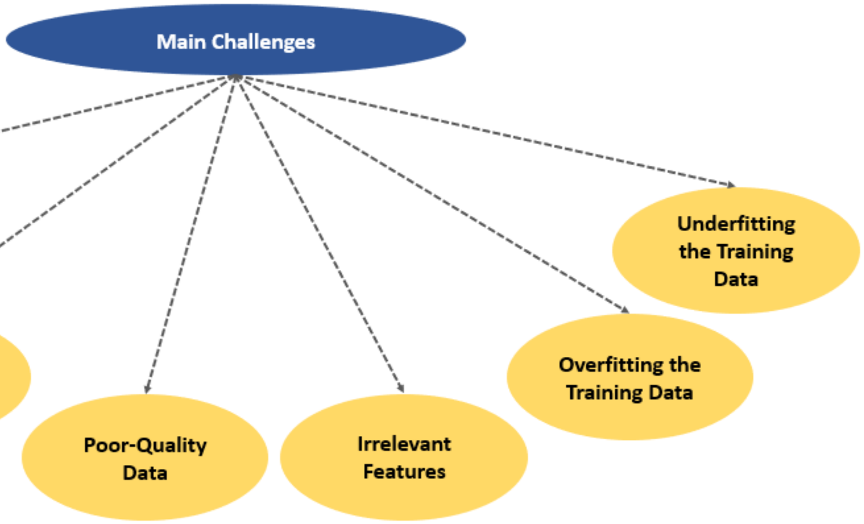
Main Challenges

Friday, December 6, 2019 6:20 PM

Note Written by: Zulfadli Zainal

In Machine Learning: Only focus on 2 problems! (Bad Data or Bad Algorithm)


Main Challenges
Insufficient Quantity of Training Data
Non Representative Training Data
Poor-Quality Data
Irrelevant Features
Overfitting the Training Data
Underfitting the Training Data

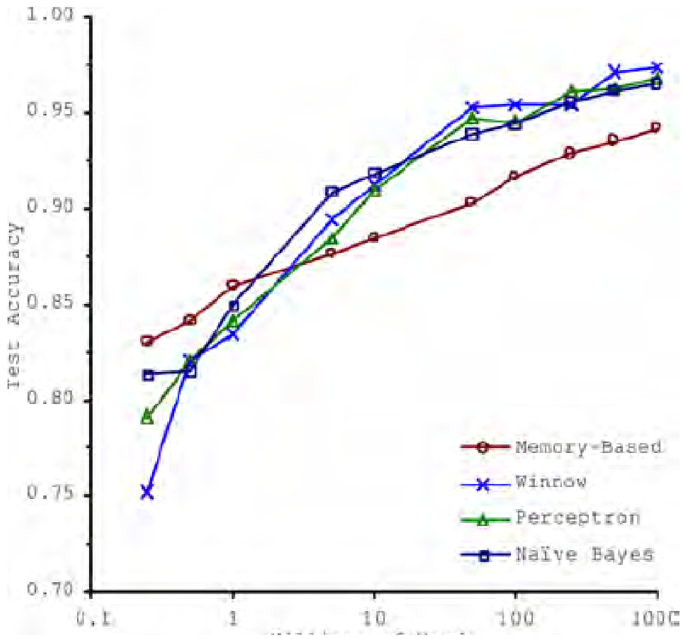
Insufficient Quantity of Training Data

Sometimes, with enough data, you don't need a complex algorithm.

In 2001, Microsoft did a study with sufficient enough data -> even simple machine learning algorithm can perform similarly good with complex machine learning algorithm.

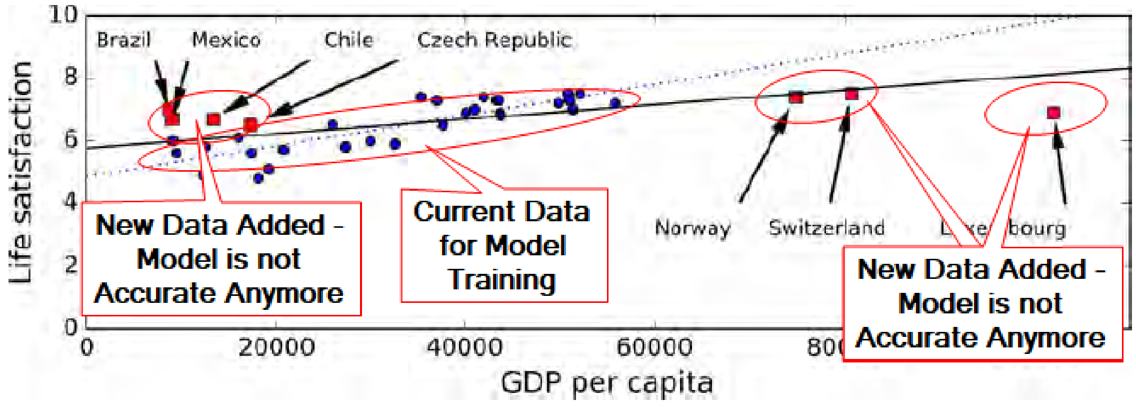
We may want to reconsider the tradeoff between spending time and money on algorithm development versus spending it on corpus development.

