Aprendizado de máquina
INF01017 Arthur Zachow, 00277951
Prof. Mariana Recamonde Mendoza Data: 20/08/2021

Relatório de implementação do algoritmo KNN e o método holdout

Instruções de execução

As instruções para execução do código podem ser encontradas no pacote que contém este arquivo.

Resultados

Os experimentos foram realizados utilizando a linguagem Python3 com uso das bibliotecas padrão inclusas na linguagem. Para o desenvolvimento foi utilizada a plataforma Google Colab onde o código e as instruções de uso foram escritas no formato notebook.

O parâmetro para o método holdout, foi de 80% dos dados utilizados para o treinamento do modelo. Os valores de K utilizados foram os sugeridos incluindo 53 e 101 que são números primos e muito maiores que os valores sugeridos. O valor de 285 foi utilizado pois representa o resultado da "maioria"de todos os dados.

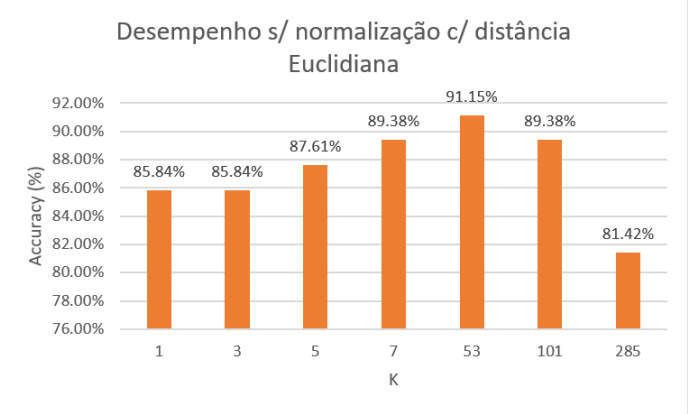

Figura 1: Desempenho, variando o hiper-parâmetro K e utilizando a métrica de distância euclidiana.

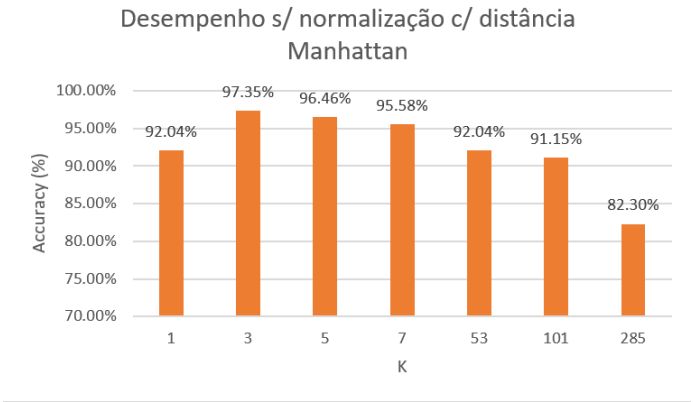

Figura 2: Desempenho, variando o hiper-parâmetro K e utilizando a métrica de distância de Manhattan.

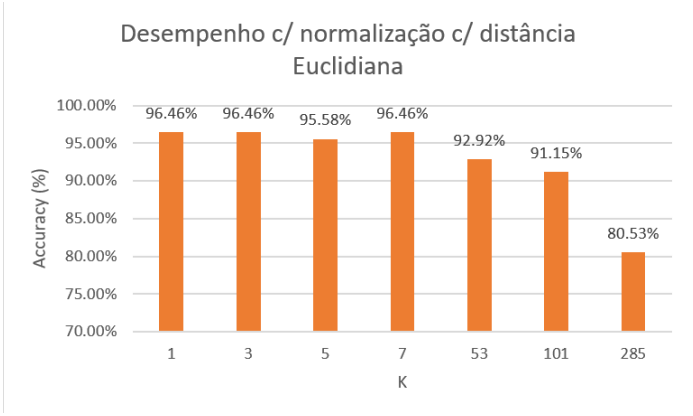

Figura 3: Desempenho, variando o hiper-parâmetro K e utilizando a métrica de distância euclidiana e normalização min-max.

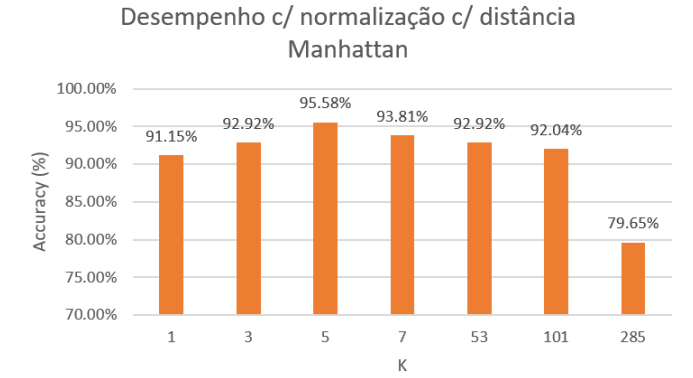

Figura 4: Desempenho, variando o hiper-parâmetro K e utilizando a métrica de distância de Manhattan e normalização min-max.

Conclusão

Baseado nos resultados apresentados nas figuras acima podemos tirar as seguintes conclusões.

Dados de entrada

Os dados de entrada possuíam escalas muito distintas entre os atributos, sendo uma diferença tão grande quanto 10^6, como na comparação entre os atributos $X5$ e $X2$.

Métrica de distância

Comparando as métricas de distância utilizadas é nítido que a distância euclidiana é mais sensível à diferença de escalas dos atributos. Neste quesito, a distância de Manhattan se mostra menos sensível à variação da escala.

Com a normalização é a distância euclidiana que possui os melhores resultados. Isto se deve ao fato de que a distância euclidiana é mais adequada para dados reais do que a distância de Manhattan. A distância de Manhattan não consegue representar tão precisamente quais pontos estão mais próximos dos outros nesta situação.

Hiper-parâmetro K

Em relação ao hiper-parâmetro K, na maior parte dos experimentos, os melhores resultados obtidos estiveram na faixa dos menores valores de K, normalmente entre os valores de 3 e 5. A única exceção à esta tendência foi na figura 1, onde o melhor resultado utilizava o valor de K igual à 53 e isto se deve à sensibilidade da métrica de distância utilizada.

A tendência é que o desempenho piore com o aumento do hiper-parâmetro K. Isto acontece porque com o aumento de K as fronteiras de decisão se tornam menos sensíveis a localidade espacial da entrada que é a base da técnica de KNN. Esta base é a ideia de que entradas parecidas (próximas no espaço de atributos) têm rótulos semelhantes.

Normalização

A normalização dos dados impactou os resultados. O impacto é dependente da métrica de distância utilizada conforme explicado na sub-seção que aborda os impactos das métricas de distância. Em geral a normalização melhorou os resultados para a métrica de distância euclidiana de forma significativa.