Introduction to Smoothing

Machine Learning - Section 3.2.1

Marc Omar Haddad

Published: 25 February, 2020

Updated: 09 March, 2020

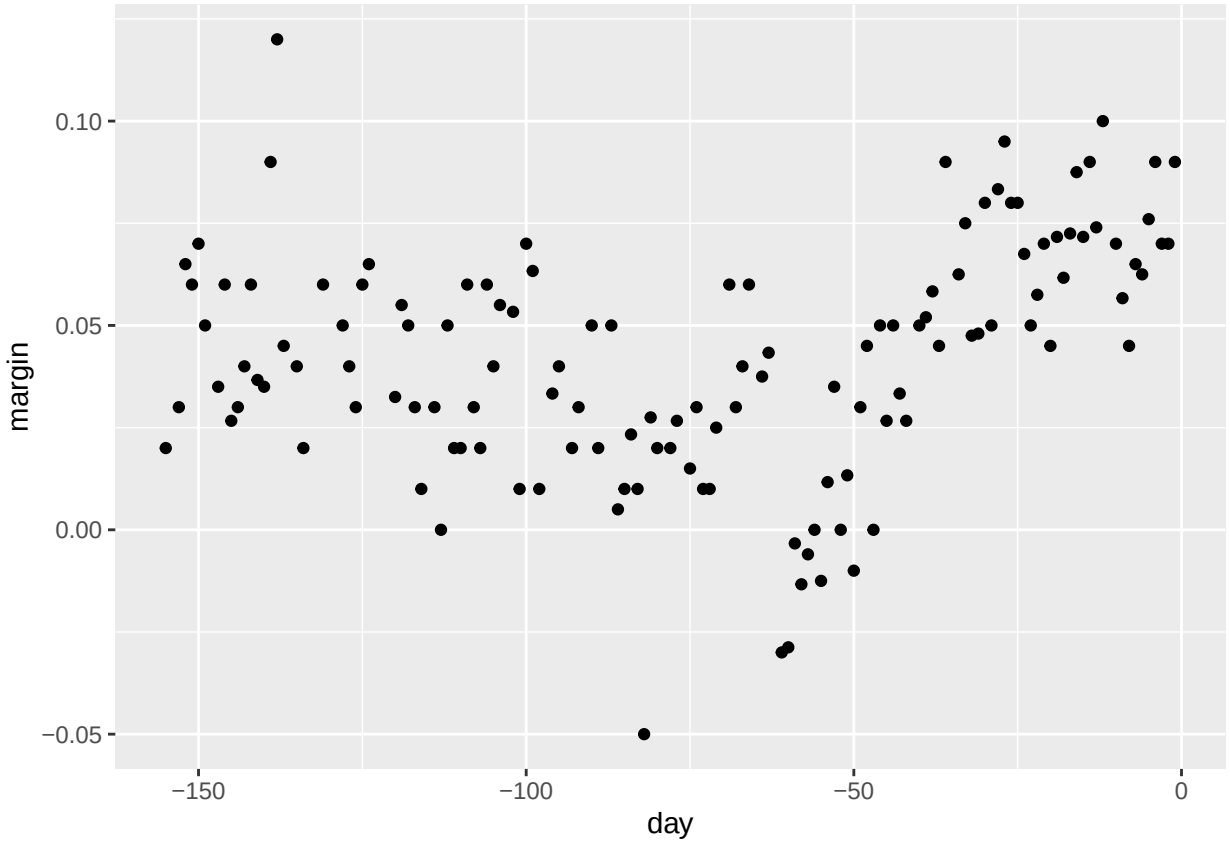
Smoothing is an extremely powerful tool that is used accross data analysis fields. Also known as "Curve Fitting" and "Low Band Pass Filtering", **Smoothing** is designed to **detect trends** in the presence of **noisy data** (where the shape of the trend is unknown). The term "Smoothing" is used to indicate that **we assume the trend of noisy data to be smooth**.

Smoothing techniques are quite useful when applied in a machine learning context. In machine learning, like in Smoothing, our goal is to **extract conditional expectations/probabilities** (the underlying trends) of an unknown shape, in the presence of **uncertainty** (noisy data).

