maximiza o resultado.

Nos gráficos abaixo observa-se a presença de vales principalmente quando o número de vizinhos é par. Quando o número de vizinhos é par, existe uma maior probabilidade de que o ponto a ser classificado pelo algoritmo esteja equidistante a duas ou mais classes, uma das maneiras que o KNN usa para desempatar a classificação é escolher aleatoriamente a qual classe aquele ponto pertence aumentando a probabilidade de erro.

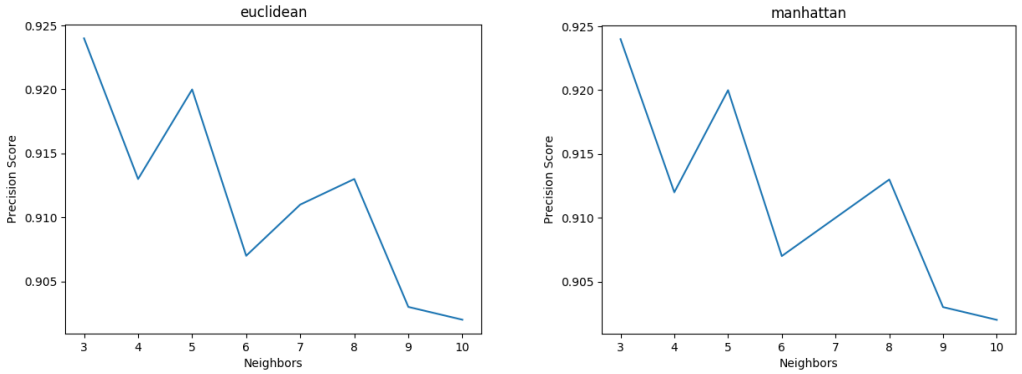

Figura 3 - Métrica vs Precisão vs Neighbors

Conclusão

Para a tarefa de classificação de dígitos, foi utilizado um algoritmo KNN com distância euclidiana e com k igual a 3. O vetor de característica foi obtido através do redimensionamento da imagem para X=25 e Y=25 e binarizando os pixels com um limiar de 250. O resultado máximo obtido foi o de 92% de precisão na utilizando os parâmetros já descritos. Não houve ganho de precisão ao modificar e/ou aumentar os parâmetros do KNN.

Mesmo com os bons resultados obtidos, fica claro que para essa tarefa é necessária uma representação dos dados e um algoritmo mais sofisticado para alcançar taxas de precisão maiores e com mais acurácia.