Escolha automática da métrica de distância em problemas de regressão

1

Abstract. *This work examines the importance of choosing the appropriate distance measure in machine learning algorithms, specifically in regression problems. The Euclidean distance is commonly used, but the study found that the distance measure can affect the accuracy of the regression model and proposes two solutions to determine the appropriate measure. The first evaluates the variance of the distances between instances, while the second uses a heuristic that considers the linearity between distances and the obtained outputs. In total, we evaluated the choice of distance measures (11 possibilities) in 10 publicly available datasets. The linearity heuristic had a higher correlation with the evaluated regressor's output and therefore was able to choose the best measures to be used in the datasets evaluated in this work.*

Resumo. *Este trabalho examina a importância da escolha da medida de distância apropriada em algoritmos de aprendizado de máquina, especificamente em problemas de regressão. A distância euclidiana é comumente usada, mas o estudo descobriu que a medida de distância pode afetar o acerto do modelo de regressão e propõe duas soluções para determinar a medida apropriada. A primeira avalia a variância das distâncias entre instâncias, enquanto a segunda usa uma heurística que considera a linearidade entre as distâncias e as saídas obtidas. No total, avaliamos a escolha das medidas de distância (11 possibilidades) em 10 conjuntos de dados disponíveis publicamente. A heurística da linearidade teve uma maior correlação com a saída do regressor avaliado e, portanto, conseguiu escolher as melhores medidas a serem utilizadas nos conjuntos de dados avaliados neste trabalho.*

1. Introdução

O aprendizado de Máquina é uma área da Inteligência Artificial (IA) cujo objetivo é o desenvolvimento de técnicas computacionais sobre o aprendizado bem como a construção de sistemas capazes de adquirir conhecimento de forma automática [Monard and Baranauskas 2003]. Dentre as aplicações para estas técnicas podemos citar: previsão de carga e consumo de energia elétrica [Conde et al. 2007], previsão do deslocamento de tempestades severas [Rozin 2018], previsão dos movimentos do Ibovespa [Finkler 2017], predição de evasão de estudantes no ensino público superior [Silva 2022], entre outras.

Para conseguir mapear as entradas e predizer uma saída, internamente, os algoritmos de forma geral precisam utilizar medidas/métricas de distância para comparar as instâncias do problema e encontrar padrões (no restante deste trabalho indicaremos simplesmente como medidas de distância).

Em diversos trabalhos encontrados na literatura a escolha da melhor medida de distância, que será utilizada pelos algoritmos, não é levada em consideração, sendo comum o uso de medidas tradicionais, tais como a Euclidiana. Em um cenário ideal a melhor medida de distância é aquela, que dado um conjunto de dados, possui maior capacidade comparativa de instâncias, sendo capaz de discriminar diferentes entradas dentro de um contexto específico. Além disso, os poucos trabalhos que tiveram a preocupação em escolher de forma mais criteriosa a medida de distância utilizada são trabalhos em bases de classificação de dados, onde se busca encontrar uma classe, dentro das possibilidades limitadas existentes.

Neste trabalho abordaremos o problema de selecionar a melhor medida de distância especificamente em problemas de regressão. Nos problemas de regressão a saída retornada pelo modelo é um valor contínuo, levando a um número ilimitado de valores possíveis previstos pelo modelo.

Selecionamos 11 medidas de distância, que serão utilizadas por algoritmos de aprendizado de máquina em 10 conjunto de dados de problemas de regressão, para avaliarmos quais são as mais apropriadas. Esperamos com esta análise responder se existe uma maneira automática de escolher a melhor medida de distância para um dado problema de regressão.

Os resultados obtidos mostram que não existe uma "bala de prata" que se destaca como obtendo melhores resultados independentemente do contexto. Concluímos então que a escolha da melhor medida de distância não é obvia, e encontrar um método que seja capaz de solucionar este problema não é uma tarefa trivial e sua solução é capaz de trazer melhorias na qualidade dos modelos produzidos.

