Comparação de Modelos de Representação Latente

José Geraldo Fernandes
Escola de Engenharia
Universidade Federal de Minas Gerais
Belo Horizonte, Brasil
josegeraldof@ufmg.br

I. Introdução

Aprendizado de representação é uma propriedade muito importante que justifica alguma parte do grande sucesso de modelos profundos [1]. Em uma rede neural a última camada é apenas um discriminador linear, cabe, portanto, às camadas anteriores aprender uma representação dos dados que modele o problema de classificação linearmente. Em redes convolucionais, há uma semelhança em o quão fundo está a camada e a complexidade de *features* que o mesmo captura.

Uma vantagem muito grande de trabalhar com espaços lineares é todo o arcabouço formal da teoria de regressão linear [2]. Não só o problema de aprendizado é mais simples como há diversas propriedades interessantes e úteis quando o problema está matematicamente bem formulado.

Outro aspecto interessante do aprendizado de representação é carregar informação *a priori* para direcionar a projeção. Em redes neurais convencionais, o interesse é no desempenho do classificador de forma direta, todavia outro direcionamento é plenamente possível como informação de vizinhança [3].

Esse trabalho explora métodos de aprendizado de representação, ou projeção, e estuda uma propriedade conveniente do regressor linear. Também propõe um método heurístico de seleção de amostras para treinamento.

II. Revisão Bibliográfica

A. Extreme Learning Machine

As *Extreme Learning Machines* (ELM) [4] foram uma abordagem para tratar o problema de alto custo computacional no treinamento de redes neurais convencionais. Em vez de ajustar os parâmetros da rede de forma iterativa, tenta-se uma projeção aleatória suficientemente grande para modelar os dados em um espaço de representação linear.

Uma matriz de projeção inicial aleatória $\boldsymbol{Z}$ é gerada e aplica-se uma não linearidade, como na Equação 1.

$$\boldsymbol{H} = S(\boldsymbol{X}\boldsymbol{Z}) \tag{1}$$

O problema de classificação é então modelado no novo espaço $\boldsymbol{H}$ e pode ser resolvido por qualquer técnica tradicional.

B. Autoencoder

Autoencoders são apenas uma rede neural convencional replicando a entrada na saída, isto é, o objetivo da rede é apenas reconstruir os dados e encontrar uma representação latente [5]. Pode-se considerar uma rede dividida em dois módulos: *encoder* $\boldsymbol{H} = f(\boldsymbol{X})$, que faz a projeção latente; e, o *decoder* $\boldsymbol{X} = g(\boldsymbol{H})$ que tenta reconstruir a saída a partir dessa representação.

É especialmente útil para encontrar uma representação latente compacta de dados muito extensos ou complexos, como imagens e sons, de outra forma intratáveis. Há trabalhos que mostram alto ganho de desempenho em comparação com outras técnicas mais clássicas de redução de dimensionalidade [6].

Apesar disso, o sentido contrário também é possível, principalmente quando há uma direção *a priori* que guie o treinamento da projeção, como dispersibilidade [7] por exemplo.

De qualquer forma, espera-se que o novo espaço $\boldsymbol{H}$ traga propriedades que facilitem em algum aspecto o problema de aprendizado.

C. Grafo de Gabriel

O Grafo de Gabriel é uma construção a partir de um conjunto de pontos $\mathcal{S}$, que define os vértices do grafo, para outro conjunto de arestas $\mathcal{E}$ tal que dois pontos $\boldsymbol{x_i}, \boldsymbol{x_j}$ são adjacentes, definem uma aresta, se não há outro ponto de $\mathcal{S}$ dentro da hiperesfera definida com a distância entre os dois pontos $\boldsymbol{x_i}, \boldsymbol{x_j}$ como diâmetro, como na Equação 2.

$$\begin{gathered}(\boldsymbol{x_i}, \boldsymbol{x_j}) \in \boldsymbol{\mathcal{E}} \leftrightarrow \\ \delta^2(\boldsymbol{x_i}, \boldsymbol{x_j}) \leq [\delta^2(\boldsymbol{x_i}, \boldsymbol{x_k}) + \delta^2(\boldsymbol{x_j}, \boldsymbol{x_k})] \ \forall \boldsymbol{x_k} \in \mathcal{S}\end{gathered} \tag{2}$$

Onde δ é a métrica de distância na construção e comumente definida como a distância euclidiana.