Introduction to our Example Problem: Popularity Over Time

We will be trying to estimate the **time trend of the popularity vote** in the 2008 Presidential election between Obama and McCain.

```
data("polls_2008")
qplot(day, margin, data = polls_2008)
```



Note: **Do not think of this example as a forecasting problem**. All we are interested in is **learning the shape of the trend *after* collecting all the data**.

We assume that **for every day** x**,** we have a **true preference among the electorate,** $f(x)$. However, due to the uncertainty of polling, each data point comes with an **error** ϵ**.**

The following is the mathematical model for the observed poll margin Y:

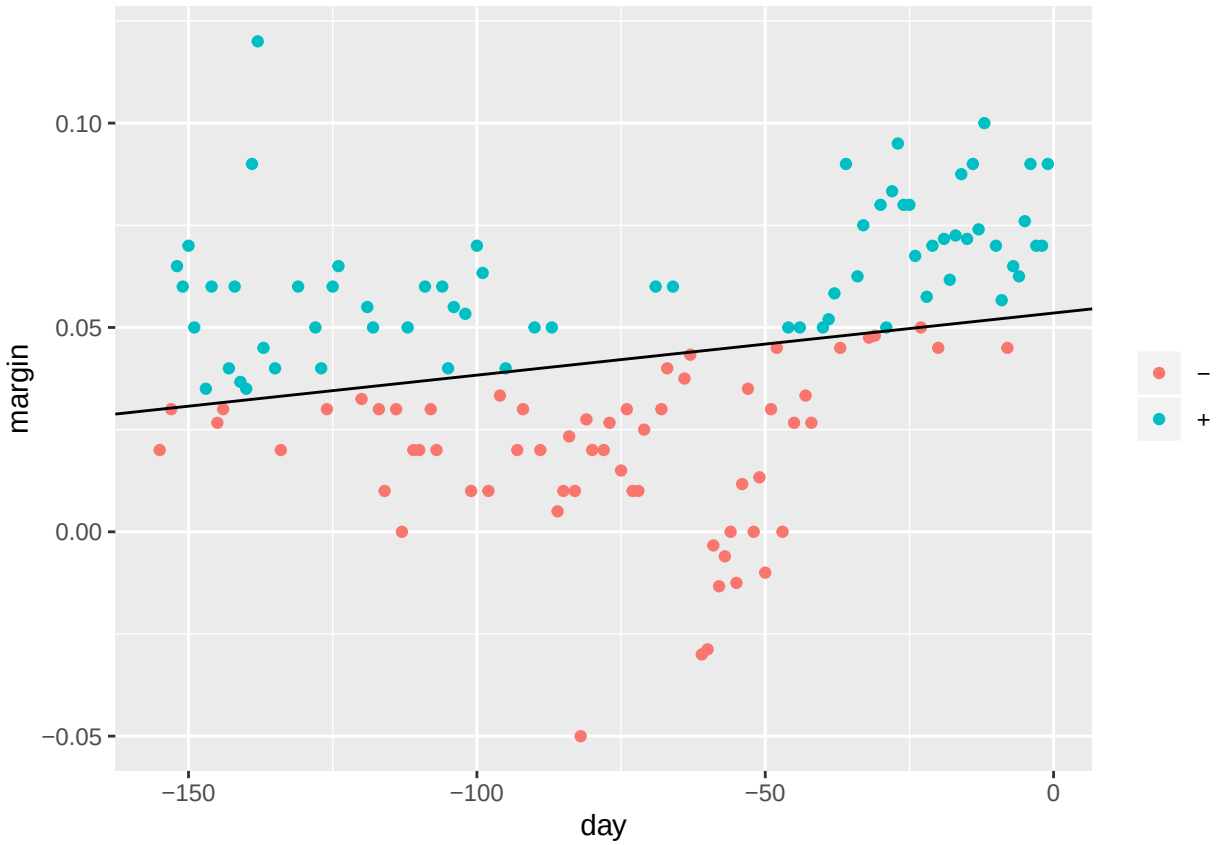
$$Y_i = f(x_i) + \epsilon_i$$

We can redefine this problem in a machine learning context: We want to **predict** Y **given day** x**, through the conditional expectation of**:

$$f(x) = \mathrm{E}(Y \mid X = x)$$

Since we *don't know* the true value of the above, we must **estimate it**.

We will first use regression for our estimates.



As we can see in our above plot, our regression line does not do a great job of describing our trend. We will need a more flexible algorithm.