Generalization

TDT17 - Sept. 10th, 2019

Erik Liodden, Simen Ullern and Vebjørn Fjeldberg

Table of Contents

1. Generalization: what is it, and why?
2. Overfitting and underfitting
3. How to measure generalization
4. How to generalize your model
    - Preparing your dataset
    - Regularization
5. Summary

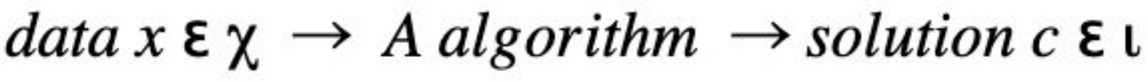
1. What is generalization?

Generalization refers to your model's ability to adapt properly to new, previously unseen data, drawn from the same distribution as the one used to create the model.

https://developers.google.com/machine-learning/crash-course/generalization

Why is this difficult? The fundamental problem of ML

Stochastic input data X ⇒ stochastic output as well



$$data\ x\ \varepsilon\ \chi \rightarrow A\ algorithm \rightarrow solution\ c\ \varepsilon\ \iota$$

What parts of the input is relevant? Can it be found with statistical induction techniques using only a subset of the data?

Algorithms that create algorithms need some unknown optimal structure function

Rudolf Carnap

The problem of induction poses the question: “What justifies us in going from the direct observation of facts to a law that expresses certain regularities of nature?”

One approach to achieve generalization

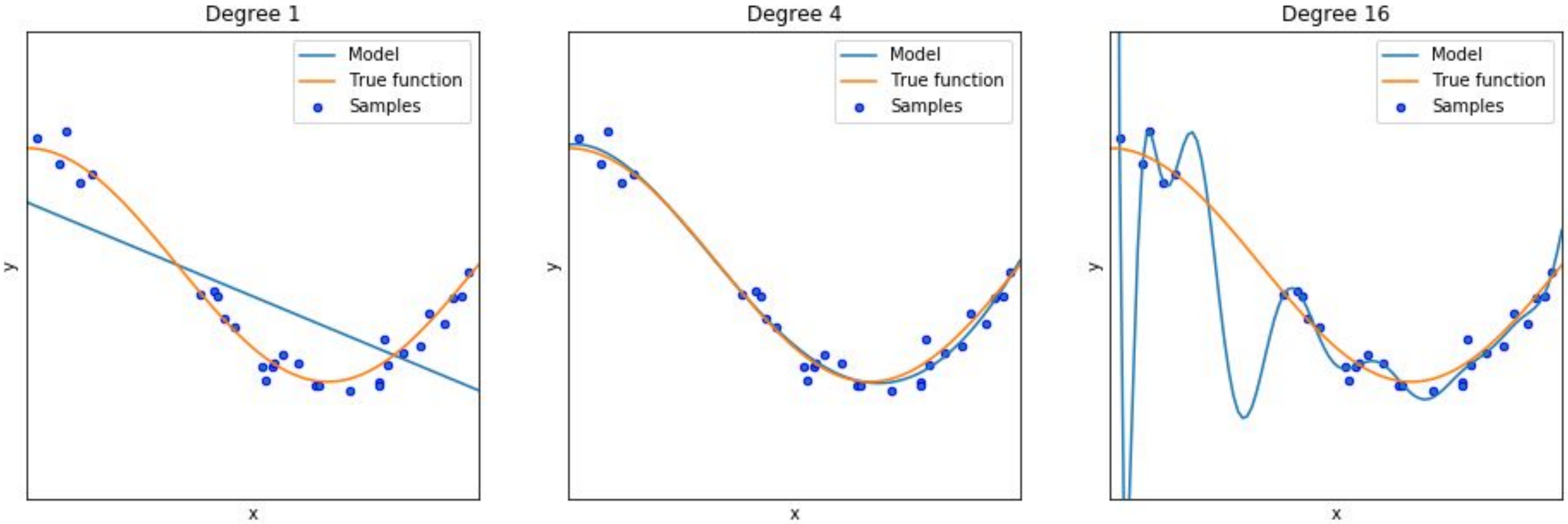
- Partition the dataset into training, validation and testing.
- The model is trained on the training set
- The validation set is tested upon regularly during training and used to tune the hyperparameters of the model
- Test set is only used at the end to give an indication about how well the model is generalizing

What happens if you peek on test data?

Test data cannot be used before the final estimator has been selected!

If we use test data for validation, then estimator adaptation introduces statistical dependencies between outcome of the learning process (estimator) and test data. This design flaw yields a too optimistic estimate of the test error.

2. Overfitting and underfitting



Occam's razor: when presented with competing hypotheses that make the same predictions, one should select the solution with the fewest assumptions.

