Big Data

Dealing with Noisy Data in Data Science

abril 29, 2019 Judith Chao Andrade 0 comentarios Noise data

Tabla de contenidos

Ankit RathiFollowSep 17, 2018

Recursos de interés

Programing & Informatica Think & Learn

Las matemáticas que nos dicen

Healthcare

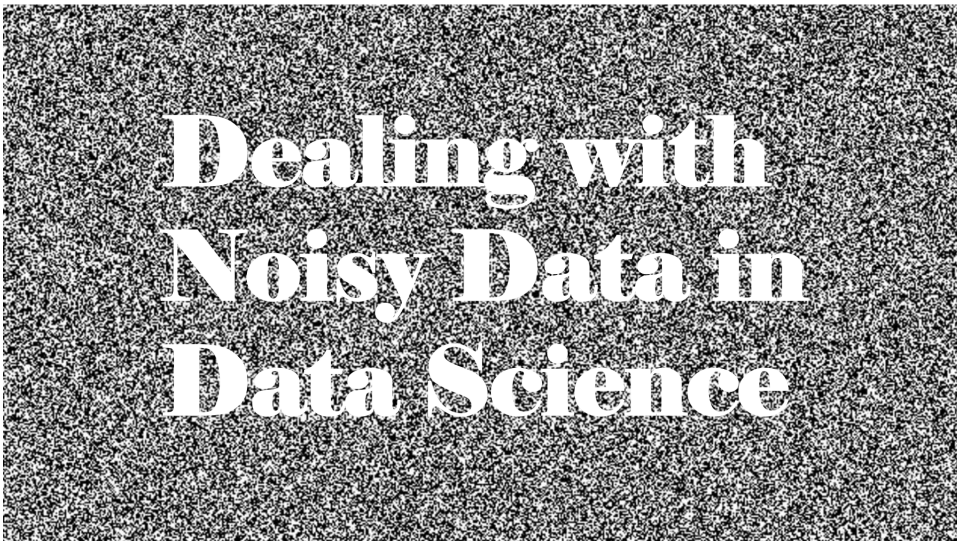
Provecto AI-ON:

Programing & Informatica Think & Learn

Los ordenadores están cada vez más presentes en nuestro día a día y nos ayudan a desarrollar numerosas tareas, como, por ejemplo, encontrar la ruta más corta a un destino

diciembre 20, 2021 Judith Chao Andrade 0

Photography Programing & Informatica



This article discusses the types of noise you encounter while working on data (tabular data) in data science projects and possible approaches you can take to deal with such noise. For detailed explanation of the methods mentioned in this post, please refer the links in 'Reference' section or explore yourself.

- *Noise in data*
- *Noise as an item (Noise1)*
- *Noise as a feature (Noise2)*
- *Noise as a record (Noise3)*
- *Unsupervised methods*

We were working on a dataset for our data science project, where we saw that our model was not performing up to the mark. While performance is a subjective term and there can be many reasons for an under-performing model, our hunch was that this is because of the noise in the dataset.

We tried many approaches to identify and reduce this noise. Some of them worked, and some of them didn't, because of the specific nature of the problem and the patterns in the data.

Based on my above experience, I am going to discuss various type of noise in data, and the approaches and methods to identify & reduce noise in a given dataset.

Understanding Noise in Data

Noise (in the data science space) is unwanted data items, features or records which don't help in explaining the feature itself, or the relationship between feature & target. Noise often causes the algorithms to miss out patterns in the data.

Un algoritmo de aprendizaje automático logra que Grand Theft Auto se vea

Curiosity Sin categoría

'Racecraft', el arte de las tecnológicas de culpar a la raza

Big Data

Curiosity Video

La película más antigua de la historia, remasterizada en 4K gracias a la inteligencia artificial

Healthcare

La 'start-up' que traza un mapa de tu corazón

octubre 4, 2020 Judith Chao Andrade 0

Conect with me at Linkedin

Judith Chao Andrade

Nombre (obligatorio)

Correo electrónico (obligatorio)

Mensaje

Noise in tabular data can be of three types:

1. Anomalies in certain data items (*Noise 1*: certain anomalies in features & target)
2. Features that don't help in explaining the target (*Noise 2*: irrelevant/weak features)
3. Records which don't follow the form or relation which rest of the records do (*Noise 3*: noisy records)

Index	Feature1	Feature2	Feature3	Feature..	Feature..	Feature..	Feature..	FeatureM-1	FeatureM	Target	
Record1											
Record2											
Record3											Noise1
Record4											Noise2
Record5											Noise3
Record6											
Record7											
Record..											
Record..											
Record..											
Record..											
RecordN-1											
RecordN											