Este trabalho está organizado da seguinte maneira: a Seção 2 apresenta os trabalhos que serviram de base para este artigo. A Seção 3 descreve as principais medidas de distância utilizadas atualmente em problemas de regressão e os conjuntos de dados selecionados para avaliação. A Seção 4 mostra os resultados obtidos sobre os possíveis processos para escolha da medida de distância a ser utilizada e a Seção 5 apresenta as conclusões e propostas de continuidade deste trabalho.

2. Trabalhos Relacionados

Em [Aggarwal et al. 2001] os autores analisaram o comportamento das medidas de distância em espaços de alta dimensionalidade. Nestes espaços os dados se tornam dispersos, o conceito de proximidade perde quantitativamente o significado e os algoritmos falham nos quesitos de eficiência e eficácia. Baseando-se na norma L_k os autores mostram que o problema da significância é sensível aos valores de k, sendo preferível nesse contexto optar por valor de k menores.

Em [Giancarlo et al. 2010] os autores investigaram qual a melhor medida de distância para se trabalhar com dados de microarrays. No estudo, diferentes medidas de distância foram aplicadas a 6 conjuntos de dados de microarrays, cuja solução "ouro" era conhecida. Os autores não conseguiram chegar a uma conclusão e definir a melhor medida, porém, os resultados obtidos entendem o trabalho de [Costa et al. 2004] — que recomenda as medidas de Pearson, Cosseno e Euclidiana para problemas desta natureza — uma vez que foram utilizados no experimento diferentes tipos de microarrays.

Table 1. Medidas de distâncias selecionadas. O parâmetro p representa a ordem da distância de Minkowski.

Id.	Medida de distância
0	Euclidiana
1	Minkowski p=0.2
2	Minkowski p=0.5
3	Minkowski p=0.8
4	Manhattan
5	Canberra
6	Bray-Curtis
7	Cosseno
8	Hamming
9	Correlação
10	Chebyshev

Em [Jaskowiak et al. 2014] os autores abordam o problema da escolha da métrica de distância ao se trabalhar com algoritmos de agrupamento voltados para análise de expressões genéticas. Segundo os autores escassos são os estudos com foco no problema em questão, não sendo encontrado nenhum que auxilie na tarefa de se escolher uma entre as várias métricas, ou indique as que devam ser preferidas, priorizadas ou evitadas. Dentre os trabalhos que comparam as medidas de distância aplicadas aos conjuntos de conjuntos, há aqueles que testaram um número reduzido de medidas e/ou conjuntos de dados, ou que trataram problemas de natureza diferentes de maneira igual, o que não é adequado.

[Jaskowiak et al. 2014] é uma complementação de um estudo anterior dos autores [Jaskowiak et al. 2013], que avaliou medidas de distância para séries temporais genéticas e dados cancerígenos, independentemente. Neste estudo os autores verificaram que a seleção de uma distância apropriada depende do cenário em questão. Além disso, em cada cenário, dado o mesmo método de agrupamento, diferenças significativas na qualidade podem surgir da seleção de medidas de distância distintas.

3. Medidas de distância avaliadas e conjuntos de dados utilizados

Conforme observado nos trabalhos anteriores, a escolha da medida de distância influência no desempenho dos algoritmos de aprendizado de máquina. O presente trabalho está voltado para medidas de distância destinados aos problemas de regressão.

Neste trabalho extrapolamos o conceito de métrica de distância incluindo também medidas que não respeitam todas as condições necessárias para caracterização de uma métrica. Todas as medidas selecionadas estão disponíveis na ferramenta Sklearn e são mostradas na Tabela 1.

Os conjuntos de dados utilizados no trabalho foram obtidos nos repositórios UCI[Dua and Graff 2017] e de Luiz Torgo [1], e são descritos na Tabela 2. Os conjuntos de dados foram normalizados utilizando um Z-Score, onde os dados são mapeados em uma distribuição onde a média é definida como 0 e o desvio padrão como 1.

