Normalización

Minería de Datos

Zavala Roman Irvin Eduardo 1270771

¿Para qué sirve?

Por la naturaleza de la normalización, no solo se ponen atributos en una misma escala, también:

- Varios algoritmos funcionan mejor con datos normalizados
- La manipulación desde bases de datos se vuelve más eficiente
- Permite ver relaciones entre algunos atributos que sin normalizarlos tal vez no se hubieran visto
- Se minimiza la cantidad de datos redundantes

¿Cuando usarse y no usarse?

En Machine Learning la normalización no es algo que debe usarse estrictamente para todas las bases de datos, sólo en aquellas donde la diferencia de escala es suficiente como para afectar el rendimiento del modelo o cuando no se sabe la distribucion de los datos.

Elevation	Aspect	Slope	Horizontal_D	Vertical_Dist	Horizontal_D	Hillshade_9a	Hillshade_N	Hillshade_3p	Horizontal_Distance_To_Fire_Points
2596	51	3	258	0	510	221	232	148	6279
2590	56	2	212	-6	390	220	235	151	6225
2804	139	9	268	65	3180	234	238	135	6121
2785	155	18	242	118	3090	238	238	122	6211
2595	45	2	153	-1	391	220	234	150	6172
2579	132	6	300	-15	67	230	237	140	6031
2606	45	7	270	5	633	222	225	138	6256
2605	49	4	234	7	573	222	230	144	6228
2617	45	9	240	56	666	223	221	133	6244
2612	59	10	247	11	636	228	219	124	6230
2612	201	4	180	51	735	218	243	161	6222
2886	151	11	371	26	5253	234	240	136	4051
2742	134	22	150	69	3215	248	224	92	6091
2609	214	7	150	46	771	213	247	170	6211
2503	157	4	67	4	674	224	240	151	5600
2495	51	7	42	2	752	224	225	137	5576
2610	259	1	120	-1	607	216	239	161	6096
2517	72	7	85	6	595	228	227	133	5607
2504	0	4	95	5	691	214	232	156	5572

First Few Rows Of Original Data

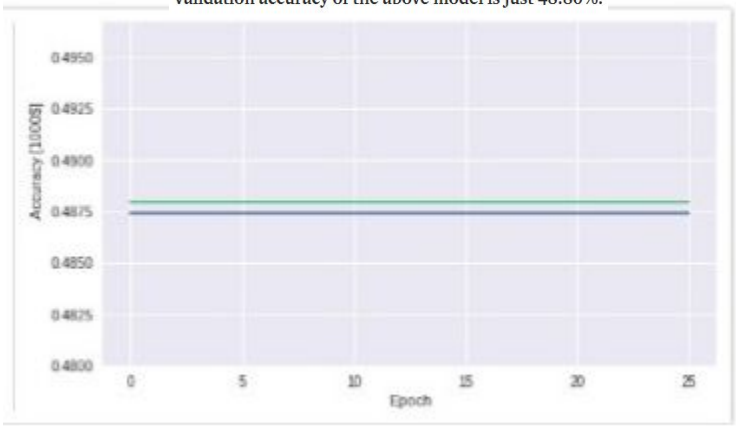
```
        Elevation    Aspect     Slope  Horizontal_Distance_To_Hydrology  \
152044   0.222366 -0.228639 -0.412503                          0.148486
363373   1.980490 -0.469989  0.255453                          3.018822
372733  -1.081933  0.271939  0.389044                         -0.867895
572846  -1.164122 -0.157128 -0.278912                         -1.267860
114145  -0.052787  0.861906  0.255453                         -0.279711

        Vertical_Distance_To_Hydrology  Horizontal_Distance_To_Roadways  \
152044                        0.149095                         1.336119
363373                        4.443372                         0.168073
372733                       -0.160093                        -0.241801
572846                       -0.795646                        -0.461170
114145                       -0.125739                         1.811419

        Hillshade_9am  Hillshade_Noon  Hillshade_3pm  \
152044       1.002687        0.539776      -0.510339
363373       1.227001       -0.270132      -1.190275
372733       0.292357        1.349684       0.378807
572846       0.965301        0.641014      -0.431885
114145      -1.090917        1.299065       1.581770

        Horizontal_Distance_To_Fire_Points     ...      Soil_Type32  \
152044                           -0.111226     ...                0
363373                           -0.703030     ...                0
372733                            0.038235     ...                0
572846                           -1.450334     ...                0
```

Output: Data after normalization

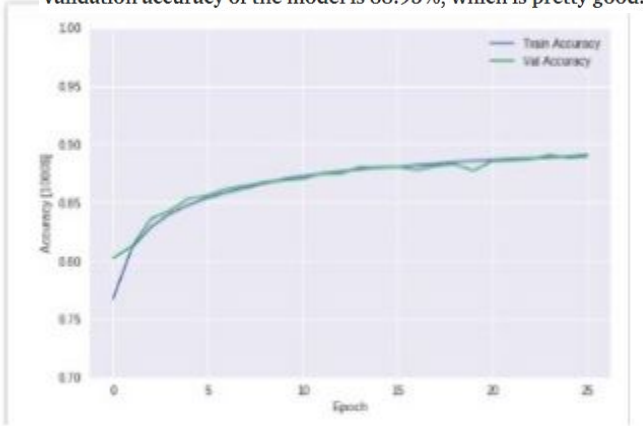
Ejemplo

Left: Model Accuracy, without normalized data
Right: Model Accuracy with normalized data

Validation accuracy of the above model is just 48.80%.



Validation accuracy of the model is 88.93%, which is pretty good.



The accuracy plot of the above 2 models

¿Qué tipos de normalización son más convenientes usar bajo qué circunstancias?

- Min-Max: Es bastante sencillo para implementar e interpretar pero solo escala de 0 a 1
- Decimal Scaling: Reduce números muy grandes a una escala de 0 a 1 de manera sencilla manteniendo la misma proporción, por lo que es fácil de entender
- Z-Score: Como utiliza el promedio y desviación, es útil cuando no se sabe el máximo y mínimo. Por como funciona la escala no es como la original

Observaciones

Al cambiar los datos originales a una escala nueva, dependiendo del método es necesario guardar ciertos parámetros para seguir guardando datos dentro de esa normalización.

Por ejemplo, en Z-Score se debe guardar la media y desviación estándar, en Min-Max se deben guardar los mínimos y máximos.

Bibliografía

GeeksforGeeks(2019). *Data normalization in data mining*. . https://www.geeksforgeeks.org/data-normalization-in-data-mining/

Galaktikasoft. (2019). *Data mining normalization*. Galaktikasoft. https://galaktika-soft.com/blog/data-mining-normalization.html

Jaitley, U. (2018). *Why Data Normalization is necessary for Machine Learning models*. Medium. https://medium.com/@urvashilluniya/why-data-normalization-is-necessary-for-machine-learning-models-681b65a05029

Shalabi, L. A., Shaaban, Z., & Kasasbeh, B. (2006). Data Mining: A Preprocessing Engine. *Journal of computer science*, *2*(9), 735–739. https://doi.org/10.3844/jcssp.2006.735.739

Sharma, R. (2020). *What is Normalization in data mining and how to do it?* UpGrad Blog; UpGrad. https://www.upgrad.com/blog/normalization-in-data-mining/

Tolety, K. (2022, mayo 27). Data Normalization Techniques in Data Mining simplified 101. *Learn | Hevo*. https://hevodata.com/learn/normalization-techniques-in-data-mining/