

0123456789

O dataset MNIST (*Modified National Institute of Standards and Technology*) é um conjunto de imagens contendo dígitos escritos à mão que é comumente utilizado para treinar vários algoritmos de aprendizado de máquina. O conjunto de imagens no banco de dados MNIST foi criado em 1994 como uma combinação de dois dos bancos de dados do NIST: O Banco de Dados Especial 1 e o Banco de Dados Especial 3 consistem em dígitos escritos por estudantes do ensino médio e funcionários do Bureau do Censo dos Estados Unidos, respectivamente. O MNIST foi criado, portanto, reagrupando as amostras dos bancos de dados originais do NIST. Os criadores consideraram que, uma vez que o conjunto de treinamento do NIST foi retirado de funcionários do Bureau do Censo dos Estados Unidos, enquanto o conjunto de teste foi retirado de estudantes do ensino médio americanos, não era adequado para experimentos de aprendizado de máquina. As imagens originais eram binárias (preto e branco) de 128x128 pixels, que foram processadas e suavizadas em imagens em tons de cinza de 28x28 pixels.

Assim, o MNIST contém 60.000 imagens de treinamento e 10.000 imagens de teste, cada uma com 28x28 pixels em tons de cinza. Os criadores originais do banco de dados mantêm uma lista de alguns dos métodos testados nele. Em seu artigo original, eles usam uma máquina de vetores de suporte (SVM) para obter uma taxa de erro de 0.8% [1]. Na Tabela 1 encontra-se o número de exemplos do MNIST.

Tabela 1: Número de exemplos de treinamento e teste por dígito do MNIST

	0	1	2	3	4	5	6	7	8	9	Total
Treinamento	5923	6742	5958	6131	5842	5421	5918	6265	5851	5949	60000
Teste	980	1135	1032	1010	982	892	958	1028	974	1009	10000

Neste trabalho, os conjuntos de treinamento e teste devem ser utilizados como fornecidos (exceto alterar a ordem entre os atributos, caso isso seja necessário) sem realizar qualquer tipo de pré-processamento nos atributos (pixels).

O algoritmo Naive Bayes deve ser utilizado para geração do classificador a partir dos dados de treinamento. Em seguida, o classificador deve ser avaliado no conjunto de teste.

a) Indique a matriz de confusão do classificador no conjunto de teste

		Classes Preditas										
		0	1	2	3	4	5	6	7	8	9	
Classes Verdadeiras	0											
	1											
	2											
	3											
	4											
	5											
	6											
	7											
	8											
	9											

b) Utilize a métrica taxa de erro (por classe e geral) para indicar o desempenho do classificador no conjunto de teste:

Taxa de Erro	0	1	2	3	4	5	6	7	8	9	Total
Teste											

- c) O que você pode concluir com os resultados obtidos? Justifique sua resposta.
- d) Relate como você fez para obter os resultados acima (quais bibliotecas ou softwares foram utilizados; caso tenha desenvolvido algum código, anexe todo o código desenvolvido, juntamente com as respostas acima, em um arquivo compactado do tipo ZIP.

[1] LeCun, Yann; Léon Bottou; Yoshua Bengio; Patrick Haffner (1998). "Gradient-Based Learning Applied to Document Recognition". Proceedings of the IEEE. 86 (11): 2278–2324. doi:10.1109/5.726791. S2CID 14542261.

Submissão

- Um único arquivo compactado no formato ZIP deve ser submetido na atividade correspondente no Tidia-AE, contendo as respostas às questões de a) até d) e código-fonte desenvolvido, se for o caso.
- O trabalho pode ser desenvolvido por equipes de até quatro alunos. Neste caso, uma única submissão deve ser efetuada no Tidia-AE contendo o nome de todos os integrantes, em ordem alfabética.

Critério de Correção

- 50% para respostas corretas em a) e b) e justificativas adequadas em c) e d)
- 50% estão reservados para a extensão, inovação, visualização dos resultados. Por exemplo:
 - O classificador NB foi alterado, desde que mantidas suas características originais? Por exemplo, uma distribuição diferente da normal foi utilizada para atributos numéricos? Qual? Justifique.
 - Visualização de componentes do classificador, por exemplo, as médias (e variâncias) por classe (e geral) podem ser representadas como uma imagem 28x28

Fonte dos Dados:

Os dados para este trabalho estão disponíveis no arquivo compactado (em formatos ARFF e CSV) junto ao enunciado e foram obtidos a partir da URL: https://git-disl.github.io/GTDLBench/datasets/mnist_datasets/

O formato CSV disponibilizado é o seguinte:

label, pixel-1, pixel-2, pixel-3, ..., pixel784

em que "label" representa um dígito entre 0 e 9 e pixel-* representa um pixel em tons de cinza, de 0 a 255, por exemplo: