offers data science lesson videos made simple! Sign up or log in to Magoosh Data Science.

Magoosh Data Science Blog

Everything you need to know about Data Science

Search...

How to manage noisy data

BY AMAN GOEL ON APRIL 26, 2018 IN DATA ANALYSIS

Noisy data is a data that has relatively signal-to-noise ratio. When data is collected, humans tend to make mistakes and instruments tend to be inaccurate, so the collected data has some error bound to it. This error is referred to as noise.

Noise creates trouble for machine learning algorithms because if not trained properly, algorithms can think of noise to be a pattern and can start generalizing from it, which of course is undesirable. We ideally want the algorithm to make sense of the data and generalize the underlying properties of the data. This is prevented by noise which tends to "fool" the algorithm to making incorrect generalizations. Therefore, it becomes important for any data scientist to take care of noise when applying any machine learning algorithm over a noisy data.

In order to manage noisy data, here are some techniques that are extensively used:

Collecting more data

The simplest way to handle noisy data is to collect more data. The more data you collect, the better will you be able to identify the underlying phenomenon that is generating the data. This will eventually help in reducing the effect of noise. Think about it – when survey companies conduct surveys, they do it on a mass scale. This is because a handful of survey responses might not be good for generalizing because humans tend to be moody and so, some may answer the survey negatively because of a possibly bad mood (noisy data). This may not reflect the actual behavior of the masses unless the survey is conducted on a really large scale.

As a rule of thumb – the larger the sample size, the better will you be able to uncover the actual behavior of the population.

PCA

The Principal Component Analysis (PCA) is a method of the family of the data analysis and more usually multivariate statistics, which consists of transforming linked variables (called "correlated" into statistics) into new variables that are uncorrelated from each other. These new variables are called "principal components", or main axes. It allows the practitioner to reduce the number of variables and make the information less redundant.

It is an approach that is both geometric (the variables are represented in a new space, according to maximum inertia directions) and statistical (the research on independent axes explaining at best the variability – the variance – Datas). When you want to compress a set of random variables, the main axes of principal component analysis are a better choice, from the point of view of inertia or variance.

PCA effectively reduces the dimension of the input data by projecting it along various axes. For instance, consider projecting a point in X-Y plane along X-axis. This way, we are able to remove the (possibly) noisy dimension – Y-Axis. This exercise is also referred to as "dimensionality reduction". PCA is therefore widely used to reduce noise from data by "forgetting" the axes that contain the noisy data.

Regularization

The core of a Machine Learning algorithm is the ability to learn and generalize from the dataset that the algorithm has seen. However, if the algorithm is given enough flexibility (more parameters), then it may happen that the algorithm "overfits" the noisy data. This means that the algorithm is fooled into believing that the noise part of data also represents a pattern. In order to avoid that, one commonly used technique is called as Regularization. In regularization, a penalty term is added to the algorithm's cost function, which represents the size of the weights (parameters) of the algorithm. This ensures that for the minimization of the cost, the weights are smaller thereby leading to lesser freedom for the algorithm. This greatly helps in avoiding overfitting. There are 2 commonly used techniques in regularization:

- **L1 regularization:** In L1 regularization, a term of $|w_i|$ is added for each *i*. The modulus function is always positive and so, the regularization term leads to an increase in the cost function.
- **L2 regularization:** In L2 regularization, a term of w_i^2 is added. Since square is a positive function, so here also the regularization term leads to an increase in the cost function.

In order to minimize the cost, the optimizer tries to take lower values of the weights, thereby leading to less flexibility of the algorithm. This avoids overfitting which in turn helps in handling noisy data easily.

Cross Validation

Cross-validation is a technique that helps in tackling with noisy data by preventing overfitting. This is just like overfitting. In cross-validation, the dataset is broken into 3 sets (rather than 2):

- Training data
- Cross validation data
- Testing data

The algorithm is trained using the training data. However, the hyper-parameters are tuned using the cross-validation data which is separate from the training data. This makes sure that the algorithm is able to avoid learning the noise present in the training data and rather generalize by a cross-validation procedure. Finally, the fresh, test data can be used to evaluate how well the algorithm was able to generalize.

It is important for all data scientists to understand the impact the noise can create on the data and so, every data scientist must take appropriate measures to design algorithms accordingly. This way, the generalizing capabilities of the algorithm on new data will be far better.

Related Posts

What Is Perceptron?

Introduction to Machine Learning

7 Best Books on Artificial Intelligence

5 Data Mining Techniques You Should Know About

5 Most Helpful Data Governance Tools

Open Source Languages

About Aman Goel

Aman Goel is a Computer Science and Engineering graduate from IIT Bombay. He secured AIR 33 in JEE Advanced 2013! He is now the co-founder of AllinCall Research & Solutions. Aman loves to write motivational articles to help students perform well in JEE and to understand the principles of Data Science.

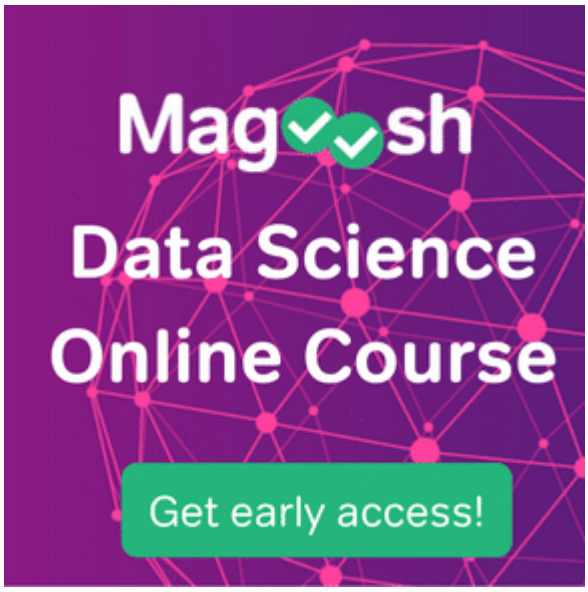
‹ Previous Next ›

Comments are closed.

Magoosh blog comment policy: To create the best experience for our readers, we will only approve comments that are relevant to the article, general enough to be helpful to other students, concise, and well-written! 😄 Due to the high volume of comments across all of our blogs, we cannot promise that all comments will receive responses from our instructors.

We highly encourage students to help each other out and respond to other students' comments if you can!

If you are a Premium Magoosh student and would like more personalized service from our instructors, you can use the Help tab on the Magoosh dashboard. Thanks!



Recent Posts

- What Is a Confusion Matrix?
- What Is Perceptron?
- What Is a Bayesian Network?
- 5 Machine Learning Algorithms to Understand
- What Is a Regression Model?

View our privacy policy.

© 2021 Magoosh Data Science Blog. All Rights Reserved.