High Dimensional Data

Cynthia Kineza

2017-11-09

High dimensional data occurs when there is more predictors / features than number of observations. The majority of datasets coming out of modern biological techniques are high-dimensional.

Examples of high dimensional data

1. Predict risk of diabetes on the basis of DNA sequence data
2. Identify regions of interest associated with brain activation during a memory task ...

This brings out statistical challenges. Too many predictors will overfit the data (overfitting occurs when a regression model is tailored to fit a particular set of data but fails to fit additional data or predict future observations reliably).

We must care about model's performance on observations not used to fit the model (for example we may want to predict the length of stay of a new individual who walks into the emergency room given his symptoms).

One could ask why the number of variables matters. Well, when the number of individuals is less than the predictors you can always get a perfect model fit to the training data but the test error will be pretty bad.

Dimensionality matters since classical statistical techniques cannot be applied. Furthermore, there are risks of overfitting, false positives and more.

Some of the techniques to estimate the test error would be

The validation set approach: Split data into two equal sets, train on one set and evaluate performance on the other.

Leave-one-out cross-validation: Fit N models, each on n-1 of the observations. Then evaluate each model on the left-out observation

K-fold cross-validation: Split data into 5 sets, train numerous times a model on 4 sets and evaluate its performance on the 5th.

For Big data (high dimensional), we cannot usually perform least squares regression to fit a model because we will get zero training error however we will get a very bad test error.

Instead we will use methods such as Variable pre-selection, Principal components regression, Ridge regression, or Lasso Regression which will fit less complex models. Those are some of the alternatives to least squares.

For neuroimaging data analysis, due to high dimensionality, I use methods mentioned above to predict regions of interest associated with tasks given individuals demographic features.

There are a lot of things that I glossed over in this post, this was just an introduction about supervised and unsupervised statistical machine learning, I will go more in depth in the future, in the meantime feel free to send me an email if you have suggestions or questions.

Thanks for those who have reached out and gave me feedbacks, I will be blogging more frequently in the future!