Exercício de Fixação de Conceitos
EFC1 - Questão 1

Nome: Guilherme Rosa RA: 157955

1. Separação das amostras de treinamento nos conjuntos de treinamento e validação

As 60.000 amostras de treinamento foram permutadas pseudo-aleatoriamente e divididas em dois conjuntos para utilizar a técnica de validação *holdout*, cuja finalidade é aumentar a capacidade de generalização do classificador linear desejado. A divisão das amostras foi feita da seguinte forma:

- 80% das amostras foram agrupadas no conjunto de treinamento (48.000)
- 20% das amostras foram agrupadas no conjunto de validação (12.000)

Em seguida foi verificado a proporção de amostras por classe em ambos os conjuntos, de modo a verificar o balanceamento e representatividade das classes. As porcentagens de cada classe variaram entre 8.99% a 11.29% no conjunto de treinamento e 9.2% a 11.02% no conjunto de validação. Nota-se que todas as classes estão representadas de forma equilibrada, o que leva a classificadores menos enviesados.

2. Busca inicial pelos melhores coeficientes de regularização

A busca inicial pelo melhor coeficiente de regularização λ foi feita utilizando dois critérios de desempenho: erro quadrático médio e a taxa de erro de classificação. De fato, os melhores coeficientes são aqueles associados aos menores erros. Essa busca foi feita considerando os seguintes valores para λ:

$$\lambda = \{2^{-10}, 2^{-8}, \dots, 2^{+8}, 2^{+10}, \dots, 2^{18}, 2^{20}\}$$

O conjunto inicial fornecido pelo professor não foi suficiente, sendo necessário estender os valores de λ na busca.

As figuras 1 e 2 apresentam os gráficos semilog das métricas de desempenho em função do coeficiente de regularização. Da Figura 1, pode-se perceber que o mínimo do erro quadrático médio ocorre em algum λ com valor dentro do intervalo [10, 110], enquanto que na Figura 2, o mínimo da taxa de erro de classificação ocorre em algum λ com valor dentro do intervalo [100, 10000]. Os melhores coeficientes de regularização da busca inicial estão apresentados na primeira linha da Tabela 1.

Tabela 1: Valores dos melhores coeficientes de regularização, antes e após o refinamento, considerando as métricas erro quadrático médio e taxa de erro de classificação

	Melhor λ para o erro quadrático médio	Melhor λ para a taxa de erro de classificação
Busca inicial	64	1024
Busca refinada	51,6181	965,8832

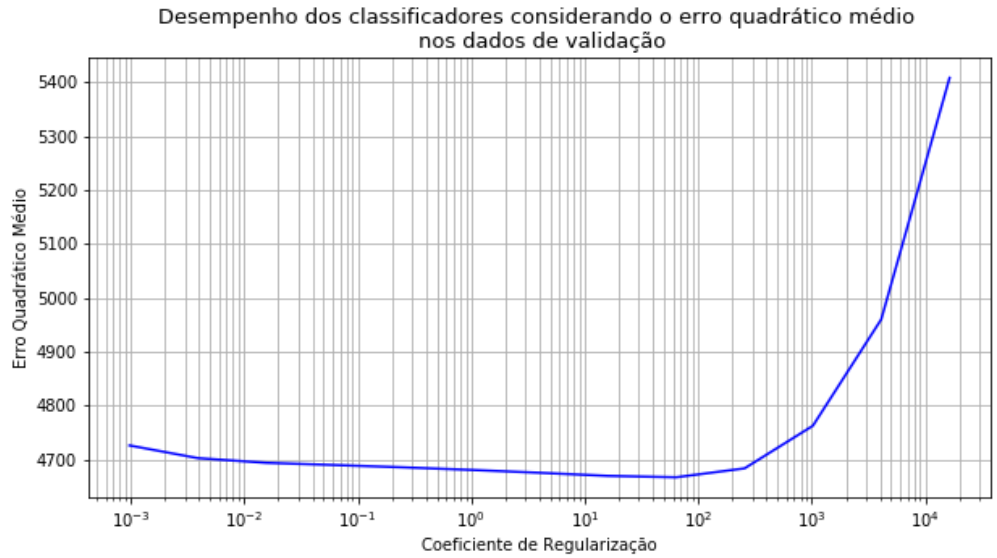

Figura 1: Gráfico semilog do erro quadrático médio em função do λ na busca inicial.