A construção desse grafo é uma técnica da Geometria Computacional e emprestada para os problemas de aprendizado como uma forma de expressar as características de vizinhança da estrutura do conjunto de dados.

1) Métrica de Qualidade: Baseado na diferença de classes entre o vértice e as amostras conectadas por suas arestas, define-se uma grandeza Q que representa a qualidade dessa amostra como na Equação 3, onde V representa o número de arestas e V_{eq} o número dessas para amostras de classe coincidente.

$$Q(\boldsymbol{x}_i) = \frac{V_{eq}(\boldsymbol{x}_i)}{V(\boldsymbol{x}_i)} \tag{3}$$

Amostras na região de fronteira e sobreposição terão sua qualidade afetada quão maior for a mistura. Também, amostras

com qualidade perfeita $Q(\boldsymbol{x}_i) = 1$ estão em regiões predominantes e longe da fronteira.

A partir do princípio que as amostras mais importantes para o treinamento são as localizadas na região de fronteira, uma estratégia de seleção é filtrar amostras com essa métrica de qualidade alta.

III. Metodologia

A. Base de Dados

Para avaliação dos índices de qualidade dos agrupamentos seleciona-se um conjunto de 15 *datasets* padrão do repositório UCI [8] em todos os testes. Esses são separados em 10 partições determinadas para validação cruzada *k-fold*.

Para confiabilidade do resultado, as bases de dados selecionadas são amplamente utilizadas em aplicações semelhantes. Como pré-processamento, fez-se uma normalização de todos os atributos e conversão dos categóricos em númericos, necessário para o algoritmo CLAS. Todos são de problemas de classificação binária. Segue a descrição da seleção: *Statlog Australian Credit Approval* (australian); *Banknote Authentication* (banknote); *Breast Cancer Wisconsin* (breastcancer); *Breast Cancer Hess Probes* (breastHess); *Liver Disorders* (bupa); *Climate Model Simulation Crashes* (climate); *Pima Indian Diabetes* (diabetes); *Fertility* (fertility); *Statlog German Credit Data* (german); *Gene Expression* (golub); *Haberman's Survival* (haberman); *Statlog Heart Dicease* (heart); *Indian Liver Patient* (ILPD); *Parkinsons* (parkinsons); *Connectionist Bench Sonar, Mines vs. Rocks* (sonar).

A Tabela I mostra as principais características dessa seleção, η representa a proporção entre as classes. Note a alta diversidade dos problemas para atestar a generalização do método.

Tabela I
Características das bases de dados selecionadas.

Dataset	**Amostras**	**Atributos**	η
australian	690	14	0.44
banknote	1372	4	0.44
breastcancer	683	6	0.65
breastHess	133	30	0.74
bupa	345	6	0.42
climate	540	18	0.91
diabetes	768	8	0.65
fertility	100	9	0.12
german	1000	24	0.70
golub	72	50	0.65
haberman	306	3	0.74
heart	270	13	0.56
ILPD	579	10	0.72
parkinsons	195	22	0.75
sonar	208	60	0.47

B. Arquitetura

Neste trabalho comparou-se duas arquiteturas parecidas: ELM; e, *autoencoder* (AE) com uma camada escondida. Na primeira, calcula-se uma projeção aleatória de tamanho p, comumente maior que a dimensão dos dados. Da mesma forma, no AE uma mesma projeção de tamanho p é treinada a partir da reconstrução do sinal de entrada, substituindo $\boldsymbol{X}$ na saída de uma rede *multilayer perceptron* convencional.