Distinguish between expected and empirical error

Empirical error can easily be measured on some test set

$$I_S[f_n] = \frac{1}{n}\sum_{i=1}^{n} V(f_n(x_i), y_i)$$

But **the expected error or risk is the quality measure which we ultimately care about** because it is calculated over all possible values

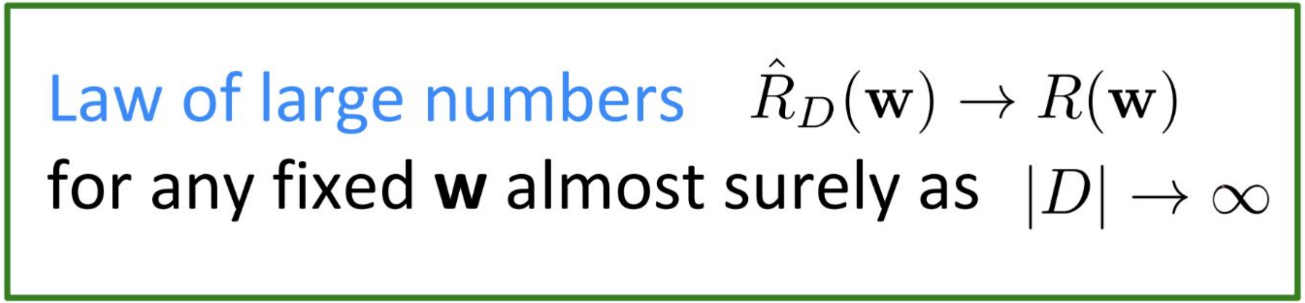
$$I[f_n] = \int_{X\times Y} V(f_n(x), y)\rho(x, y)dxdy,$$

where V denotes a loss function and p(x,y) is the unknown probability distribution

Estimating the generalization error

Estimate the true error by the empirical error on a sample data set D

Why might this work?



Law of large numbers $\hat{R}_D(\mathbf{w}) \rightarrow R(\mathbf{w})$

for any fixed **w** almost surely as $|D| \rightarrow \infty$

Generalization in general

The generalization error is the difference between the expected and empirical error.

An algorithm is said to generalize if the generalization error approaches 0 as the number of data points goes to infinity.

How to generalize your model

preparing your dataset and regularization

Preparing your dataset

Attribute sampling: assume which attributes are critical and which are going to add more dimensions and complexity to your dataset without any predictive contribution.

Record sampling: remove less representative records to make prediction more accurate.

Data normalization: change values of attributes in the dataset to common scale, without distorting differences in the ranges of values

Data augmentation: Increase the diversity of the dataset used for training by generating new data. In image classification, this can be done by rotating and mirroring images and adding noise etc.

Regularization techniques

Dropout: a technique used in DL where randomly selected neurons are ignored during training

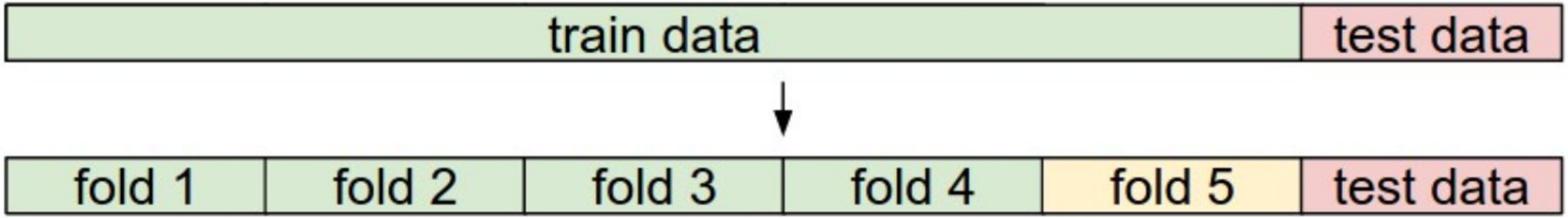
L2:

$$Min(\sum_{i=1}^{n}(y_i - w_i x_i)^2 + \boxed{p\sum_{i=1}^{n}(w_i)^2)}$$

L1:

$$Min(\sum_{i=1}^{n}(y_i - w_i x_i)^2 + \boxed{p\sum_{i=1}^{n}|w_i|)}$$

K-fold cross-validation



http://cs231n.github.io/classification

Useful to shuffle around the validation data. This usually helps the model to generalize better.

Split the training data into different folds and alternate which fold is used as validation set.

No free lunch theorem

The no free lunch theorem states that there is no one model that works best for every problem. The assumptions of a great model for one problem may not hold for another problem.

One must always assess trade-offs between speed, accuracy and complexity in practical ML problems.

Summary

Minimize error, maximize generality

Prepare your dataset properly and regularize to avoid too optimistic estimates

Partition your dataset into a training set and testing set. A validation set can be used to guide estimator selection. Estimate the true error by the empirical error on your test set

Unless you have limited computational power, you have no excuse not to do cross validation

Be aware of the tradeoffs that you either explicitly or implicitly have to make