1. Introdução

Neste trabalho construímos um classificador usando regressão logística multinomial com penalização ridge e lasso baseado em *Friedman et al. (2010)*. Para melhor demonstrar as propriedades de ambas técnicas, fizemos uma simulação que foi capaz de mostrar a diferença entre elas. No exemplo prático, utilizamos os dados de dígitos do MNIST, separados em conjunto de treino (70%) e teste (30%). Os resultados de acurácia nos dados de teste foram 90,87% para ridge e 90,27% para lasso.

1.1 Banco de Dados

Os dados de dígitos MNIST (Modified National Institute of Standards and Technology) é considerado o "Hello World" no campo da Visão Computacional e do Machine Learning. Os dados foram criados em 1998, como uma combinação de duas bases de dados e desde então tem sido usado como um benchmark de performance de algoritmos, variando desde Classificação Linear (Taxa de Erro de 7.6%) até Rede Neural de Convolução (Taxa de Erro de 0.17%).

2. Metodologia

O modelo de regressão logística multinomial pertence à classe dos modelos lineares generalizados. Nesse modelo, a variável resposta Y é categórica com mais de duas categorias. Suponha que a variável resposta tenha J categorias e temos um vetor x de covariáveis. Construímos o modelo de regressão multinomial através dos $J-1$ logitos

$$\log \frac{\Pr(Y=j \mid x)}{\Pr(Y=J \mid x)} = \beta_{0j} + x^\top \beta_j,\ j = 1, \ldots, J-1, \tag{1}$$

onde β_ℓ é vetor de coeficientes de dimensão p. No caso de modelos penalizados, escolhemos outra parametrização:

$$\Pr(Y=j \mid x) = \frac{e^{\beta_{0j} + x^\top \beta_j}}{\sum_{j=1}^{J} e^{\beta_{0j} + x^\top \beta_j}}. \tag{2}$$

Essa parametrização não é estimável sem restrições, porque para qualquer valor dos parâmetros $\{\beta_{0\ell}, \beta_\ell\}_1^J$, $\{\beta_{0\ell} - c_0, \beta_\ell - c\}_1^J$ fornecem as mesmas probabilidades. A regularização lida com essa ambiguidade de maneira natural.

Agora, precisamos construir a função onde estimamos os parâmetros do modelo. Primeiro construímos a função de verossimilhança para a distribuição multinomial. Então, considere que temos N subpopulações em nossos dados, onde cada subpopulação é representada por um vetor de covariáveis x_i, $i = 1, \ldots, N$, tamanho n_i e y_{ij} observações na categoria $j = 1, \ldots, J$. Considere ainda que $\pi_{ij} = \Pr(Y = j \mid x_i)$. Logo, a função densidade de probabilidade conjunta é:

$$f(y \mid \beta) = \prod_{i=1}^{N} \left[\frac{n_i!}{\prod_{j=1}^{J} y_{ij}!} \cdot \prod_{j=1}^{J} \pi_{ij}^{y_{ij}} \right], \tag{3}$$

onde $\boldsymbol{\beta}$ é o vetor de parâmetros do modelo de regressão. Como queremos maximizar a Equação (3) com respeito à $\boldsymbol{\beta}$, os termos fatoriais que não têm os termos π_{ij} podem ser tratados como constantes. Então, podemos escrever a função de verossimilhança para o modelo de regressão logística multinomial como:

$$L(\beta \mid y) \propto \prod_{i=1}^{N} \prod_{j=1}^{J} \pi_{ij}^{y_{ij}}. \tag{4}$$

Substituindo a Equação (2) na Equação (4), temos que

$$L(\beta \mid y) \propto \prod_{i=1}^{N} \prod_{j=1}^{J-1} \left(\frac{e^{\beta_{0j} + x^\top \beta_j}}{\sum_{k=1}^{J} e^{\beta_{0k} + x^\top \beta_k}} \right)^{y_{ij}}. \tag{5}$$