[1] `https://www.dcc.fc.up.pt/~ltorgo/`

Table 2. Conjuntos de dados selecionados.

Id.	Conjunto de dados	# Instâncias	# Atributos
0	diabetes	43	3
1	mcs_ds_edited_iter_shuffled	107	6
2	airfoil_self_noise	1503	6
3	qsar_aquatic_toxicity	546	9
4	bank8FM	4499	9
5	concrete	1030	9
6	energy_efficience	768	10
7	daily_demand_forecasting_orders	60	13
8	auto_price	159	16
9	bank32nh	4500	33

Table 3. Resultados KNN: RMSE. Melhores resultados em destaque.

	Medida de distância										
Dataset	0	1	2	3	4	5	6	7	8	9	10
0	**0.8345**	0.919	0.8633	0.8481	0.8681	0.9447	0.9204	0.8871	1.0279	1.1863	0.8748
1	0.5391	0.5503	0.5821	0.5416	**0.5287**	0.5706	0.5589	0.5449	0.8728	0.5351	0.6232
2	0.6292	0.7586	0.6868	0.6285	**0.6149**	0.6844	0.6443	0.7129	1.1146	0.9869	0.7335
3	0.8155	0.8656	0.8057	0.8037	0.7945	0.8081	**0.7872**	0.8215	1.0454	0.8178	0.7901
4	**0.3769**	0.6794	0.5578	0.4431	0.4061	0.5876	0.3842	0.3892	1.3572	0.5406	0.3975
5	0.7457	0.7936	0.7453	0.7352	**0.7319**	0.7561	0.7700	0.7518	1.1798	0.7696	0.7856
6	0.2269	0.2286	0.2092	0.1975	**0.1921**	0.2166	0.1986	0.2404	0.2401	0.2602	0.3421
7	0.4871	0.6348	0.5301	0.5061	0.5018	0.484	**0.472**	0.5748	1.0821	0.7667	0.6479
8	0.6200	0.5954	0.5996	0.5812	**0.5731**	0.6005	0.6448	0.677	0.8643	0.6759	0.76
9	0.9144	1.0423	0.9898	0.9734	0.9564	1.0239	0.9376	0.8988	1.1563	**0.8936**	0.9106

Os códigos desenvolvidos neste trabalho serão disponibilizados após o processo de revisão.

4. Estudo de caso: Regressão por KNN

Para avaliarmos qual a medida de distância mais apropriada em cada um dos conjuntos de dados selecionados, inicialmente construímos um modelo regressor utilizando a implementação do algoritmo Regressor KNN (*K-Nearest Neighbors Regressor*) fornecida pelo Sklearn [2].

O KNN utilizado em nossos estudos foi parametrizado com k=3 vizinhos e durante o processo de validação cruzada dividimos os dados em 5 porções para então calcularmos o erro de previsão modelo por meio do RMSE (*root-mean-square error*). Os resultados obtidos são mostrados na Tabela 3.

Como podemos verificar não existe uma medida de distância que pode ser considerada a melhor em todos os casos. Observamos que a distância Euclidiana (medida padrão do algoritmo KNN na implementação fornecida pelo Sklearn) é superada por outras medidas em grande parte dos conjuntos de dados, assim reafirmamos o uso indiscriminado desta métrica, como comumente encontramos na literatura, é inadequado, impactando di-

