Machine Learning Course - CS-433

Overfitting

Oct 3, 2017

minor changes by Martin Jaggi 2016

small changes by Rüdiger Urbanke 2017

Last updated on: September 25, 2017

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Motivation

Most models can be *too limited* or they can be *too rich*. In the first case we are likely to underfit and in the second to overfit. We will discuss this second concept now by looking at linear models.

Can Linear Models Overfit? At first it might seem that linear models are too simple to ever overfit. But in fact, linear models are highly prone to overfitting, much more so than complicated models like neural nets. The aim of this lecture is to understand why.
Consider linear regression. To keep things simple assume that the input x_n is one-dimensional. In order to have any representational power with linear models we typically "augment" the input. E.g., if the input (feature) is one-dimensional we might add a polynomial basis (of arbitrary degree M),

$$\boldsymbol{\phi}(x_n) = [1,\, x_n,\, x_n^2,\, x_n^3,\, \ldots,\, x_n^M]$$

so that we end up with an extended feature vector.
We then fit a linear model to this extended feature vector $\boldsymbol{\phi}(x_n)$:

$$y_n \approx w_0 + w_1 x_n + w_2 x_n^2 + \ldots + w_M x_n^M = \boldsymbol{\phi}(x_n)^\top \mathbf{w}.$$