Com isso, a função de logverossimilhança para o modelo de regressão logística multinomial é:

$$l(\beta) = \sum_{i=1}^{N} \sum_{j=1}^{J-1} \left[y_{ij}(\beta_{0j} + x^\top \beta_j) - y_{ij} \log \left(\sum_{k=1}^{J} e^{\beta_{0k} + x^\top \beta_k} \right) \right]. \tag{6}$$

Para o modelo de regressão logística multinomial regularizada, estimamos os parâmetros do modelo através da função de logverossimilhança penalizada, que é dada por:

$$l(\beta, \lambda, \alpha) = l(\beta) - \lambda P_\alpha(\beta), \tag{7}$$

onde

$$\begin{aligned} P_\alpha(\beta) &= (1 - \alpha)\|\beta\|_{\ell_2}^2 + \alpha\|\beta\|_{\ell_1} \\ &= \sum_{j=1}^{p} \left[(1 - \alpha)\beta_j^2 + \alpha \left|\beta_j\right| \right]. \end{aligned} \tag{8}$$

$P_\alpha(\beta)$ é a penalização *elastic-net*. Nesse trabalho consideramos a penalização ridge ($\alpha = 0$) e a penalização lasso ($\alpha = 1$).

Portanto, para cada valor de λ, encontramos os estimadores de β ao maximizar a logverossimilhança penalizada. Vale ressaltar que os valores de β são encontrados através de métodos numéricos, pois não há solução analítica para encontrar os estimadores de β. Além disso, as variáveis preditoras são padronizadas antes do processo de otimização. Para a implementação do algoritmo usamos o pacote **`glmnet`** do software estatístico **`R`**.

2.1 Selecionando o melhor valor de λ

Os valores das estimativas de β são obtidos de acordo com o valor de λ. Logo, para escolher um λ ótimo, podemos usar um erro de predição para nos guiar. Nesse trabalho, separamos os dados em treinamento e teste, sendo 70% dos dados para treino e 30% para teste. Nos dados de treino, aplicamos a regressão ridge e lasso, e escolhemos o valor de λ através da validação cruzada. O conjunto de teste é usado para avaliar os modelos de regressão ridge e lasso obtidos nos dados de treinamento. Além disso, como as classes do nosso banco de dados são balanceadas, usamos a acurácia como erro de predição.

3. Simulação

De acordo com *Friedman et al. (2010)*, as regressões ridge e lasso são úteis quando temos mais variáveis preditoras do que observações, ou qualquer situação onde há muitas variáveis preditoras correlacionadas. *Friedman et al. (2010)* ainda argumenta que a regressão ridge é conhecida por encolher os coeficientes de preditores correlacionados, um em direção à outro, como se tivessem o mesmo peso no modelo. Por sua vez, a regressão lasso é, de certa maneira, indiferente à preditores altamente correlacionados, e tende a escolher um preditor e ignorar o restante.

Com isso, vamos construir a regressão ridge e lasso para dois casos: com preditores correlacionados e preditores maior que observação. A variável resposta é multinomial com 3 classes, cada classe rotulada como 1, 2 e 3, respectivamente.

3.1 Preditoras correlacionadas

Por simplicidade, consideramos apenas dez preditoras, geradas a partir de uma distribuição normal, com 1000 repetições, e introduzimos correlação entre algumas dessas variáveis. Para o i-ésimo vetor de

preditoras x_i, seja $\pi_{ij} = P(Y = j \mid x_i)$, $j = 1, 2, 3$, onde Y é a variável resposta. Então, consideramos o modelo:

$$\begin{aligned} \log\left(\frac{\pi_{i1}}{\pi_{i3}}\right) &= 1 + x_i^T\beta_1 = 1 + 2x_{i1} + 3x_{i2} - x_{i3} + 4x_{i4} + 10x_{i5} - 2x_{i6} - 4x_{i7} + 7x_{i8} - 2x_{i9} + x_{i10}, \\ \log\left(\frac{\pi_{i2}}{\pi_{i3}}\right) &= 3 + x_i^T\beta_2 = 3 + x_{i1} - x_{i2} - 5x_{i3} + 2x_{i4} + 15x_{i5} + 3x_{i6} - 7x_{i7} + 4x_{i8} + 6x_{i9} + 9x_{i10}, \end{aligned} \tag{9}$$

onde

$$\begin{aligned} \pi_{i1} &= \frac{e^{1+x_i^T\beta_1}}{1 + e^{(1+x_i^T\beta_1)+(3+x_i^T\beta_2)}} \\ \pi_{i2} &= \frac{e^{3+x_i^T\beta_2}}{1 + e^{(1+x_i^T\beta_1)+(3+x_i^T\beta_2)}} \\ \pi_{i3} &= \frac{1}{1 + e^{(1+x_i^T\beta_1)+(3+x_i^T\beta_2)}}. \end{aligned} \tag{10}$$

A i-ésima observação da variável resposta foi gerada a partir de uma distribuição multinomial com as proporções obtidas na Equação (10). Esse processo resultou em 340 variáveis com valor 1, 452 variáveis com valor 2 e 208 variáveis com valor 3. A Figura 1 mostra a estrutura de correlação entre as 10 variáveis, onde algumas estão altamente correlacionadas entre si.

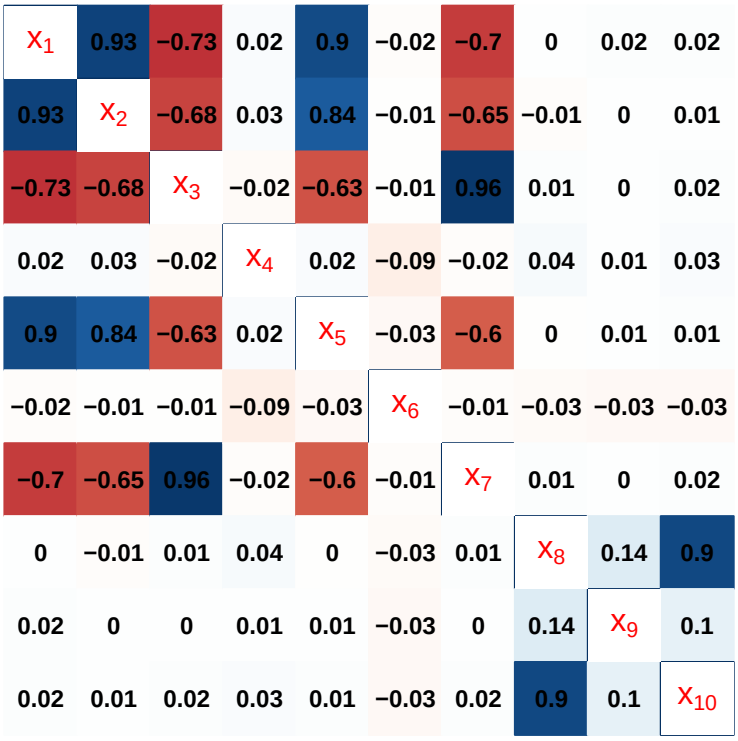

Figura 1: Correlação entre as variáveis preditoras.

