Dealing With Noisy Data in Data Science

Tejas Vijayprakash Desai

M.Sc. in Information Technology, Keraleeya Samajam's Model College, Maharashtra, India.

---***---

Abstract:

As the word itself tells a noisy data are data with a large amount of additional meaningless information in it. In general noise is a random error or variance which may include faulty data collection instruments, technology limitations, resource limitation and data entry problems. Normally we work on specific datasets for our data science project, where we apply a particular model and check whether the model is performing up to the mark or not. If any kind of noises are available then we try to identify and reduce the noisy data. But during this process some of the things work and some don't due to patterns and specific nature in the datasets. Mostly we have to use more data in order to clear or reduce the noisy ratio in the available data. Sometimes this noisiness in data can also provide some irrelevant features. In this study I have summarized some types of noise in data and the methods and approaches to identify and reduce the noise in the data. I have also discussed the type of noise can an individual encounter while working on data in data science projects and possible approaches can take to deal with such data.

Information

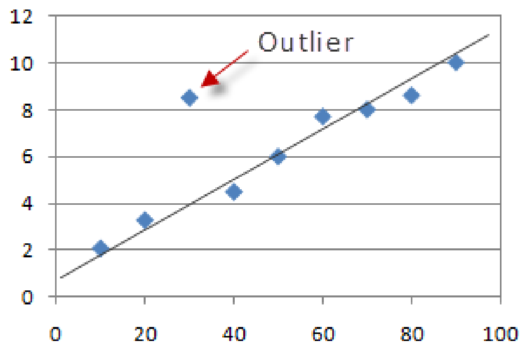
Understanding Noise in Data:

Normally noise is unwanted data items, features and records which don't help the relationship between feature and targets or in explaining the feature. Due to the noise algorithms often miss out data patterns. I have segmented noises into mainly three following categorize with relation to the data, they are

- **Noise as an item :-** i.e. anomalies in certain data items (certain anomalies in features and target).
- **Noises as a feature :-** i.e. features that don't help in explaining the target (irrelevant/weak feature).
- **Noise as a record :-** i.e. records which don't follow the form or relation which rest of the record or datasets do (noisy records).

Noise as an item

It is a type in which the noise can be analyzed as an item to its uncertain feature or a target. This includes outlier data detection & treatment i.e. data shown in the datasets are extremely different than the other remained data in the data sets. This kind of data or records either removed or put to upper or lower ceiling (related to ceiling effect in the mathematics). For example following diagram has shown the outlier data or record.



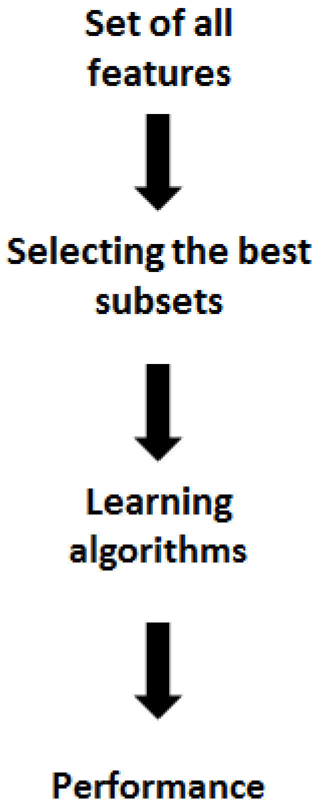
In general the noise as an item is the data or particular records in the datasets which are uncertain or very different than the remaining data.

Noise as a feature

In this kind of noise the data shows some feature which is not intended i.e. the noise as a feature introduced when the data shows some features (generally 'bugs' from the programming world) which are not related to target or doesn't help in explaining the target. In this type of noise the user can either select the feature which is raised or eliminate it.

For feature selection or elimination we can perform various tests to identify which feature is of further use, which feature is more relevant than the others. I have personally created criteria for selecting the feature, the criteria is as follows:

Feature Selection or Elimination:



There may be condition in which not all features are important, so to get the best feature there is various methods an individual can use.

Filter method

An individual can perform this or various other statistical tests methods between feature and response to identify the relevant or most useful feature which are different than others. The filter method don't identify Multicollinearity

Feature/Response	Continuous	Categorical
Continuous	**Some Data**	**Some Data**
Categorical	**Some Data**	**Some Data**

Wrapper Method

In this particular method an individual can compare the performance of the model by adding/removing features.

We can perform three operations using this method.

- ➢ Forward Selection

In this particular selection an individual can start with a null model and after then can start fitting the model with each individual feature at a single time and then select that feature or any feature with minimum p-value. Then model should fit with two features by trying combinations of the earlier selected feature with all other remaining feature. After this again select the feature with the minimum p-value but this time fit the model with the three features by trying the combination of previously selected two combinations.

Steps for forward selection:

1. Choose a Significant Level (SL).
2. By considering one feature fit the possible simple regression model. And select the feature with lowest p-value.
3. Fit all model with one extra feature added to the previously selected features.

   Again, select the feature with the minimum p-value. If p-value < SL then go to third step otherwise terminate the process

- ➢ Backward Elimination

In backward elimination an individual can start with the full model and then remove the insignificant feature with the highest p-value (i.e. p-value> SL). This is redundant process until user gets final set of significant features.

Steps for backward elimination:

1. Choose a Significant Level.
2. Including all the features fit the full model.
3. Consider the feature with the highest p-value.
4. If the p-value > SL then go through step 4 again otherwise terminate the process.
5. Remove a feature which is under consideration
6. Fit a model without this feature.

