feature_engineering, machine_learning, dimensionality, pca

Dimensionality Reduction is good or bad

Steve **Jul '15**

Whenever we have more number of features, we apply dimensionality reduction techniques like PCA, forward and backward elimiation and others.

Here by reducing the number of features or creating a new vector, we are losing information about data or generate a new variable which does not has any practical meaning.

Do you think that it is better to apply these techniques? Is there any other method, which can solve issues to deal with more number of features?

Regards,
Steve

---

karthe1 **Jul '15**
Awesome Hackers!

@Steve - In large datasets, very often the variables are highly correlated. The very purpose of applying dimension reduction techniques like PCA is to identify those highly correlated variables or variables which are not related to the target variable and drop them out from further analysis.

Having highly correlated variables leads to overfitting and hence the accuracy of the model will suffer. So it is good to do the dimension reduction wherever possible.

---

hinduja1234 **Jul '15**

@Steve
Dimensionality reduction helps in
Visualization: projection of high-dimensional data onto 2D or 3D.
Data compression: efficient storage and retrieval.
Noise removal: positive effect on query .

So it is good to use dimensionality reduction technique than any other techniques to deal with more number of features.

Regards,
Rohit

---

kunal **Jul '15**

@Steve

There is no global right or wrong answer. It depends on situation to situation. Let me cover a few examples to bring out the spectrum of possibilities. These may not be comprehensive, but would give you a good flavor of what should be done.

Example 1: You are building a regression with 10 variables, some of them are mildly correlated.
In this case, there is no need to perform any dimensionality reduction. Since there is little correlation among the variables, all of them are bringing in new information. You should keep all the variables in the mix and build a model.

Example 2: Again, you are building a regression model. This time with 20 variables and some of these variables would be highly correlated.
For example, in case of telecom - number of calls made by a customer in a month and the monthly bill he / she receives. Or in case of insurance, it could be number of policies and total premium sold by an agent / branch. In these cases, because you only have limited number of variables you should only add one of these variables to your model. There is limited / no value you might get by bringing in all the variables. And a lot of this additional value can be noise. You can still do away with applying any formal dimensional reduction technique in this case (even though you are doing it for all practical reasons).

Example 3: Now assume that you have 500 such variables with some of them being correlated with each other.
For example, data output of sensors from a smartphone. Or a retail chain looking at the performance of a store manager with tons of similar variables - like total number of SKUs sold, number of bills created, number of customers sold to, number of days present, time spent on the aisle etc. etc. All of these would be correlated and are way too many to individually figure out which are correlated to which ones. In this case, you should definitely apply dimension reduction techniques. You may or may not make actual sense of these vectors, but you can still understand them by looking at the result.
In a lot of scenarios like this, you will also see that you will retain more than 90% of information with less that 15 - 20% of variables. Hence, these can be good applications of dimensionality reduction.

Hope these examples help.

Kunal

---

karthiv **Jul '15**

Hi sir,

Nice Explain.