O modelo obtido por regressão ridge fornece o seguinte resultado:

$$\begin{aligned}
\log\left(\frac{\pi_{i1}}{\pi_{i3}}\right) &= 0,826 + 0,431z_{i1} + 0,422z_{i2} - 0,315z_{i3} + 0,341z_{i4} + 0,863z_{i5} \\
&\quad - 0,227z_{i6} - 0,417z_{i7} + 0,528z_{i8} - 1,176z_{i9} + 0,260z_{i10}, \\
\log\left(\frac{\pi_{i2}}{\pi_{i3}}\right) &= 1,301 + 0,392z_{i1} + 0,256z_{i2} - 0,431z_{i3} + 0,121z_{i4} + 0,822z_{i5} \\
&\quad + 0,06z_{i6} - 0,482z_{i7} + 0,434z_{i8} + 1,845z_{i9} + 0,422z_{i10}.
\end{aligned} \tag{11}$$

onde z_{ij} é a variável padronizada. O λ escolhido pela validação cruzada é 0,0404. No geral, não vemos muitos coeficientes encolhidos para próximo de zero e o peso das variáveis correlacionadas não estão próximos. Isso pode ser explicado pelo valor de λ ser próximo de zero. Se aumentarmos o valor de *lambda* para, digamos, dez, temos o seguinte modelo:

$$\begin{aligned}
\log\left(\frac{\pi_{i1}}{\pi_{i3}}\right) &= 0,495 + 0,028z_{i1} + 0,028z_{i2} - 0,025z_{i3} + 0,008z_{i4} + 0,030z_{i5} \\
&\quad - 0,006z_{i6} - 0,025z_{i7} + 0,009z_{i8} - 0,025z_{i9} + 0,009z_{i10}, \\
\log\left(\frac{\pi_{i2}}{\pi_{i3}}\right) &= 0,780 + 0,028z_{i1} + 0,025z_{i2} - 0,026z_{i3} + 0,002z_{i4} + 0,028z_{i5} \\
&\quad - 0,001z_{i6} - 0,025z_{i7} + 0,018z_{i8} + 0,04z_{i9} + 0,016z_{i10}.
\end{aligned} \tag{12}$$

Agora os resultados aproximam-se daqueles comentado por *Friedman et al. (2010)*: os coeficientes foram encolhidos para próximo de zero e os coeficientes das variáveis altamente correlacionadas estão próximos uns dos outros.

Por sua vez, o modelo obtido por regressão lasso fornece o seguinte resultado:

$$\begin{aligned}
\log\left(\frac{\pi_{i1}}{\pi_{i3}}\right) &= 0,401 + 0,033z_{i1} + 0z_{i2} - 0,106z_{i3} + 0z_{i4} + 0,634z_{i5} \\
&\quad - 0z_{i6} - 0,08z_{i7} + 0z_{i8} - 0,543z_{i9} + 0z_{i10}, \\
\log\left(\frac{\pi_{i2}}{\pi_{i3}}\right) &= 0,712 + 0,033z_{i1} + 0z_{i2} - 0,106z_{i3} + 0z_{i4} + 0,634z_{i5} \\
&\quad - 0z_{i6} - 0,08z_{i7} + 0z_{i8} + 1,333z_{i9} + 0z_{i10}.
\end{aligned} \tag{13}$$

Não usamos o λ obtido pela validação cruzada, pois era praticamente zero e as estimativas do modelo eram próximas ao modelo sem penalização. Consideramos, então, $\lambda = 0.1$. Não consideramos o mesmo valor de λ da regressão ridge ($\lambda = 10$), pois mesmo para valores pequenos de λ, os coeficientes das preditoras eram zero. Dito isso, os resultados do modelo com $\lambda = 0.1$ aproximam-se daqueles comentado

por *Friedman et al. (2010)*: entre as preditoras altamente correlacionadas, o modelo manteve uma e discartou o restante.

Note que em nenhum momento propromos um conjunto de candidatos para λ. Ao invés disso, implementamos o algoritmo com os candidatos para λ padrão do pacote **`glmnet`**. Seguimos com a mesma abordagem para a aplicação em dados reais.

3.1 Quantidade de preditoras maior que observações

Nesse caso não avaliamos a forma das estimativas do coeficientes do modelo de regressão como no caso das preditoras correlacionadas. O principal argumento para considerar a regressão penalizada por ridge ou lasso quando a quantidade de preditoras é maior que as observações, é que a estimativa do coeficientes é única, enquanto que a regressão sem penalização não tem solução única para a estimativa dos coeficientes. Portanto, não simulamos dados nesse cenário, pois o interesse não está em saber os valores dos coeficientes.