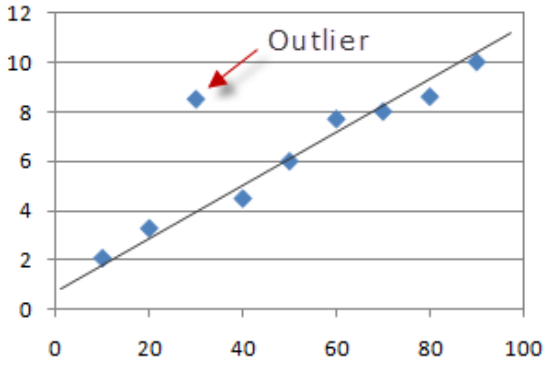
Benefits of identifying & treating noise in data:

- enables the DS algorithm to train faster.
- reduces the complexity of a model and makes it easier to interpret
- improves the accuracy of a model if the right subset is chosen
- reduces overfitting

These are the ways of dealing noise within data based on the type of noise:

Noise as an item

We can analyse the features & target and identify the noise in terms of outliers.



Outlier detection & treatment: either remove the records or put upper and lower ceiling.

Noise as a feature

Contáctanos

TAGS

AI algoritm Art Artificial Inteligence Blockchain cancer Color Data Science Dataset Diagnose Digital Healthcare Digitization fact-check Formats GIS Google healthack Health Data History of cinema IA Image analysis Infección desease Innovation Machine learning medical image medical records Medical Technique Medicine for all mHealth New medicine developments Open data Pharmacology Portal del Paciente Preserve Productividad Python Research resistance antibiotics Resources Search engines Social Media Technical Photography Twitter Vaccines X-ray

This type of noise is introduced when there are features in the data which are not related to target or doesn't help explaining target.

Feature Selection or Elimination

Not all features are important, so we can use various methods to find the best subset of features:

Filter method

We can perform various statistical tests between feature & response to identify which features are more relevant than others.

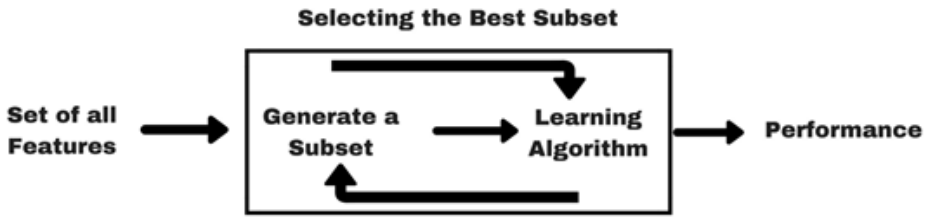
Feature\Response	Continuous	Categorical
Continuous	Pearson's Correlation	LDA
Categorical	Anova	Chi-Square

Please note that above methods don't identify or deal with multicollinearity, we need to figure that out separately.

Wrapper method

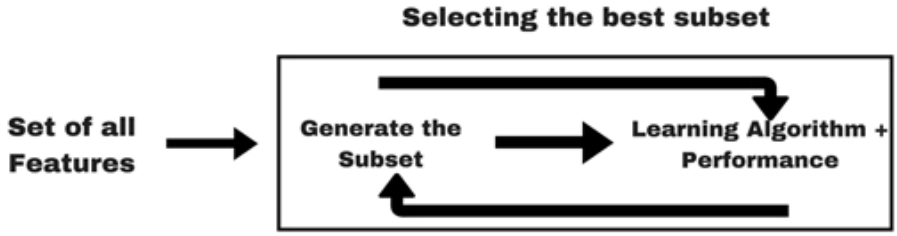
Here we add/remove features to baseline model and compare the performance of the model:

- Forward selection
- Backward elimination
- Recursive elimination



Embedded Methods (Regularization)

This method make use of filter & wrapper method, it is implemented using algos which have its own built-in feature selection methods.

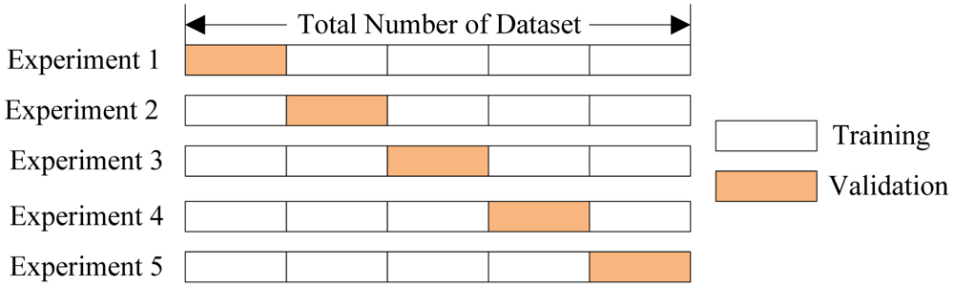


Noise as a record

In these methods, we can try to find the set of records which have noise.