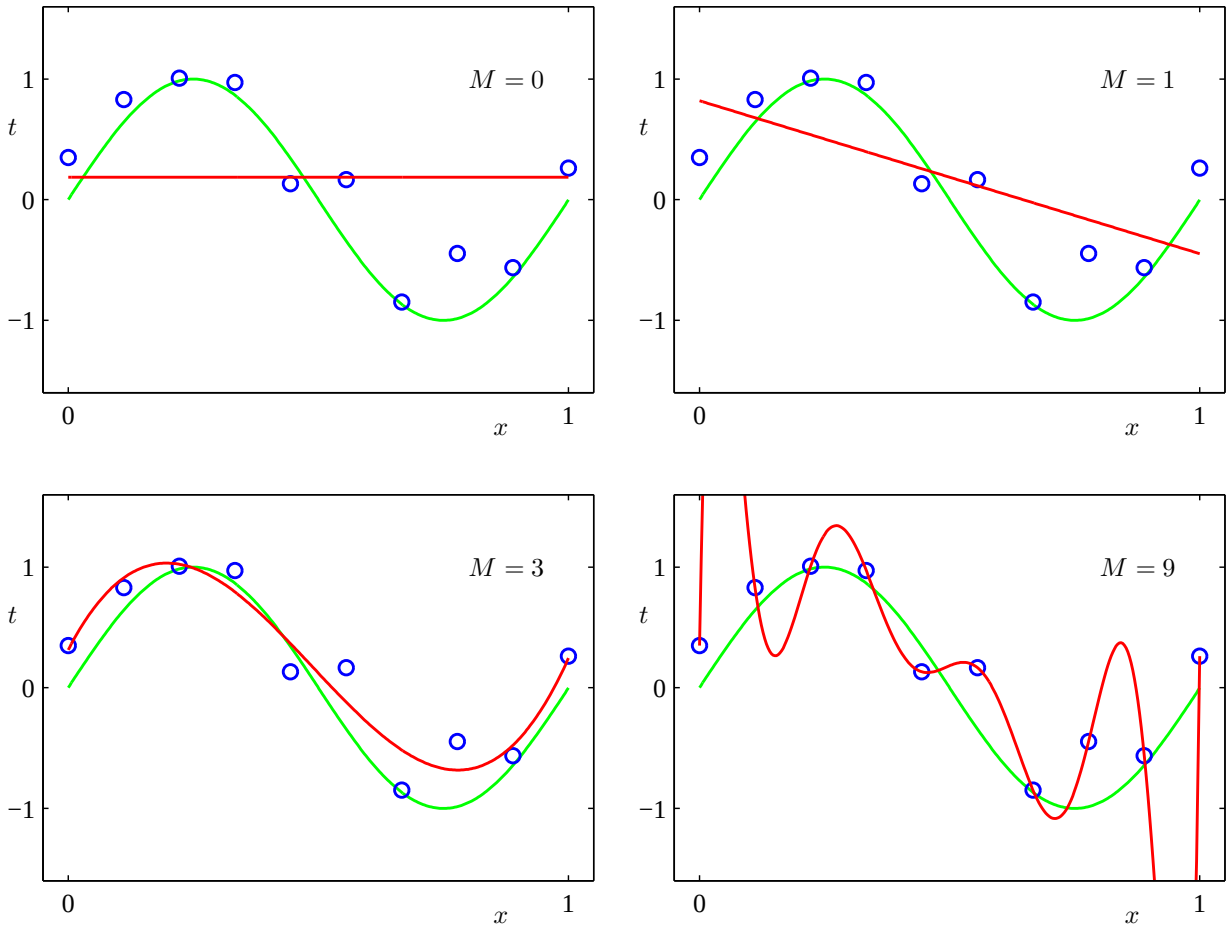
Overfitting and Underfitting

We say that a model overfits if it does not only fit the signal but also fits the noise. We speak of underfitting if the model does not fit the signal well. If we are given the data we of

course do not know what part is the signal and what part is the noise.

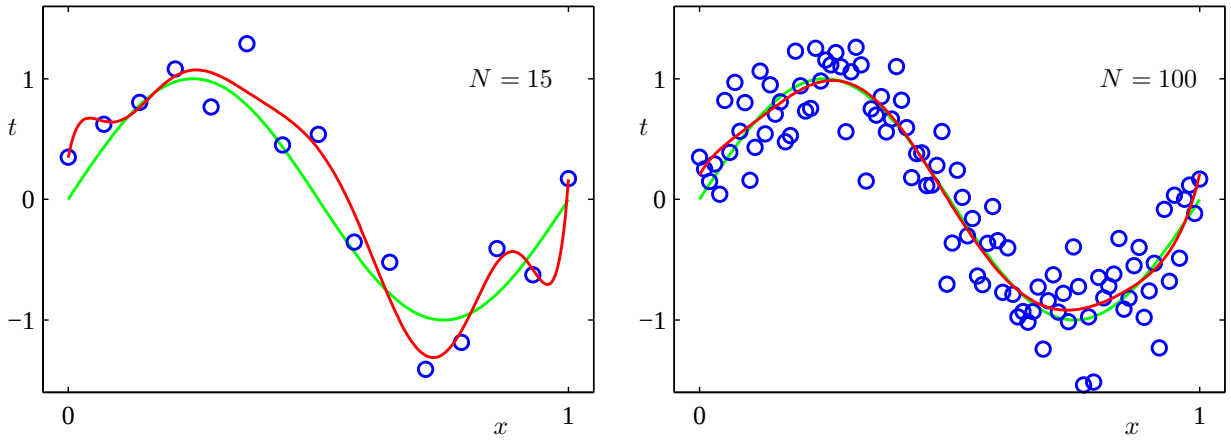
Complex Models Overfit Easily

In the following four figures, circles are data points, the green line represents the "true function", and the red line is the model. The parameter M is the maximum degree in the polynomial basis.



For $M = 0$ (the model is a constant) the model is underfitting and the same is true for $M = 1$. For $M = 3$ the model fits the data fairly well and is not yet so rich as to fit in addition the small "wiggles" caused by the noise. But for

$M = 9$ we now have such a rich model that it can fit every single data point and we see severe overfitting taking place. What can we do to avoid overfitting? If you increase the amount of data (increase N, but keep M fixed), overfitting *might* reduce. This is shown in the following two figures where we have $M = 15$ and even $M = 100$ but we have a signficant amount of extra data.



Occam's Razor

A second approach to avoid overfitting is to apply some sort of regularization. Such a regularization forces the model to be not too complex.
One solution is dictated by Occam's razor which states that "Plurality is not to be posited without necessity" or rephrased "Simpler models are better - only use complicated ones if strictly necessary". So when unsure, choose a simple model over a complicated one.
We can choose simpler models by adding a regularization term which "penalizes" complex models. E.g., consider the

cost function

$$\min_{\mathbf{w}} \quad \frac{1}{2N} \sum_{n=1}^{N} \left[y_n - \boldsymbol{\phi}(\mathbf{x}_n)^\top \mathbf{w} \right]^2 \; + \; \lambda \|\mathbf{w}\|_2^2.$$

By picking λ larger and larger we emphasize models with smaller and smaller weights.

ToDo

Read about overfitting in the paper by Pedro Domingos (Sections 3 and 5 of "A few useful things to know about machine learning").