- ➢ Bi-directional Elimination

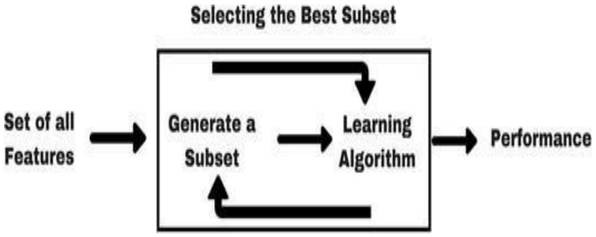
It is similar to forward selection but in this process while adding a new feature it also checks the significance of already added features.

Steps for backward elimination:

1. Choose a significant level to enter and exit the model.
2. Perform the next step of forward selection (newly added feature must have p-value < SL).
3. Perform all steps of backward elimination (any previously added feature with p-value > SL)

The following diagram shows the operation that we can perform by using wrapper method.

The following diagram shows the operation that we can perform.

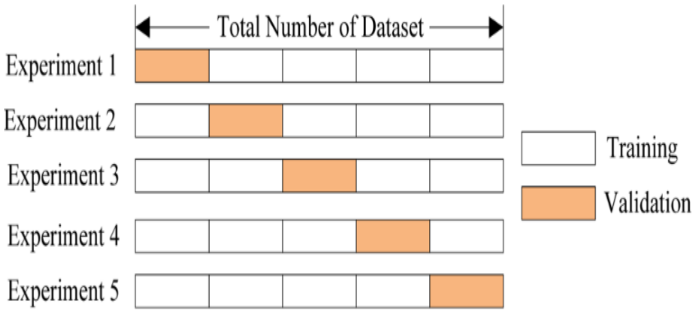


Noise as a record

In this particular type a user can find the set of records or data which have noise. For finding noisy records K-fold validation method can be used.

K-fold validation

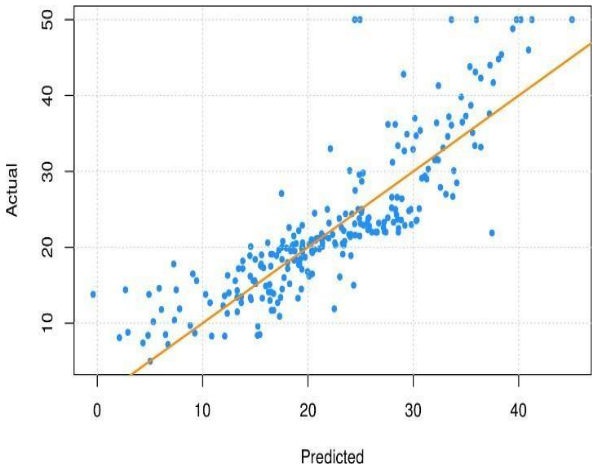
By using this method an individual can find records by taking the cross validation score (i.e. CV score) of each fold and then analyze those folds which have poor CV score. An user can also find common attributes of records having poor scores.



Manual Method

By using this method an individual can evaluate cross validation of each predicted and actual records and then further filter them for separating records which having poor cross validation score.

Let's take an example by using following diagram.



The diagram shows which records are nearby the prediction and which are not. A user can further analyze these dots (records) and select the feature accordingly.

Unsupervised Method (Anomaly Detection):

To identify anomalies in the data an individual can also use unsupervised learning algorithms, these algorithms are also known as Anomaly Detection Techniques.

Density-based anomaly detection

It assumes normal data points which are raised around dense abnormalities are far away. Some methods like kNN & LOF based methods.

Clustering based anomaly detection

A user can analyze the data through various clusters in the clustering. The instances or particular datasets falling outside the clusters can be summarized as anomalies. The most common clustering method is k-mean clustering method.

SVM based anomaly detection

In this method the detection process includes SVM i.e. Support Vector Machine to learn soft boundary in training set and tune on validations sets to identify anomalies. Sometimes it also reduces usage of large samples from previous cluster based anomaly detection method. The common method in this approach is One-class SVM.

Auto-encoder based anomaly detection

This method is advanced and outperforms all other old anomaly detection methods. This method mostly used for unsupervised learning in deep learning techniques. The common method is Variational Auto-encoder based Anomaly Detection using Reconstruction Probability.

Benefits of identifying and reducing noisy data:

- It makes it easier to interpret by reducing complexity of a model.
- By choosing right subset an individual can improve accuracy of their model.
- It reduces over fitting by reducing unwanted or false noisy data.
- It also enables any data science model to train faster.
- It also helps in the data cleaning by maintaining the accuracy of the data.

Conclusion

I have mentioned various methods and types above, but not every method suits in every problem or conditions. Users have to first analyze what kind of noise they have in their data and try to remove it. If an individual determines or finds the noise in the data then they can try to either minimize or remove it by corresponding these methods. A user has to try these methods once & check whether these methods work on the specific pattern in their records.

References

- **Ananlyticsvidhya.com/**Introduction to feature selection methods with an example
- **Ananlyticsvidhya.com/**Introduction to anomaly detection
- **en.wikipedia.org**
- **www.youtube.com/Data Science Dojo**
- **ieeexplore.ieee.org**
- **www.datascience.com**

AUTHOR

Name: Tejas Vijayprakash
Desai B.Sc. (computer
science)
Pursuing M.Sc. (information technology)