4. Aplicação em dados reais

As informações do banco de dados MNIST foram extraídas de imagens, onde essas imagens são dígitos escritos à mão de 0 à 9. Cada imagem foi traduzida em uma matriz de 20 linhas e 20 colunas, onde cada elemento dessa matriz é um pixel da imagem. A banco de dados foi construído de tal forma que cada coluna é um pixel, o que resulta em 400 colunas. Logo, temos 400 variáveis preditoras. Além disso, o conjunto de dados tem 5000 imagens, com 500 imagens para cada dígito. A Figura 2 apresenta uma amostra das imagens do banco de dados.

Para aplicar a regressão logística multinomial penalizada como um classificador, separamos 70% dos dados para o conjunto de treino e 30% para o conjunto de teste. Após aplicar validação cruzada para escolher o valor de λ que minimiza o erro de predição, temos que $\lambda = 0,036$ para a regressão ridge e $\lambda = 0,002$ para a regressão lasso. O modelo de regressão ridge teve uma acurácia de 90,87% nos dados de testes, enquanto que a regressão lasso 90,27%. As porcentagens são muito próximas. Isso pode ser pelo fato do valor de λ para as duas penalizações serem próximas de zero, o que pode fazer com que a acurácia seja próxima à do modelo não penalizado ($\lambda = 0$). De fato, a acurácia do modelo de regressão logística multinomial sem penalização é 90,67%. Logo, considerando a capacidade preditiva dos modelos através da acurácia, não temos uma ganho considerável ao aplicar a penalização.