A esperança dessa projeção é representar os dados em um espaço $\boldsymbol{H}$ mais tratável onde o problema é linearmente separável. Dessa forma, conclui-se o treinamento com uma regressão linear padrão como na Equação 4, onde λ é um parâmetro de regularização L2.

$$\boldsymbol{W} = (\boldsymbol{H}^T\boldsymbol{H} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{H}^T\boldsymbol{y} \tag{4}$$

A síntese do modelo é direta, faz-se a projeção de novos dados a partir da rede aprendida, e calcula-se a regressão linear com os parâmetros $\boldsymbol{W}$, como na Equação 5. No problema de classificação, aplica-se também uma função linear como degrau ou sigmoide.

$$\hat{\boldsymbol{y}} = \text{projecao}(\boldsymbol{X})\boldsymbol{W} = \boldsymbol{H}\boldsymbol{W} \tag{5}$$

C. Seleção de Parâmetros

Nesse tipo de modelo, combinação de projeção e regressão linear, há apenas dois principais parâmetros de ajuste: o tamanho da projeção p; e, o peso da regularização λ.

Uma das formas de fazer o ajuste é a partir de validação cruzada com conjuntos separados de teste. Contudo, esse método pode ser custoso computacionalmente além de reduzir o tamanho efetivo de dados no treinamento.

Nesse sentido, e em virtude de aproveitar a teoria fundamentada de regressão linear, utiliza-se uma equação fechada para o erro de validação cruzada *leave-one-out* (LOO) do regressor nos dados da projeção, como na Equação 6.

$$\begin{aligned} \boldsymbol{A} &= (\boldsymbol{H}^T\boldsymbol{H} + \lambda\boldsymbol{I}_p)^{-1} \\ \boldsymbol{P} &= \boldsymbol{I}_N - \boldsymbol{H}\boldsymbol{A}\boldsymbol{H}^T \\ \boldsymbol{\sigma}_i &= \text{diag}(\boldsymbol{P})_i^{-1}(\boldsymbol{P}\boldsymbol{y})_i \\ \text{LOO} &= N^{-1}\boldsymbol{\sigma}^T\boldsymbol{\sigma} \end{aligned} \tag{6}$$

Espera-se haver uma correspondência entre otimizar essa função objetivo LOO, menos custosa, e a generalização, desempenho em novos dados.

A otimização multiobjetivo foi feita de forma simples e em duas etapas. Primeiro encontrou-se p ótimo em varredura com $\lambda = 0$, e, em seguida, variou-se o parâmetro de regularização também em varredura.

D. Classificação

O problema de classificação segue o padronizado. Para cada *dataset* separou-se 10 *folds* fixos, no conjunto de treinamento treinou-se a rede de projeção e o regressor com os parâmetros p, λ ótimos, no conjunto de teste avaliou-se o desempenho.

E. Amostragem

Para tratar o problema de classes desbalanceadas, há alguns exemplos graves como exposto na Tabela I, utilizou-se duas estratégias de amostragem.

Como referência fez-se uma amostragem aleatória da classe minoritária até manter o número de amostras balanceado,

upsample. E, como método proposto, utilizou-se o filtro de seleção a partir do Grafo de Gabriel retirando amostras com qualidade total $Q(\boldsymbol{x}_i) = 1$, *downsample*.

IV. Resultados

A. Seleção de Parâmetros

Para assegurar a esperança de correspondência entre a função objetivo LOO e o desempenho em um conjunto de teste, avaliou-se simultaneamente ambos atributos. Na validação cruzada, a métrica foi o erro quadrático médio (MSE).

O comportamento observado é variado em função do *data-set*, as Figuras 1 e 2 mostram um experimento de varredura em p onde as curvas parecem próximas. Contudo, há exemplos, como nas Figuras 3 e 4, onde, apesar de alguns traços serem bem convergentes, a curva LOO diverge quando MSE parece ainda decrescer.

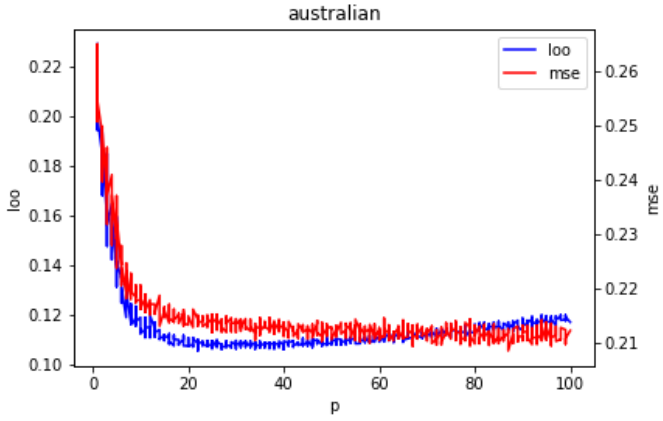

Figura 1. Curvas do custo LOO, no conjunto de treinamento, e MSE, no conjunto de teste; *dataset* australian e avaliando p na ELM.