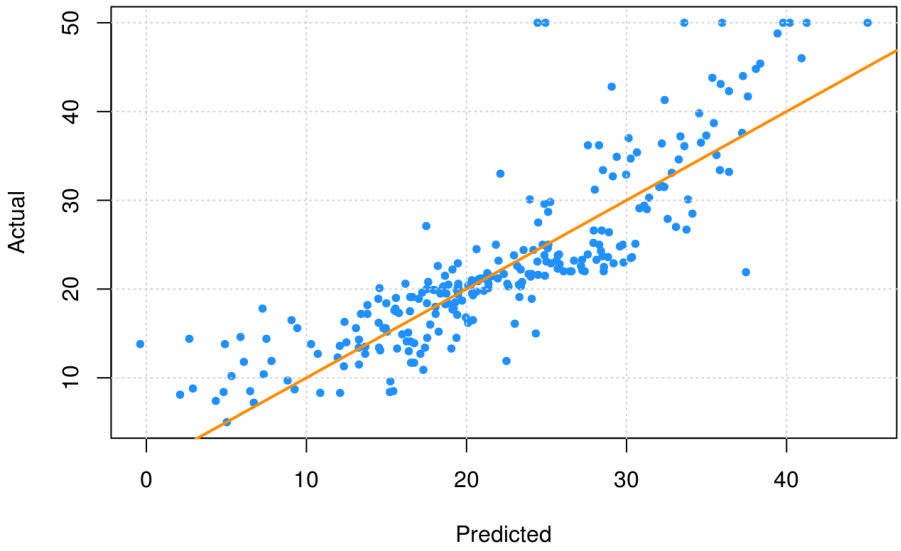
K-fold validation

In this method, we can look at the cross validation score of each fold and analyse the folds which have poor CV scores, what are the common attributes of records having poor scores, etc.



Manual method

Here we can evaluate CV of each record (predicted vs. actual) and filter/analyse the records having a poor CV score. This will help us in analyzing why this is happening in the first place.



Unsupervised Methods (Anomaly Detection)

We can also use unsupervised learning algorithms to identify anomalies in data, these are mostly categorized as Anomaly Detection techniques.

Density-based anomaly detection

This method assumes normal data points occur around a dense neighborhood and abnormalities are far away. *i.e. kNN & LOF based methods*

Clustering-based anomaly detection

Using clustering technique, we can analyse the clusters to analyse which has noise. Data instances falling outside the clusters can be marked as anomalies. *i.e. k-Means clustering*

SVM-based anomaly detection

This technique uses SVM to learn the soft boundary in the training set and tune on validation set to identify anomalies. In this approach, the need of large samples by the previous approach is reduced by using Support Vector Machine while maintaining the high quality of clustering-based anomaly detection methods. *i.e. One-class SVM*

Autoencoder-based anomaly detection

Auto-encoders are used in deep learning for unsupervised learning, we can use them for anomaly detection to identify noisy data-set. These methods are advanced and outperforms traditional anomaly detection methods. *i.e. Variational Autoencoder based Anomaly Detection using Reconstruction Probability.*

Conclusion

Not every method mentioned above suits in every situation or problem. We need to analyse what kind of noise we have in our data, and try corresponding methods to remove or minimize it. In our project some of methods we tried & worked based on the specific patterns in our data-set.

References

Introduction to Feature Selection methods with an example (or how to select the right variables?)
Introduction One of the best ways I use to learn machine learning, is by benchmarking myself against the best data…
www.analyticsvidhya.com**Introduction to Anomaly Detection**
Experience with the specific topic: Novice Professional experience: No industry experience This overview is intended…
www.datascience.com**Anomaly detection using Support Vector Machine classification with k-Medoids clustering – IEEE…**
*Anomaly based Intrusion Detection System, in the recent years, has become more dependent on learning methods …*ieeexplore.ieee.org

Couldn't preview file
You may be offline or with limited connectivity. Try downloading instead.

Thank you for reading my post. I regularly write about Data & Technology on LinkedIn & Medium. If you would like to read my future posts, then simply 'Connect' or 'Follow'. Also feel free to listen to me on SoundCloud.

Source: Analytic Vidhaya

← The Library of Congress Lets You Stream Hundreds of Free Films

Conserve the sound, an online Museum Preserves the Sound of Past Technologies-from Typewriters, Electric Shavers and Cassette Recorders, to cameras & Classic Nintendo →

Judith Chao Andrade

Apasionada del conocimiento, de compartirlo y de aprender de todo lo que me rodea, disfruto aprendiendo y realizando actividades. Actualmente estoy aprendiendo programación pero me fascinan los temas relacionados con los materiales especiales, las cuiriosidades, el humor, los eventos, las redes sociales ... Mi mayor interés podría decir que es no perder nunca la cuiriosidad por lo que si tienes un plan en mente solo proponlo !.

Deja una respuesta

Tu dirección de correo electrónico no será publicada. Los campos obligatorios están marcados con *

Comentario *

Nombre *

Correo electrónico *

Web

Guarda mi nombre, correo electrónico y web en este navegador para la próxima vez que comente.

Recibir un correo electrónico con los siguientes comentarios a esta entrada.

Recibir un correo electrónico con cada nueva entrada.

Publicar el comentario