Data Volume vs Algorithm Efficiency



Nonrepresentation Training Data

Sometimes the data not represent what we want to seek. This is Bias!



New Data Added - Model is not Accurate Anymore

Current Data for Model Training

New Data Added - Model is not Accurate Anymore

Poor Quality Data

Obviously, if your training data is full of errors, outliers, and noise (e.g., due to poor quality measurements), it will make it harder for the system to detect the underlying patterns, so your system is less likely to perform well.

Irrelevant Features

As the saying goes: garbage in, garbage out.

Your system will only be capable of learning if the training data contains enough relevant features and not too many irrelevant ones.

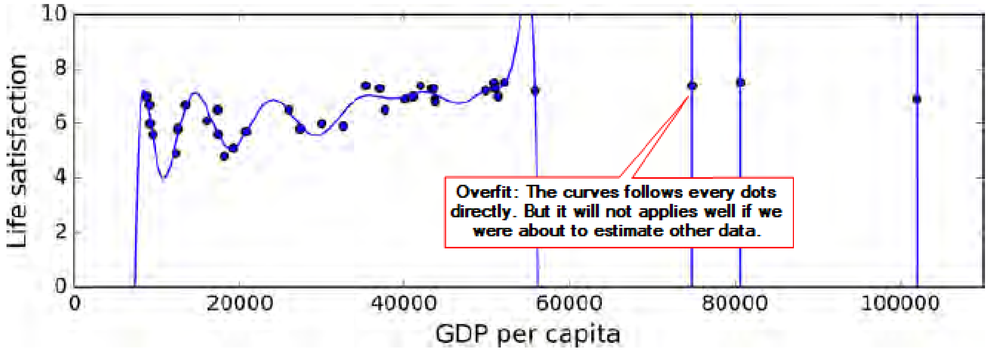
To come with a good features:

1. Feature selection: selecting the most useful features to train on among existing features.
2. Feature extraction: combining existing features to produce a more useful one (as we saw earlier, dimensionality reduction algorithms can help).
3. Creating new features by gathering new data.

Overfitting the Training Data

Overfitting -> Simply means over generalizing.

In Machine Learning this is called overfitting: it means that the model performs well on the training data, but it does not generalize well.

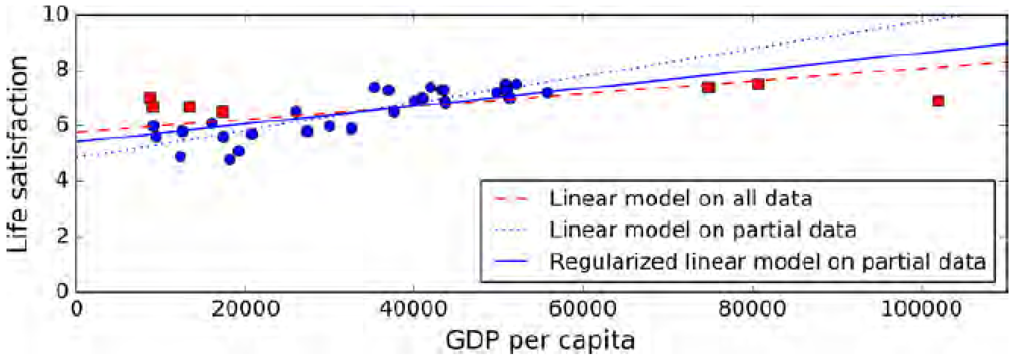


Overfit: The curves follows every dots directly. But it will not applies well if we were about to estimate other data.

Overfitting happens when the model is too complex relative to the amount and noisiness of the training data. The possible solutions are:

• To simplify the model by selecting one with fewer parameters (e.g., a linear model rather than a high-degree polynomial model), by reducing the number of attributes in the training data or by constraining the model
• To gather more training data
• To reduce the noise in the training data (e.g., fix data errors and remove outliers)

Constraining a model to make it simpler and reduce the risk of overfitting is called regularization.



Regularization reduce the risk of overfitting!

Underfitting the Training Data

Opposite of overfitting.

When it occurs? When model is too simple.

How to fix the problem:

- Selecting a more powerful model, with more parameters
- Feeding better features to the learning algorithm (feature engineering)
- Reducing the constraints on the model (e.g., reducing the regularization hyper parameter)