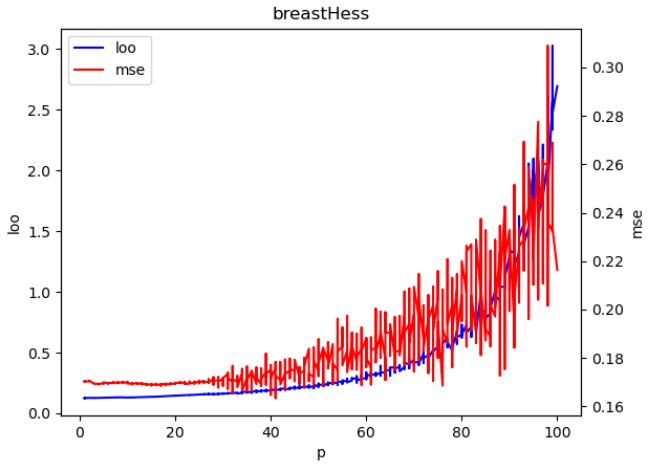

Figura 2. Curvas do custo LOO, no conjunto de treinamento, e MSE, no conjunto de teste; *dataset* breastHess e avaliando p no AE.

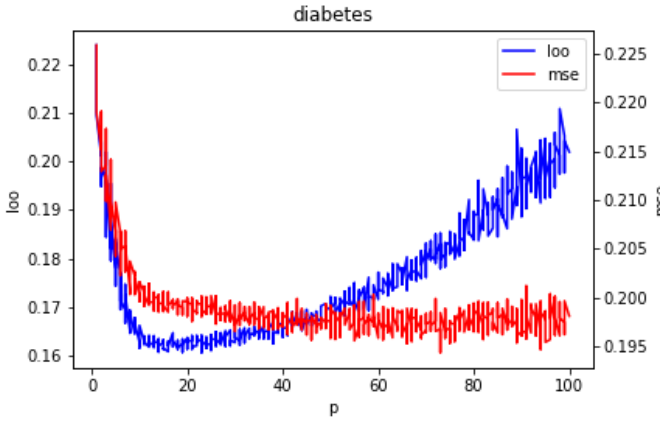

Figura 3. Curvas do custo LOO, no conjunto de treinamento, e MSE, no conjunto de teste; *dataset* diabetes e avaliando p na ELM.

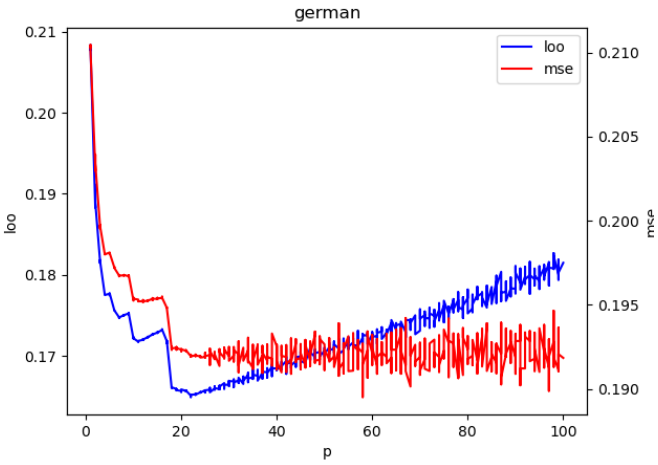

Figura 4. Curvas do custo LOO, no conjunto de treinamento, e MSE, no conjunto de teste; *dataset* german e avaliando p no AE.

Na varredura de λ esse fenômeno também acontece. As Figuras 5 e 6 mostram curvas similares das funções de custo, enquanto as Figuras 7 e 8 mostram um comportamento divergente ainda mais agressivo, apesar da escala de variação ser menor.