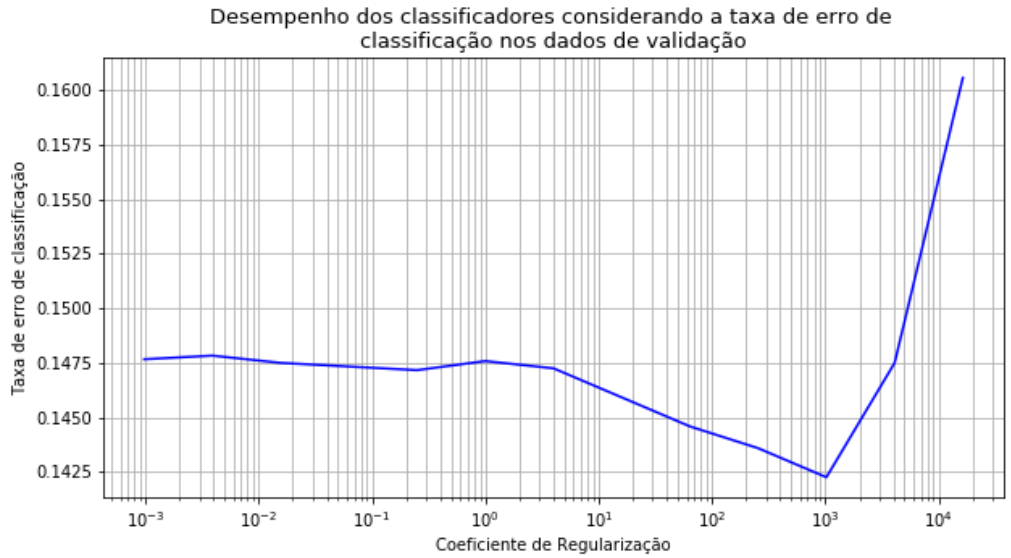

Figura 2: Gráfico semilog da taxa de erro de classificação em função do λ na busca inicial.

3. Busca refinada pelo melhor λ para o erro quadrático médio:

A busca refinada pelo coeficiente λ foi feita considerando 200 valores linearmente espaçados no intervalo [10, 110]. A Figura 3 apresenta o gráfico semilog do erro quadrático médio em função do coeficiente. O valor do melhor coeficiente obtido está apresentado na segunda linha da Tabela 1.

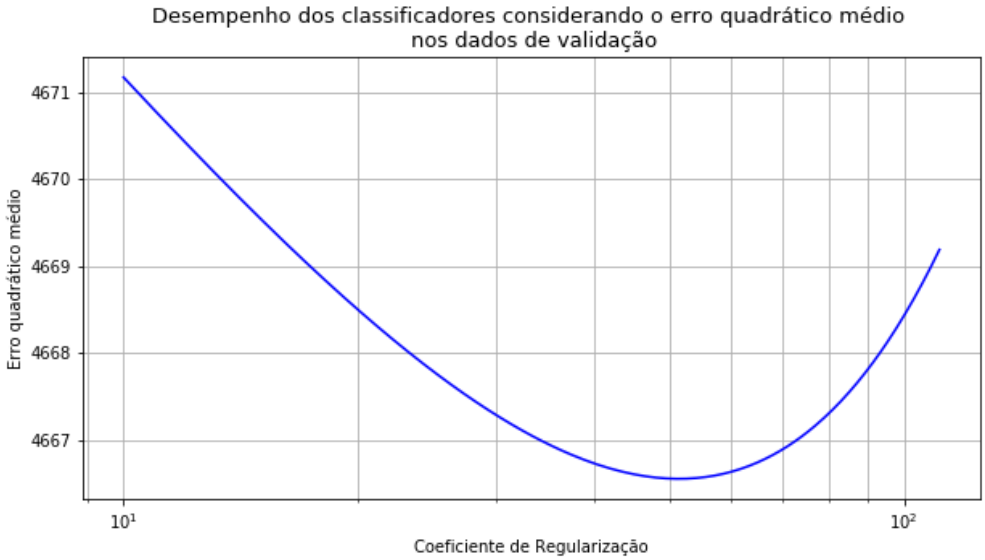

Figura 3: Gráfico semilog do erro quadrático médio em função do λ na busca refinada.