[2] `https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html`

Table 4. Resultados Variância

Dataset	Medida de distância										
	0	1	2	3	4	5	6	7	8	9	10
0	0.2449	0.3357	0.2874	0.2689	0.2655	0.2875	0.2201	0.6976	0	**1**	0.2167
1	0.1548	0.9306	0.5263	0.3713	0.3072	0.371	0.3302	0.959	0	**1**	0.1109
2	0.2703	0.5472	0.3615	0.3201	0.3097	0.3874	0.2376	0.913	0	**1**	0.2625
3	0.0817	0.245	0.2085	0.1769	0.16	0.4436	0.4861	**1**	0.1381	0.9763	0
4	0.2926	0.606	0.4844	0.4439	0.4195	0.4282	0.3077	0.9066	0	**1**	0.1627
5	0.3155	0.8098	0.5508	0.4486	0.4053	0.329	0.3764	0.991	0	**1**	0.4303
6	0.2644	0.79	0.6612	0.5563	0.4905	0.6036	0.0654	0.9454	0	**1**	0.108
7	0.4938	0.618	0.5552	0.5112	0.5007	0.4074	0.7425	**1**	0	0.9137	0.6146
8	0.2234	0.5479	0.4291	0.3533	0.3157	0.4698	0.7065	**1**	0	0.9488	0.2464
9	0.4192	**1**	0.6814	0.5839	0.544	0.3532	0.8167	0.9375	0	0.9544	0.5896

retamente no resultado do modelo. Além disso, concluímos que a escolha automática da melhor métrica de distância é uma tarefa difícil.

Agora tentaremos automatizar o processo de escolha da métrica de distância utilizada em cada um dos conjuntos de dados. Para tanto, duas tentativas serão realizadas. Na primeira a escolha será guiada pela variância das distâncias entre as instâncias de treinamento (subseção 4.1) e na segunda propomos uma heurística que considera a linearidade das distâncias em relação a saída dos dados (subseção 4.2).

4.1. Primeira tentativa para escolha da medida de distância: Variância

Em nossa primeira tentativa para escolha da métrica de distância partimos da premissa de que a variância, calculada a partir das distâncias entre as instâncias do conjunto de dados, serviria como indicador para o desempenho da métrica utilizada. A variância é uma medida de dispersão que mostra o quão distante cada valor do conjunto de dados (instância) está do valor central (médio). Acreditamos que quanto maior a variância obtida, maior a capacidade da métrica em diferenciar as instâncias dos problemas, logo melhor seu desempenho. Assim, procuramos medidas de distância que retornam grandes variâncias ao comparar instâncias do mesmo conjunto de dados.

Os resultados obtidos são apresentados na tabela 4. Para uma melhor visualização, análise e comparativo de desempenho foi realizada uma normalização MIN-MAX dos valores. Notamos que a distância Euclidiana não se destaca em relação as demais medidas e que a melhor métrica avaliada neste contexto foi 9 (correlação), sendo a melhor em 60% dos conjuntos de dados avaliados. Outra métrica que também conseguiu alcançar bons resultados nos conjuntos de dados avaliados foi 7 (distância de cosseno).

4.2. Segunda tentativa para escolha da medida de distância: Heurística

Em nossa segunda tentativa utilizamos uma heurística para nos apoiar na escolha da melhor métrica. Nesta heurística estamos avaliando o comportamento da medida de distância em relação a saída das instâncias. Assim, considerando inicialmente um comportamento linear, onde instâncias com grandes diferenças na saída também representariam grandes diferenças em termos da distância dos atributos que as descrevem utilizamos o Algoritmo 1. Este algoritmo já foi proposto inicialmente por [Filho 2023], mas foi aprimorado durante os testes aqui realizados.

Basicamente o procedimento realizado cria dois vetores ordenados. No primeiro é selecionada a instância de menor saída (linha 3) e verificada a distância entre essa

Table 5. Resultados Heurística

Dataset	Medida de distância										
	0	1	2	3	4	5	6	7	8	9	10
0	0.9706	0.9118	0.9328	0.9328	0.9454	0.895	**1**	0.9874	0	0.1597	0.9832
1	0.873	0.9312	0.9691	0.9916	0.9874	0.8947	**1**	0.9081	0	0.9705	0.0414
2	0.9368	0.2001	0.4018	0.5811	0.6741	0.5703	0.9098	0.9668	0	**1**	0.5952
3	0.833	0.3173	0.538	0.6416	0.6909	0.5756	0.8658	0.9362	0	0.8132	**1**
4	0.8362	0.855	**1**	0.9987	0.9773	0.9133	0.837	0.748	0	0.5726	0.4719
5	0.9207	0.4159	0.6557	0.7933	0.8427	0.5212	0.8197	**1**	0	0.9029	0.9197
6	0.8644	0.693	0.8081	0.8761	0.9013	0.8283	**1**	0.8795	0.6419	0.9006	0
7	0.9234	0.7772	0.9501	0.9643	0.9733	0.8859	**1**	0.918	0	0.4545	0.4831
8	0.7444	0.9068	0.9688	0.9545	0.9211	0.9684	**1**	0.8045	0.294	0.5624	0
9	0.5433	**1**	0.8463	0.737	0.685	0.6327	0.3007	0.1797	0.5454	0	0.1565