Os parâmetros foram selecionados por inspeção visual do mínimo da curva LOO. Para ambas redes são expostos nas Tabelas II e III.

B. Classificação

As Tabelas IV e V mostram os resultados obtidos com três métricas de desempenho em todos os conjuntos de dados. Número um, o LOO, único calculado a partir do conjunto de treinamento, usado para seleção de parâmetros. Número dois e três, o custo MSE e a área sob a curva ROC (AUC), calculado no conjunto de teste.

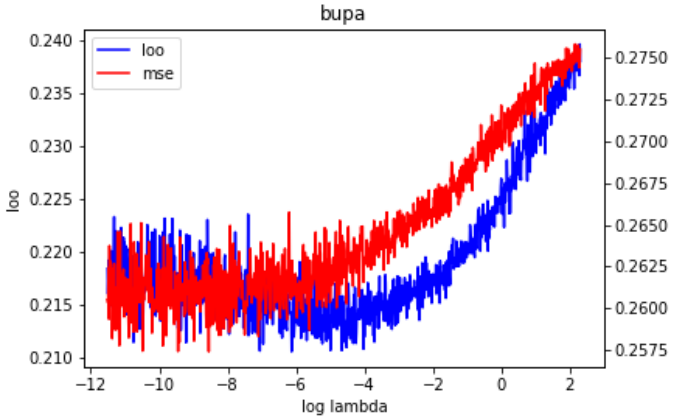

Figura 5. Curvas do custo LOO, no conjunto de treinamento, e MSE, no conjunto de teste; *dataset* bupa e avaliando λ na ELM.

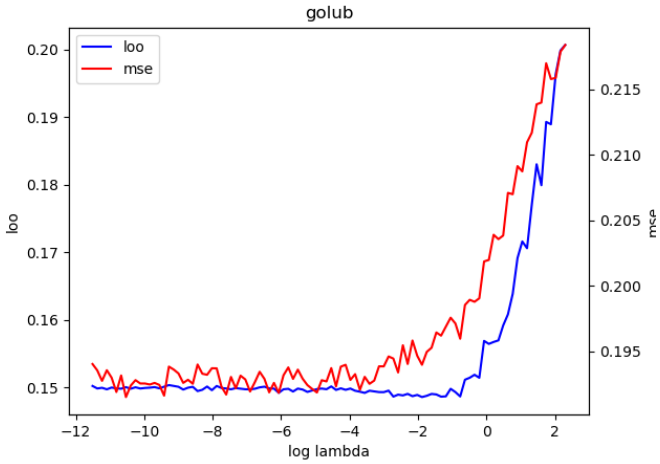

Figura 6. Curvas do custo LOO, no conjunto de treinamento, e MSE, no conjunto de teste; *dataset* golub e avaliando p no AE.

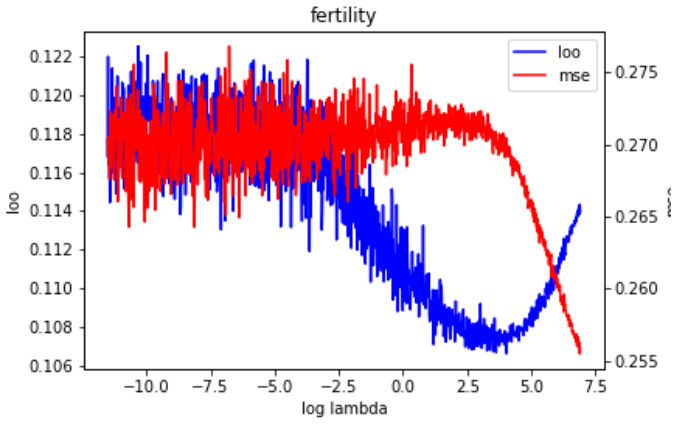

Figura 7. Curvas do custo LOO, no conjunto de treinamento, e MSE, no conjunto de teste; *dataset* fertility e avaliando p na ELM.

