Classificação Fashion-MNIST

Farzin Shams
farzinshams95@gmail.com

I. Introdução

Neste trabalho, temos como objetivo comparar diferentes modelos de classificação, tanto lineares quanto não-lineares, para classificar as 10 diferentes classes do dataset *Fashion-MNIST*. As imagens escalas de cinza de tamanho 28x28 (= 784 entradas), com 60000 dados de treinamento e 10000 de teste.

Figura 1. Algumas imagens do dataset

II. Modelos

Os modelos testados foram: regressão logística *one-vs-all* (LR), regressão logística multinomial (MLR), redes neurais com uma (NN 1) e duas (NN 2) camadas, com diferentes funções de ativação.

LR é um modelo linear de classificação binária em que é preciso gerar n modelos diferentes, um para cada classe, onde n é o número de classes. A classe final predita, para um dado qualquer, é igual àquela prevista pelo modelo com a maior convicção (probabilidade).

MLR é um modelo linear de classificação multinomial. Ele possui n saídas que passam pela transformação *softmax*, que converte as predições de cada camada de saída ($scores$) em probabilidades. A classe final predita, então, é igual à classe correspondente à saída com maior probabilidade.

Por fim, temos as redes neurais de uma e duas camadas, NN 1 e NN 2, que são semelhentes à MLR. A principal diferença entre elas é a presença das funções de ativações, que são responsáveis pela transformação não-linear de combinações lineares dos pesos e entradas.

III. Preparação dos Dados e Treinamento dos Modelos

Como as imagens estão na escala de cinza, os valores dos pixels variam entre 0 e 255. A fim de melhor assegurar a convergência da otimização dos modelos, dividimos todos os pixels por 255 para que eles variassem de 0 a 1. Não há dados faltantes no dataset.

Modelo	Acurácia no Conjunto de Teste	Tempo (segundos)	Neurônios (em ordem)
LR	0.7936	510.01	-
MLR	0.7963	902.96	-
NN 1 (Relu)	0.8064	1336.03	30
NN 2 (Sigmoid)	0.7355	1456.91	30, 30
NN 2 (Relu)	0.7385	1405.75	30, 30
NN 2 (Tanh)	0.7669	1488.56	30, 30

Tabela 1: Resumo do desempenho dos modelos

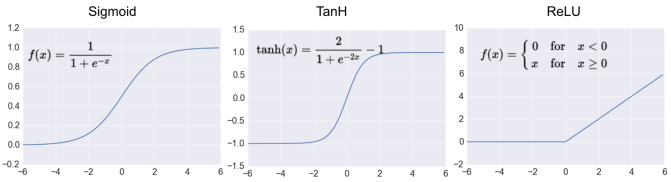

Figura 2. Funções de ativações usadas

Um resumo dos desempenhos dos modelos é apresentado na tabela 1. Todos os modelos foram treinados durante 1000 épocas, que era o suficiente para que o erro de validação estabilizasse (variação do *log loss* < 0.0001). O número de neurônios usados nas camadas das redes neurais foram mantidas constantes para que pudessemos avaliar a mudança das funções de ativação no treinamento. Os *learning rates* usados variaram entre 0.1 e 0.3, que era pequeno o suficiente para que os modelos não divergissem, mas grande o suficiente para que convergissem em menos de 1000 épocas.

Em todos os nossos experimentos, percebemos que o erro de validação dificilmente aumentava, o que indica que os modelos usados não estavam sobretreinando. Devido a isso, decidimos usar poucos dados de validação (5% do conjunto de treino) nos modelos finais para que tivessemos mais dados de treinamento e, assim, melhorar o resultado no conjunto de teste.

O LR e MLR tiveram desempenho semelhante, sugerindo

que não há muita diferença entre as abordagens one-vs-all e multinomial. A NN 1 com função de ativação *Relu* teve o melhor desempenho. As NN 2, por outro lado, tiveram o pior desempenho, sugerindo que a adição de uma camada adicional após a primeira camada intermediária reduz o poder de predição destas redes neste dataset.

A mudança de função de ativação nas NN 2 não foi o suficiente pra compensar a diferença de acurácia destes modelos com os demais, apesar de a NN 2 com $Tanh$ aumentar a acurácia em 3% em comparação aos NN 2 com funções de ativação $Relu$ e $Sigmoid$.

Uma possível explicação para os resultados obtidos é que as NN, por terem a capacidade de modelar as relações não-lineares, têm maior poder de predição que os modelos lineares. Porém, elas possuem um número elevado de parâmetros a serem otimizados, o que demanda um maior número de dados e tempo de treinamento.

IV. Matriz de Confusão

Usando as predições do melhor modelo, a NN 1 com $Relu$, plotamos a matriz de confusão na figura 3, onde cada classe possui 1000 exemplares. Como pode ser visto, a classe 6, que é a camisa, é a mais difícil de ser identificada pelo modelo, acertando apenas 30.8% dos exemplares. Ao invés dela, o modelo dá maior prioridade às classes 0, 2 e 4, que são a camiseta, blusa e jaqueta, respectivamente.

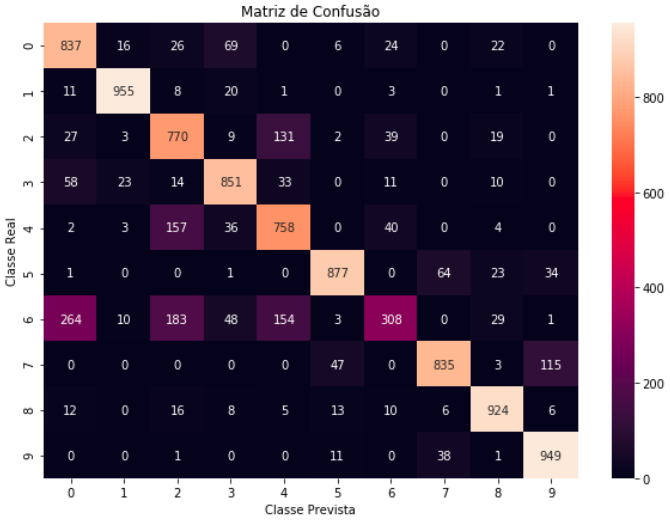

Figura 3. Matriz de Confusão