instância e todas as demais conforme a medida de distância avaliada. O segundo vetor é composto peles instâncias ordenadas simplesmente pelas saídas (linha 5). Por fim, um contador verifica a semelhança entre esses dois vetores, somando um cada posição verificada onde a instância correspondente não é encontrada no outro vetor (laço de repetição entre as linhas 7 e 17). Assim, quando menor for o valor final do contador (linha 18) mais semelhante serão os dois vetores e mais apropriada será a métrica para o conjunto de dados específico.

Algorithm 1 Heurística para escolha de medida de distância

Require: I (intâncias), Y (saídas)

1: Padroniza os dados utilizando o Z-Score
2: $Y_{ordenado}$ = ordena as saídas(Y)
3: I_{menor} = retorna a instância de menor saída($Y_{ordenado}$, I)
4: $Distanciass$ = calcula distâncias entre a instância de menor saída e todas as demais (I_{menor}, I)
5: $Distancias_{ordenado}$ = ordena as distâncias calculadas ($Distancias$)
6: $contador$ = 1
7: **for** cada p $\in Y_{ordenado}$ **do**
8: aux = 0
9: **for** cada p' $\in Distancias_{ordenado}$ **do**
10: **if** p == p' **then**
11: $contador$ += aux
12: Remove(p' in $Distancias_{ordenado}$)
13: **else**
14: aux += 1
15: **end if**
16: **end for**
17: **end for**
18: **return** $contador$

Os resultados obtidos são apresentados na tabela 5. Assim como na tentativa anterior os valores foram normalizados por MIN-MAX. Observe que neste caso destacamos os menores valores obtidos e assim temos como melhores resultados da medida 8 (Hamming), sendo superior em 70% dos conjuntos avaliados.

Table 6. Média da Correlação de Pearson entre os resultados obtidos pelos métodos para cada conjunto de dados

Conjunto de dados	Heurística/Variância	Heurística/KNN	Variância/KNN
0	0,19	0,85	0,55
1	0,56	0,82	0,51
2	0,53	0,39	0,08
3	0,32	0,83	0,15
4	0,35	0,74	0,39
5	0,50	0,81	0,51
6	0,37	0,87	0,15
7	0,43	0,96	0,43
8	0,34	0,84	0,28
9	0,19	0,57	0,60
Média dos valores absolutos	0,38	0,77	0,35
Desvio padrão	0,13	0,16	0,19

4.3. Análise e comparativo

Para comparar os resultados obtidos com as propostas para escolha de medida de distância, frete ao modelo KNN, calculamos o modulo do coeficiente de correlação de Pearson entre as alternativas. A tabela 6 apresenta os coeficientes da relação existente entre os métodos.

Como podemos observar, os resultados obtidos pela heurística avaliada obteve uma maior correlação em relação aos resultados obtidos pelo regressor KNN, sendo portanto a melhor alternativa dentre nossas tentativas de escolher a melhor medida de distância a ser utilizada. Os desvios encontrados foram semelhantes. Observe que esse resultado não pode ser generalizado para qualquer modelo regressor. Nossos testes obtiveram bons resultados para escolha específica do modelo utilizando KNN. Outras avaliações devem ser feitas a fim de constatar a pertinência do uso da heurística para diferentes modelos.