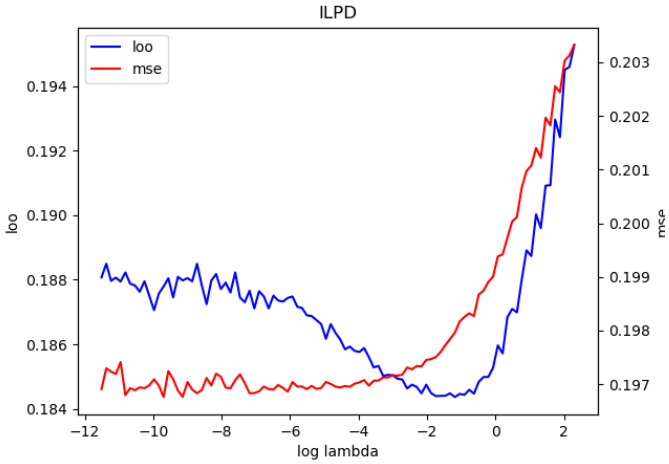

Figura 8. Curvas do custo LOO, no conjunto de treinamento, e MSE, no conjunto de teste; *dataset* ILPD e avaliando p no AE.

Tabela II
PARÂMETROS SELECIONADOS ELM.

Dataset	p	λ
australian	20	0
banknote	100	-12
breastcancer	20	-2
breastHess	10	0
bupa	15	-5
climate	25	-2
diabetes	15	-2
fertility	5	3
german	25	2
golub	5	1
haberman	5	-2
heart	15	0
ILPD	10	-1
parkinsons	30	-3
sonar	30	2

Tabela III
PARÂMETROS SELECIONADOS AE.

Dataset	p	λ
australian	20	-12
banknote	100	-12
breastcancer	40	-8
breastHess	5	-1
bupa	5	-1
climate	20	-1
diabetes	10	-1
fertility	5	2
german	20	-1
golub	5	-2
haberman	5	-2
heart	10	-1
ILPD	10	-1
parkinsons	10	-1
sonar	10	-1

Tabela IV
RESULTADOS DE DESEMPENHO ELM.

Dataset	**LOO**	**MSE**	**AUC**
australian	0.11	0.21	0.92
banknote	0.01	0.17	1.00
breastcancer	0.03	0.15	0.99
breastHess	0.14	0.18	0.86
bupa	0.22	0.26	0.71
climate	0.06	0.11	0.92
diabetes	0.16	0.20	0.83
fertility	0.11	0.27	0.57
german	0.18	0.20	0.75
golub	0.17	0.21	0.79
haberman	0.18	0.19	0.67
heart	0.14	0.20	0.90
ILPD	0.19	0.20	0.69
parkinsons	0.10	0.16	0.89
sonar	0.18	0.24	0.79

Tabela V
RESULTADOS DE DESEMPENHO AE.

Dataset	**LOO**	**MSE**	**AUC**
australian	0.11	0.21	0.93
banknote	0.02	0.18	1.00
breastcancer	0.03	0.14	0.99
breastHess	0.13	0.17	0.91
bupa	0.24	0.27	0.59
climate	0.06	0.11	0.96
diabetes	0.16	0.20	0.83
fertility	0.11	0.27	0.56
german	0.17	0.19	0.80
golub	0.15	0.20	0.85
haberman	0.18	0.19	0.69
heart	0.13	0.20	0.89
ILPD	0.18	0.20	0.72
parkinsons	0.12	0.17	0.87
sonar	0.17	0.24	0.84

C. Amostragem

A Tabela VI mostra o número médio de amostras e de proporção entre as classes η no conjunto de treinamento após o *downsample*.

As Tabelas VII e VIII mostram os resultados obtidos das duas estratégias de amostragem a partir da área sob a curva ROC (AUC), calculado no conjunto de teste.

V. DISCUSSÕES

Os resultados de desempenho mostraram que a projeção aleatória é suficientemente tão boa quanto a aprendida pelo AE. Isso é razoável principalmente nos casos em que a dimensão é aumentada, se a função de custo é apenas a reconstrução então é simples encontrar uma função inversível que modele o AE.

Sobre o desempenho, o modelo se mostrou competitivo apesar de um fenômeno incômodo na saída. A Figura 9 mostra um experimento com boa modelagem da resposta, a variação entre classes é clara mas também pequena. Seria interessante que a resposta fosse mais polarizada demonstrando mais confiança do classificador.

