Generalized Additive Models

Generalized Additive Models

- So far we have seen a number of approaches for flexibly predicting a response y on the basis of a single predictor x. These approaches may be seen as extensions of simple linear regression.
- Here we explore the problem of flexibly predicting y on the basis of several predictors, $x_1, \ldots, x_p$.
- Generalized additive models (GAMs) provide a general framework for extending a standard linear model by allowing non-linear functions of each of the variables, while maintaining additivity.
- The beauty of GAMs is that we can use splines and local regression as building blocks for fitting an additive model

Generalized Additive Models

- A natural way of extending the multiple linear regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

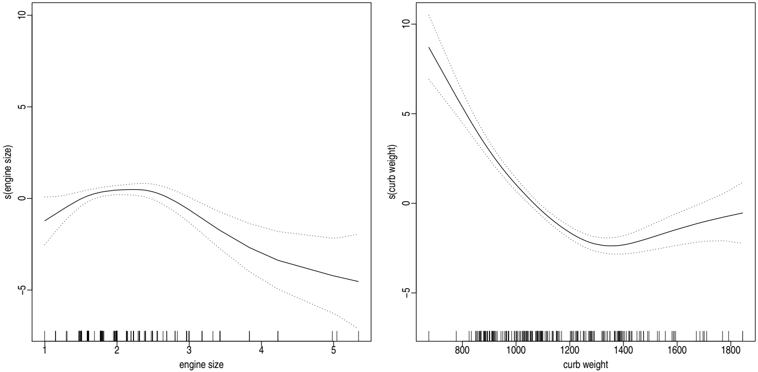
  in order to allow for nonlinear relationships between each feature and the response is to replace each linear component $\beta_j x_{ij}$ with a smooth nonlinear function $f_j(x_{ij})$.
- We can then write

$$y_i = \beta_0 + \sum_{i=j}^{p} f_j(x_{ij}) + \varepsilon_i = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \varepsilon_i$$

- this is a Generalized Additive Model (GAM).
- It is called additive because we calculate a separate f_j for each x_j and then add together all of their contributions.

Example

Estimate of city distance according to engine size and curb weight by an additive model with a spline smoother



Generalized Additive Models

GAM important properties:

- GAMs allow us to fit a non-linear f_j to each x_j, so that we can automatically model non-linear relationships that standard linear regression will miss. This means that we do not need to manually try out many different transformations on each variable individually.
- The non-linear fits can potentially make more accurate predictions for the response y.
- Because the model is additive, we can still examine the effect of each x_j on y individually while holding all of the other variables fixed.
- If we are interested in inference, GAMs provide a useful representation.

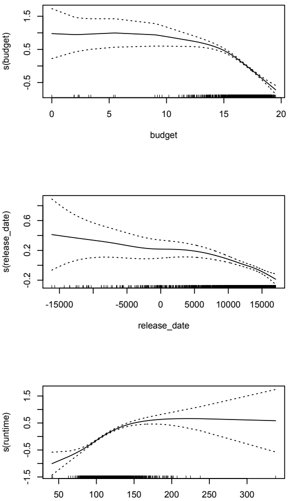
Example

Rating movies

- We consider the case of a dataset about movies
- We are interested in understanding the variable 'average vote' obtained by movies
- We want to study the relationship with other variables such as 'budget', 'popularity', 'revenues', 'runtime'

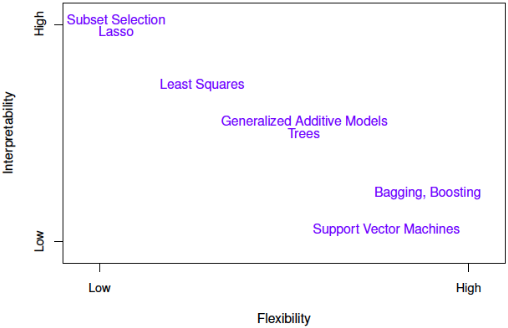
Generalized Additive Models

we may appreciate some results obtained with GAM



Flexibility vs Interpretability of models

There is a trade-off between flexibility and interpretability



Example

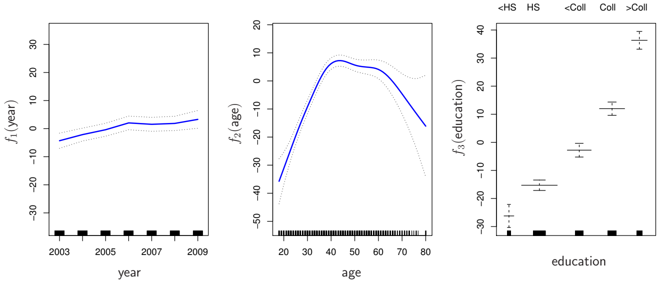
Let us consider the task of fitting the model

