Relatório Técnico: Classificação com Redes Neurais MLP

Dataset: Fashion-MNIST | **Metodologia:** Back-propagation e Otimização de Hiperparâmetros

1. Introdução e Contextualização do Problema

O objetivo deste estudo foi desenvolver e otimizar um modelo de rede neural do tipo *Multilayer Perceptron* (MLP) capaz de classificar corretamente imagens do dataset *Fashion-MNIST*. Este conjunto de dados apresenta um desafio de visão computacional mais complexo que o clássico MNIST de dígitos, contendo 70.000 imagens em tons de cinza (28x28 pixels) divididas em 10 categorias de vestuário, como camisetas, calças e botas.

A abordagem adotada seguiu um fluxo incremental de experimentação, partindo de uma arquitetura base para investigar a estabilidade, a influência de hiperparâmetros (taxa de aprendizado e momentum), a topologia da rede e, finalmente, a robustez estatística do modelo final através de validação cruzada.

2. Definição da Arquitetura e Dinâmica de Treinamento

2.1. Exploração Inicial e Escolha da Função de Ativação

Iniciamos os experimentos com uma topologia base (entrada de 784 neurônios, camadas ocultas de 64 e 32 neurônios, e saída de 10 neurônios). O foco inicial foi avaliar a estabilidade da inicialização de pesos e a escolha da função de ativação ideal para este cenário específico.

Ao contrário da literatura padrão que frequentemente favorece a ReLU para redes profundas, nossos experimentos empíricos demonstraram que a função **Sigmoide** ofereceu um comportamento mais estável e uma acurácia levemente superior (média de 95,21% no treino preliminar) em comparação à ReLU (94,73%) e à Tanh (94,99%). A ReLU, neste cenário de rede rasa e dados normalizados, apresentou sinais de estagnação em alguns neurônios. Portanto, optou-se pela Sigmoide para as camadas ocultas e Softmax para a camada de saída, garantindo uma distribuição de probabilidade clara para a classificação multiclasse.

A função de perda utilizada foi a *Entropia Cruzada Categórica Esparsa*, ideal para este tipo de classificação, otimizada pelo algoritmo Adam.

2.2. Sintonia Fina de Hiperparâmetros

Com a função de ativação definida, investigamos o impacto da Taxa de Aprendizado (*Learning*

Rate) e do termo *Momentum* através de uma busca em grade (*Grid Search*).

Os resultados foram reveladores quanto à sensibilidade da rede:

1. **Instabilidade com Altas Taxas:** Taxas de aprendizado elevadas (0.1) levaram à divergência completa do modelo, resultando em uma "adivinhação aleatória" (acurácia de ~10%), independentemente do momentum.
2. **O "Ponto Ótimo":** A combinação mais eficiente foi uma **Taxa de Aprendizado de 0.001** com um **Momentum moderado de 0.5**. Esta configuração, aliada a um *batch size* de 64 e treinamento por 40 épocas, permitiu uma convergência suave e consistente, atingindo uma acurácia de validação próxima de 89.5% com baixo desvio padrão.

3. Investigação da Topologia: "Menos é Mais"

Uma das descobertas mais interessantes deste estudo ocorreu durante a exploração do número de camadas e neurônios. Testamos 20 configurações diferentes, variando de redes rasas (1 camada oculta) a redes profundas (4 camadas ocultas), e de 32 a 1024 neurônios.

Observou-se um fenômeno claro de **retornos decrescentes**. O aumento da complexidade não se traduziu linearmente em desempenho. Pelo contrário, as arquiteturas mais profundas (3 ou 4 camadas) sofreram com o problema do *Vanishing Gradient* (gradiente desvanecente), exacerbado pelo uso da função Sigmoide, resultando em aprendizado mais lento e menor acurácia final.

As arquiteturas que melhor equilibraram desempenho e custo computacional foram as "piramidais decrescentes" e rasas. Quatro finalistas foram selecionadas para a fase de teste:

1. **Topologia A:** 2 Camadas: 256, 128
2. **Topologia B:** 3 Camadas: 512, 256, 128
3. **Topologia C:** 1 Camada: 256
4. **Topologia D:** 4 Camadas: 1024, 512, 256, 128

4. Validação e Resultados Finais

4.1. Influência dos Dados e Teste Cego

Antes da seleção final, analisamos o impacto do volume de dados. Os testes confirmaram que o uso de 100% do dataset de treino (estratificado) foi mandatório. Embora isso aumente linearmente o tempo de processamento, observou-se que o ganho em estabilidade e a redução da variância compensam o custo computacional adicional.

Ao submeter as quatro topologias finalistas ao **Conjunto de Teste** (dados nunca vistos pelo modelo) utilizando todo o dataset disponível, a **Topologia C (1 Camada Oculta com 256**

neurônios) surpreendeu ao superar as arquiteturas mais complexas.

Modelo	Configuração	Acurácia (Teste)	F1-Score	Tempo de Treino
Modelo C	**1 Camada (256 neurônios)**	**87,77%**	**0,8786**	**26,1s**
Modelo A	2 Camadas (256, 128)	87,53%	0,8756	27,9s
Modelo B	3 Camadas (512, 256, 128)	87,24%	0,8722	48,1s
Modelo D	4 Camadas (1024...128)	87,21%	0,8736	126,7s

O Modelo C venceu não apenas em acurácia, mas também em eficiência, treinando em metade do tempo do Modelo B e um quinto do tempo do Modelo D. Isso valida a tese de que, para este problema específico e usando ativação Sigmoide, uma rede mais simples facilita a propagação do erro e o ajuste dos pesos.

4.2. Robustez Estatística (Validação Cruzada)

Para garantir que o desempenho do Modelo C não foi fruto de uma divisão de dados favorável ("sorte"), submetemos esta configuração vencedora a uma Validação Cruzada *k-fold* (com k=5).

Os resultados confirmaram a solidez da solução:

- **Média de Acurácia:** 88,83%
- **Desvio Padrão:** 0,43%
- **Consistência:** Todas as 5 partições apresentaram acurácia entre 88,3% e 89,3%.

O baixíssimo desvio padrão (< 0,5%) indica que o modelo é extremamente estável e generaliza bem para diferentes subconjuntos de dados, sem apresentar sinais significativos de *overfitting* ou dependência de amostras específicas.

5. Conclusão

O desenvolvimento deste classificador para o Fashion-MNIST ilustrou na prática conceitos fundamentais de Redes Neurais. A principal lição aprendida foi que a complexidade arquitetural

nem sempre é sinônimo de melhor desempenho.

A configuração final escolhida destaca-se pela eficiência:

- **Topologia:** 1 Camada Oculta (256 neurônios).
- **Ativação:** Sigmoide (Oculta) e Softmax (Saída).
- **Otimização:** Adam (LR=0.001, Beta1=0.5).
- **Desempenho Final:** ~88.8% de acurácia com alta estabilidade.

Este modelo oferece o melhor equilíbrio entre precisão preditiva, tempo de treinamento e simplicidade de implementação, cumprindo com êxito os requisitos do estudo dirigido.