Tabela VI
CARACTERÍSTICAS DO CONJUNTO DE TREINAMENTO NA TÉCNICA DE *downsampling*.

Dataset	**Amostras**	η
australian	427.8	0.61
banknote	586	0.94
breastcancer	243.1	0.12
breastHess	76.6	0.62
bupa	299.6	0.42
climate	476.6	0.91
diabetes	660.7	0.67
fertility	40.5	0.09
german	795.1	0.67
golub	52.6	0.58
haberman	169.2	0.71
heart	206	0.62
ILPD	362.9	0.65
parkinsons	83.9	0.49
sonar	178.8	0.44

Tabela VII
RESULTADOS DE DESEMPENHO DAS TÉCNICAS DE AMOSTRAGEM, ELM.

Dataset	**Base**	***Upsample***	***Downsample***
australian	0.92	0.92	0.92
banknote	1.00	1.00	0.96
breastcancer	0.99	1.00	0.99
breastHess	0.86	0.88	0.86
bupa	0.71	0.74	0.71
climate	0.92	0.90	0.92
diabetes	0.83	0.83	0.83
fertility	0.57	0.56	0.65
german	0.75	0.78	0.79
golub	0.79	0.87	0.75
haberman	0.67	0.70	0.65
heart	0.90	0.89	0.87
ILPD	0.69	0.71	0.71
parkinsons	0.89	0.89	0.87
sonar	0.79	0.81	0.83

Tabela VIII
RESULTADOS DE DESEMPENHO DAS TÉCNICAS DE AMOSTRAGEM, AE.

Dataset	**Base**	***Upsample***	***Downsample***
australian	0.93	0.92	0.92
banknote	1.00	1.00	1.00
breastcancer	0.99	1.00	0.99
breastHess	0.91	0.91	0.91
bupa	0.59	0.60	0.59
climate	0.96	0.96	0.96
diabetes	0.83	0.83	0.83
fertility	0.56	0.53	0.59
german	0.80	0.79	0.80
golub	0.85	0.87	0.84
haberman	0.69	0.69	0.68
heart	0.89	0.90	0.90
ILPD	0.72	0.73	0.72
parkinsons	0.87	0.87	0.88
sonar	0.84	0.84	0.84

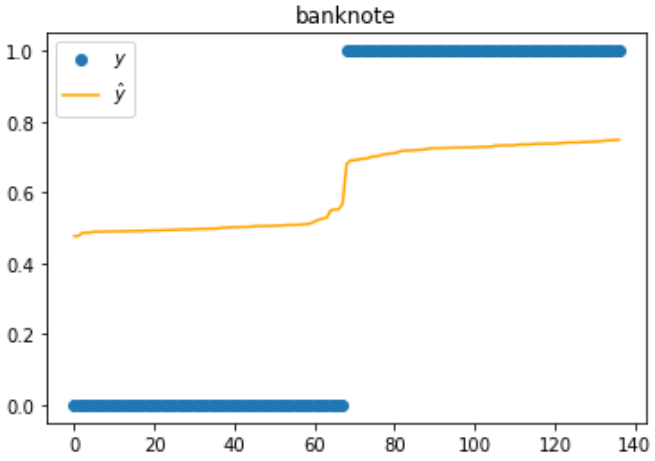

Figura 9. Resposta do modelo e classes em dispersão, *dataset* banknote.

Esse comportamento foi comum independente do conjunto de dados e do modelo, a Figura 10 mostra um desempenho pior. A Figura 11 mostra a curva ROC correspondente.

Uma hipótese para esse fenômeno é o uso de um regressor linear não ser tão polarizado mesmo no espaço latente.