$$\text{wage} = \beta_0 + f_1(\text{year}) + f_2(\text{age}) + f_3(\text{education}) + \varepsilon$$

- year and age are quantitative variables
- education is qualitative with five levels: <HS, HS, <Coll, Coll, >Coll, referring to the amount of high school or college education that an individual has completed.
- we fit a model where f_1 and f_2 are smoothing splines

Example

Wage data



- holding age and education fixed, wage tends to increase slightly with year
- holding education and year fixed, wage tends to be highest for intermediate values of age, and lowest for the very young and very old.
- holding year and age fixed, wage tends to increase with education: the more educated a person is, the higher their salary, on average.

Example

Wage data: GAM with smoothing splines for year and age

```
Anova for Parametric Effects
              Df  Sum Sq Mean Sq F value    Pr(>F)
s(year, 4)     1   27162   27162  21.981 2.877e-06 ***
s(age, 5)      1  195338  195338 158.081 < 2.2e-16 ***
education      4 1069726  267432 216.423 < 2.2e-16 ***
Residuals   2986 3689770    1236
---

Anova for Nonparametric Effects
             Npar Df Npar F  Pr(F)
(Intercept)
s(year, 4)         3  1.086 0.3537
s(age, 5)          4 32.380 <2e-16 ***
education
```

$AIC = 29887.75$

Example

Wage data: GAM with smoothing spline for age

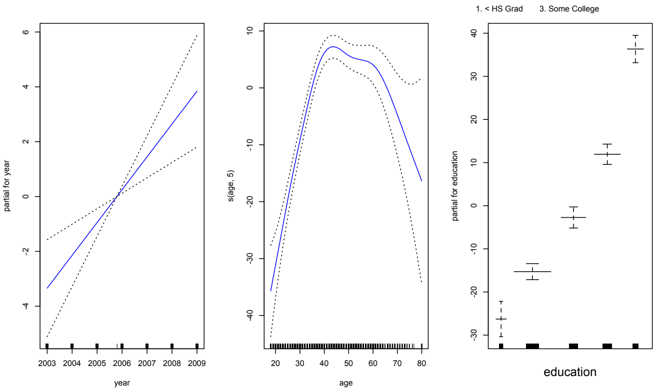
```
Anova for Parametric Effects
            Df  Sum Sq Mean Sq F value    Pr(>F)
year         1   27154   27154  21.973  2.89e-06 ***
s(age, 5)    1  194535  194535 157.415 < 2.2e-16 ***
education    4 1069081  267270 216.271 < 2.2e-16 ***
Residuals 2989 3693842    1236
---


Anova for Nonparametric Effects
            Npar Df Npar F     Pr(F)
(Intercept)
year
s(age, 5)         4  32.46 < 2.2e-16 ***
education
```

$AIC = 29885.06$

Example

Wage data



Example

Wage data: GAM with smoothing spline for year and local regression for age

```
Anova for Parametric Effects
                          Df  Sum Sq Mean Sq F value    Pr(>F)
s(year, df = 4)          1.0   25188   25188  20.255 7.037e-06 ***
lo(age, span = 0.7)      1.0  195537  195537 157.243 < 2.2e-16 ***
education                4.0 1101825  275456 221.511 < 2.2e-16 ***
Residuals             2988.8 3716672    1244
---

Anova for Nonparametric Effects
                      Npar Df Npar F  Pr(F)
(Intercept)
s(year, df = 4)           3.0  1.103 0.3464
lo(age, span = 0.7)       1.2 88.835 <2e-16 ***
education
```