4. Busca refinada pelo melhor λ para a taxa de erro de classificação:

A busca refinada pelo coeficiente λ foi feita considerando 200 valores espaçados logaritmicamente no intervalo [100, 10000]. A Figura 4 apresenta o gráfico semilog do erro quadrático médio em função do coeficiente. O valor do melhor coeficiente obtido está apresentado na segunda linha da Tabela 1.

Figura 4: Gráfico semilog do erro quadrático médio em função do λ na busca refinada.

5. Classificador linear final

Para a implementação do classificador linear final utilizamos o coeficiente de regularização da busca refinada associado a menor taxa de erro de classificação. Em seguida, treinamos novamente o modelo com todas as 60.000 amostras de treinamento e aplicamos o modelo final obtido aos dados do conjunto de teste, constituído por 10.000 amostras.

Tabela 2: Métricas de desempenho do classificador linear final com λ = 965,8832

Parâmetro/Métrica	Valor
Erro quadrático médio	3873,1678
Taxa de erro de classificação	13,46%
Taxa de acerto (acurácia global)	86,54%

A Tabela 2 apresenta algumas informações do classificador linear obtido. Nota-se que a taxa de acerto ou acurácia global do classificador é de 86,54%, sendo este valor calculado pela razão entre o número de amostras classificadas corretamente e o número total de amostras.

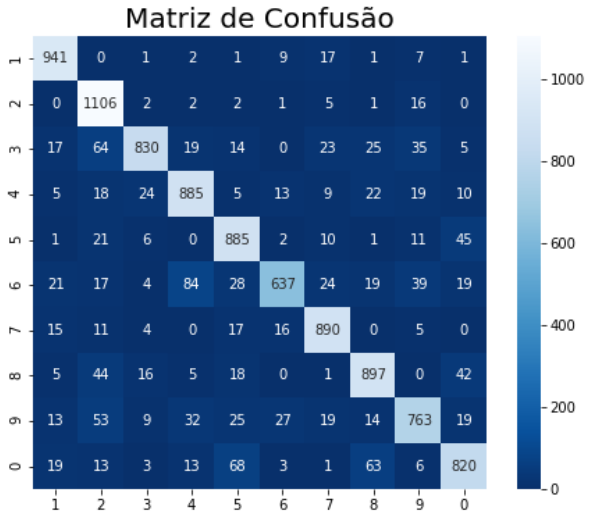

Figura 5: Matriz de confusão do classificador linear.

A matriz de confusão do classificador linear é apresentada na Figura 5. Os índices da esquerda indicam as classes verdadeiras da amostra, enquanto os índices inferiores indicam as classes estimadas pelo classificador. Por exemplo, considerando as imagens pertencentes a classe 4, pode-se observar que:

- 5 imagens foram classificadas incorretamente como pertences a classe 1.
- 18 imagens foram classificadas incorretamente como pertencentes a classe 2.
- 24 imagens foram classificadas incorretamente como pertencentes a classe 3
- 885 imagens foram classificadas corretamente como pertencentes a classe 4 e assim sucessivamente.

- A mesma análise pode ser feita para as demais classes.

A Figura 6 apresenta os mapas de calor de cada um dos classificadores lineares (um para cada classe) produzidos. Esses mapas foram gerados tomando cada uma das colunas da matriz de parâmetros W do modelo (eliminando o primeiro elemento, pois este é referente ao termo de polarização).

Podemos observar que a ativação dos parâmetros de cada um dos classificadores apresenta uma distribuição ou comportamento semelhante ao formato dos dígitos a serem classificados. Isso é bem evidente nos mapas de calor referentes as classes 1, 2, 3, 6, 8 e 0.