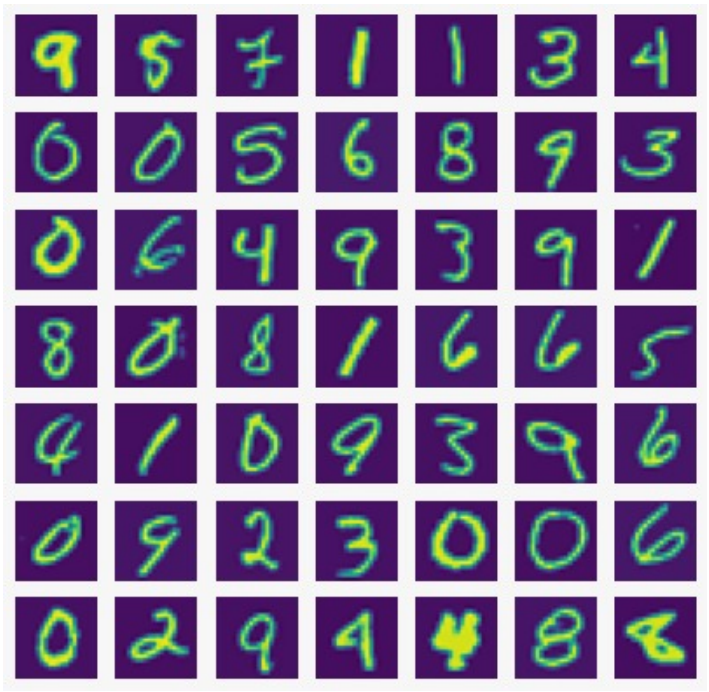

Figura 2: Amostra das imagens no banco de dados

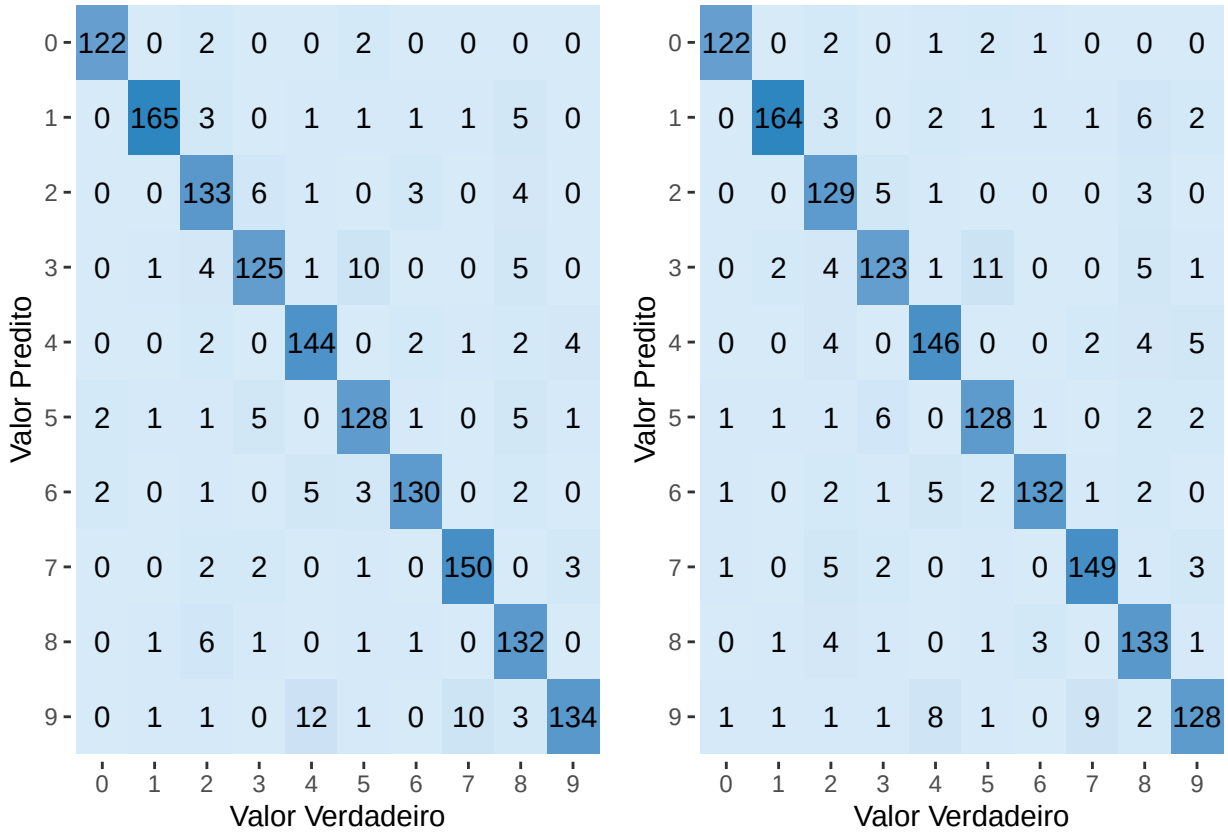

Figura 3: Esquerda: matriz de confusão da regressão ridge para o conjunto de teste. Direita: matriz de confusão da regressão lasso para o conjunto de teste.

A Figura 3 mostra a classificação dos dígitos no conjunto de teste para as regressões ridge e lasso. No geral, o erro de classificação ocorre em dígitos com escritas parecidas, como os pares (4,9) e (7,9). O que nos causa surpresa é que um dos erros de classificação mais altos ocorre entre os pares (3,5), pois não são números com escritas similares.

Vale ressaltar que, como temos a estimativa de milhares de coeficientes, é inviável apresentar as estimativas de cada um. Portanto, o principal objetivo foi contruir um modelo de regressão logística multinomial penalizado como um classificador.

Referências

Friedman, J., Hastie, T. and Tibshirani, J. (2010). **Regularization Paths for Generalized Linear Models via Coordinate Descent.** *Journal of Statistical Software*, Articles 33.