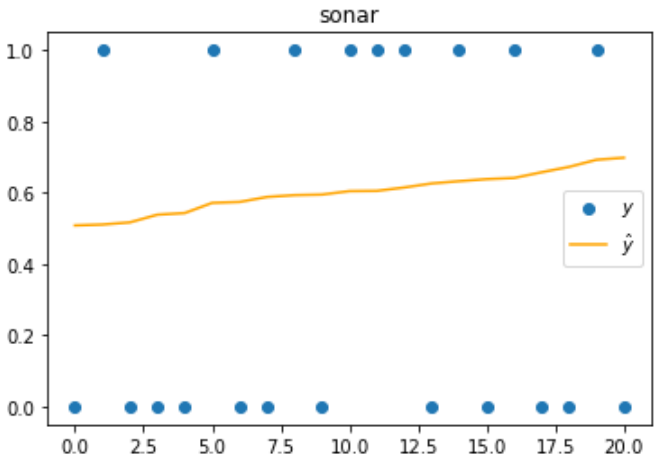

Figura 10. Resposta do modelo e classes em dispersão, *dataset* sonar.

Sobre as técnicas de amostragem houve um aumento razoável em desempenho inclusive do método proposto. Note, contudo, que esse não tratou o problema de desbalanceamento em si, parâmetro η, mesmo conseguindo um ganho de desempenho.

VI. Conclusões

Este trabalhou apresentou uma comparação entre os modelos de projeção ELM e AE. Os resultados sugerem pouca diferenciação, apesar do segundo ter uma versatilidade de carregar informação *a priori* no treinamento da projeção, algo que pode ser explorado a partir da função de custo.

Além disso, foi estudado o uso de uma métrica para estimar o desempenho de generalização durante o treinamento.

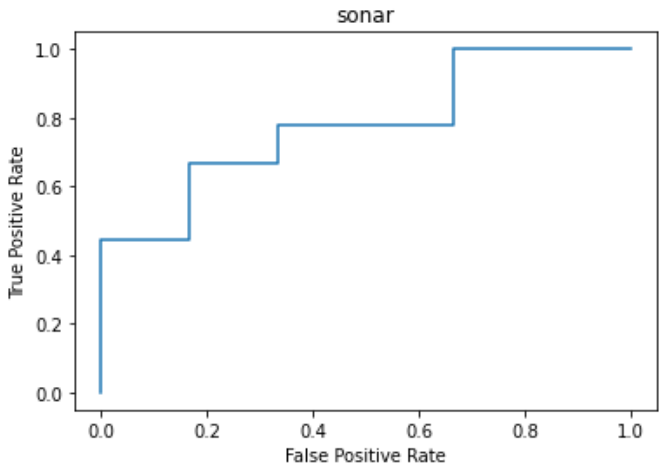

Figura 11. Curva ROC, *dataset* sonar.

Experimentos mostram uma convergência do risco real com a métrica mas carece uma demonstração mais fundamentada. O processo de seleção de parâmetros também foi muito simples. Técnicas mais avançadas de otimização não linear ou até uma varredura dupla simultânea certamente indicaram um resultado melhor.

Sobre fenômeno de respostas não polarizadas uma sugestão é o uso de um regressor logístico que também tem uma teoria matemática rica e uma resposta mais polarizada, mais comum em problemas de classificação.

Finalmente, o método de *downsampling* não parece recomendado para o problema de desbalanceamento mas se mostrou promissor na direção de seleção de amostras para treinamento. Note que não só o desempenho aumentou mas também o número de amostras foi reduzido, o que acelerou o treinamento.

Agradecimento

Este trabalho foi possível pela disponibilização das bases de dados pelo repositório *UCI Machine Learning* [8].

Referências

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[2] G. Casella and R. L. Berger, *Statistical inference*. Cengage Learning, 2021.

[3] R. Salakhutdinov and G. Hinton, "Learning a nonlinear embedding by preserving class neighbourhood structure," in *Artificial Intelligence and Statistics*, pp. 412–419, PMLR, 2007.

[4] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: a new learning scheme of feedforward neural networks," in *2004 IEEE international joint conference on neural networks (IEEE Cat. No. 04CH37541)*, vol. 2, pp. 985–990, Ieee, 2004.

[5] M. A. Kramer, "Nonlinear principal component analysis using autoassociative neural networks," *AIChE journal*, vol. 37, no. 2, pp. 233–243, 1991.

[6] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[7] A. Ng *et al.*, "Sparse autoencoder," *CS294A Lecture notes*, vol. 72, no. 2011, pp. 1–19, 2011.

[8] D. Dua and C. Graff, "UCI machine learning repository," 2017.