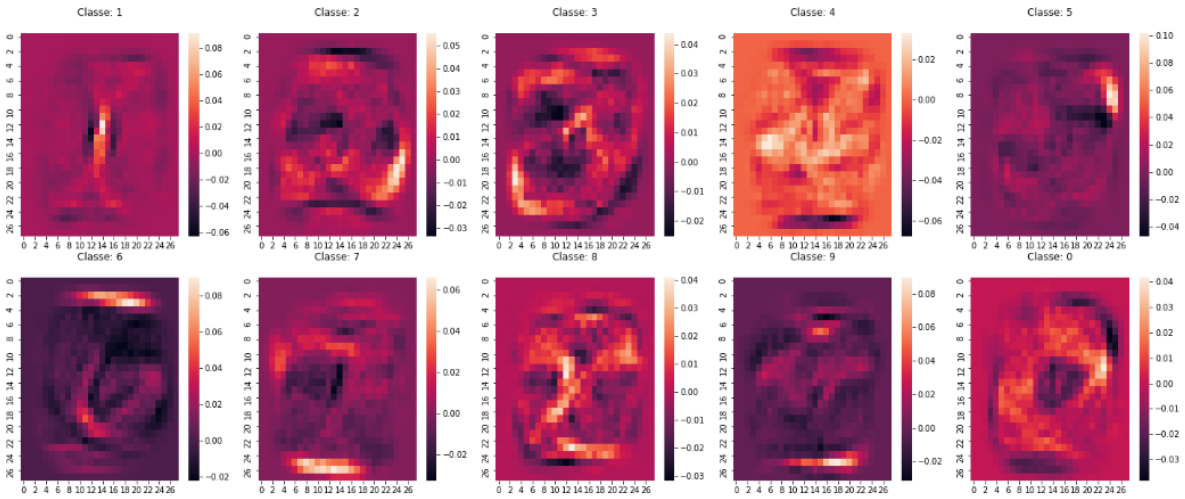

Figura 6: Mapas de calor dos parâmetros dos classificadores lineares de cada uma das classes.

Por fim, a Figura 7 mostra alguns exemplos de imagens de dígitos classificados incorretamente.

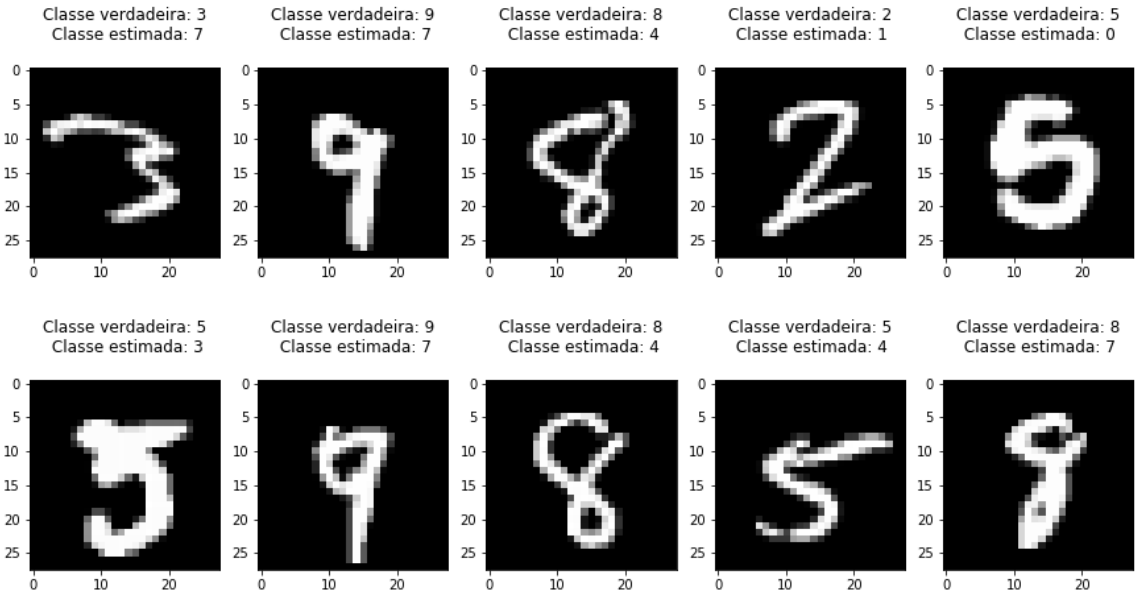

Figura 7: Exemplos de imagens classificadas incorretamente.