$AIC = 29903.95$

Example

Wage data: linear model

```
Anova for Parametric Effects
            Df  Sum Sq Mean Sq F value    Pr(>F)
year         1   22434   22434  17.421  3.08e-05 ***
age          1  195045  195045 151.460 < 2.2e-16 ***
education    4 1150320  287580 223.317 < 2.2e-16 ***
Residuals 2993 3854286    1288
```

$$AIC = 30004.62$$

Prophet model

Prophet

Prophet: forecasting model developed by Facebook Data Science Team
Let us take a time series model with trend, seasonality and other components (e.g. holidays)

$$y(t) = g(t) + s(t) + h(t) + \varepsilon(t)$$

This is a specification similar to a generalized additive model, GAM

Prophet

Some advantages of this formulation:

- Flexibility
- Interpretability
- Easy to manage

Prophet

Trend

Often the trend component is modeled according to a logistic equation

$$g(t) = \frac{C}{1 + e^{-k(t)}}$$

where C is the carrying capacity and k is the growth rate
The trend can also be linear or constant.

Prophet

Change points

- It is possible to incorporate changes in the trend with change points, to account for a non-constant growth rate
- Suppose there are S change points at $s_j, j = 1, \ldots, S$
- We may define a vector of adjustments δ, where δ_j is the change at time s_j.
- So the growth rate is given by k plus adjustments: $k + \sum_{j:t>s} \delta_j$
- Change points can be manually specified by the analyst, given prior knowledge, or automatically selected.

Prophet

Seasonality

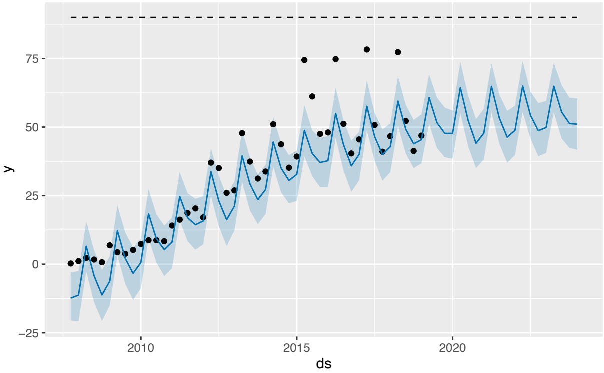
The seasonality component may be defined with a combination of Fourier series

$$s(t) = \sum_{n=1}^{N} \left(a \cos\left(\frac{2\pi n t}{P}\right) + b \sin\left(\frac{2\pi n t}{P}\right)\right)$$

Prophet

Example: Apple iPhone

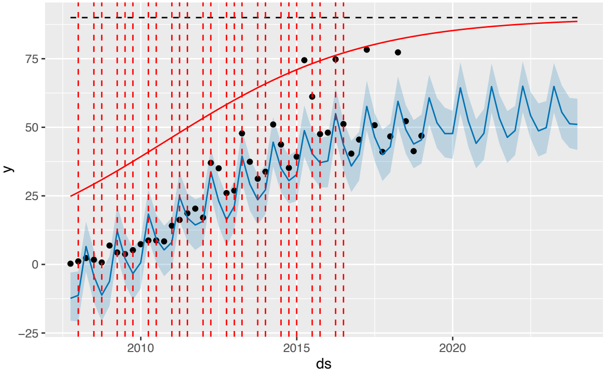
Nonlinear logistic trend and additive seasonality



Prophet

Example: Apple iPhone

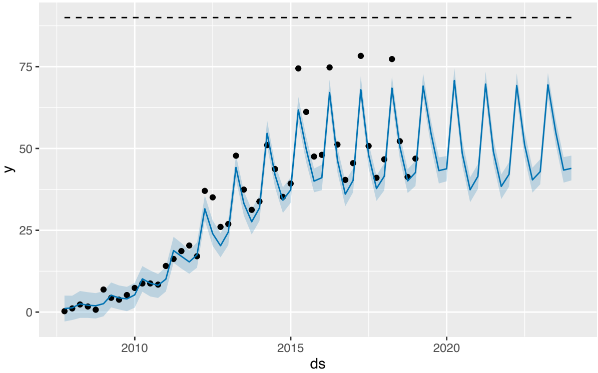
Change points in the first 80% datapoints



Prophet

Example: Apple iPhone

Nonlinear logistic trend and multiplicative seasonality



Prophet

Example: Apple iPhone

Change points