5. Conclusões e Trabalhos Futuros

Algoritmos de aprendizagem de máquinas são cada vez mais utilizados em nosso cotidiano e prover mecanismos capazes de aumentar a precisão dos resultados obtidos tem sido grande objetivo da comunidade científica.

Assim, este trabalho observou que a medida de distância utilizada também é capaz de afetar a qualidade de um modelo de regressão, podendo contribuir na obtenção de melhores resultados quando existe uma escolha cuidadosa. Propomos duas soluções para determinar a medida a ser utilizada dependendo do conjunto de dados. Na primeira avaliamos a variância calculada a partir das distancias entre as instancias, procurando maximizar este valor. Na segunda, utilizamos uma heurística que considera a linearidade entre as distâncias das instancias e as saídas obtidas por cada uma delas.

Como resultados observamos que a heurística da linearidade possui maior correlação com a saída do regressor avaliado (KNN) e, portanto é uma boa alternativa para escolha da medida de distância a ser utilizada.

Como trabalhos futuros, pretendemos avaliar a heurística em outros algoritmos de regressão para verificar se os resultados obtidos neste trabalho podem ser generalizados para outros modelos. Além disso, pretendemos melhorar a heurística utilizada a fim de capturar também a não linearidade entre as distâncias das instâncias em relação aos atributos e a saída do modelo.

References

[Aggarwal et al. 2001] Aggarwal, C. C., Hinneburg, A., and Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. In *International conference on database theory*, pages 420–434. Springer.

[Conde et al. 2007] Conde, G. A. B., Ádamo L. de Santana, Francês, C. R. L., Rocha, C. A., Rego, L., and Gato, V. (2007). Estratégias de previsão de carga e de consumo de energia elétrica baseadas em modelos estatísticos e redes neurais artificiais: Um estudo de caso nas concessionárias de energia do estado do pará. In *Anais do 8 Congresso Brasileiro de Redes Neurais*, pages 1–6, Florianópolis, SC. SBRN.

[Costa et al. 2004] Costa, I. G., de Carvalho, F. d. A., and de Souto, M. C. (2004). Comparative analysis of clustering methods for gene expression time course data. *Genetics and Molecular Biology*, 27:623–631.

[Dua and Graff 2017] Dua, D. and Graff, C. (2017). UCI machine learning repository.

[Filho 2023] Filho, R. M. (2023). *Explaining Regression Models Predictions*. PhD thesis, Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Programa de Pós-Graduação em Ciência da Computação, Belo Horizonte.

[Finkler 2017] Finkler, A. C. (2017). Aprendizagem de máquina aplicada à previsão dos movimentos do ibovespa.

[Giancarlo et al. 2010] Giancarlo, R., Lo Bosco, G., and Pinello, L. (2010). Distance functions, clustering algorithms and microarray data analysis. In *International Conference on Learning and Intelligent Optimization*, pages 125–138. Springer.

[Jaskowiak et al. 2013] Jaskowiak, P. A., Campello, R. J., and Costa, I. G. (2013). Proximity measures for clustering gene expression microarray data: a validation methodology and a comparative analysis. *IEEE/ACM transactions on computational biology and bioinformatics*, 10(4):845–857.

[Jaskowiak et al. 2014] Jaskowiak, P. A., Campello, R. J., and Costa, I. G. (2014). On the selection of appropriate distances for gene expression data clustering. In *BMC bioinformatics*, volume 15, pages 1–17. Springer.

[Monard and Baranauskas 2003] Monard, M. C. and Baranauskas, J. A. (2003). Conceitos sobre aprendizado de máquina. In *Sistemas Inteligentes Fundamentos e Aplicações*, pages 89–114. Manole Ltda, Barueri-SP, 1 edition.

[Rozin 2018] Rozin, N. A. (2018). Previsão do deslocamento de tempestades severas : abordagens por aprendizado de máquina.

[Silva 2022] Silva, J. J. d. (2022). *Uma comparação de técnicas de Aprendizado de Máquina para predição de evasão de estudantes no ensino público superior*. PhD thesis, Universidade de São Paulo.