**Overfiting**

In statistics, overfitting is the production of an analysis that corresponds too closely or exactly to a particular set of data, and may therefore fail to fit additional data or predict future observations reliablyChart

Description automatically generated

An overfitted model is a statistical model that contains more parameters than can be justified by the data

The essence of overfitting is to have unknowingly extracted some of the residual variation as if that variation represented underlying model structure

In other words, the model remembers a huge number of examples instead of learning to notice features.

The possibility of over-fitting exists because the criterion used for selecting the model is not the same as the criterion used to judge the suitability of a model.

**For example:**

if the number of parameters is the same as or greater than the number of observations, then a model can perfectly predict the training data simply by memorizing the data in its entirety. (For an illustration, see Figure 2.) Such a model, though, will typically fail severely when making predictions.

Chart, line chart

Description automatically generated

To lessen the chance of, or amount of, overfitting, several techniques are available

* model comparison
* cross-validation,
* regularization,
* early stopping,
* pruning,
* Bayesian priors,
* dropout.

**Consequences**

The most obvious consequence of overfitting is poor performance on the validation dataset. Other negative consequences include:

* A function that is overfitted is likely to request more information about each item in the validation dataset than does the optimal function; gathering this additional unneeded data can be expensive or error-prone, especially if each individual piece of information must be gathered by human observation and manual data-entry.
* A more complex, overfitted function is likely to be less portable than a simple one. At one extreme, a one-variable linear regression is so portable that, if necessary, it could even be done by hand. At the other extreme are models that can be reproduced only by exactly duplicating the original modeler's entire setup, making reuse or scientific reproduction difficult.

### Remedy

* minimum spanning tree
* life-time of correlation

a correlation matrix can be created by calculating a coefficient of correlation between investigated variables. This matrix can be represented topologically as a complex network where direct and indirect influences between variables are visualized.