Lecture 18: Finish GAMs, Text mining and topic models

Reading: Sections 7.7

STATS 202: Data mining and analysis

November 4, 2019

Generalized Additive Models (GAMs)

Extension of non-linear models to multiple predictors:

$$\texttt{wage} = \beta_0 + \beta_1 \times \texttt{year} + \beta_2 \times \texttt{age} + \beta_3 \times \texttt{education} + \epsilon$$

$$\longrightarrow \quad \texttt{wage} = \beta_0 + f_1(\texttt{year}) + f_2(\texttt{age}) + f_3(\texttt{education}) + \epsilon$$

The functions $f_1, \ldots, f_p$ can be polynomials, natural splines, smoothing splines, local regressions...

Fitting a GAM

- If the functions f_j have a basis representation, we can simply use least squares:
  - Natural cubic splines
  - Polynomials
  - Step functions

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

Fitting a GAM

- Otherwise, if the fitting procedure for a single f_j is more complex, we can use **backfitting**:

  1. Keep $f_2, \ldots, f_p$ fixed, and fit f_1 using the partial residuals:

     $$y_i - \beta_0 - f_2(x_{i2}) - \cdots - f_p(x_{ip}),$$

     as the response.

  2. Keep $f_1, f_3, \ldots, f_p$ fixed, and fit f_2 using the partial residuals:

     $$y_i - \beta_0 - f_1(x_{i1}) - f_3(x_{i3}) - \cdots - f_p(x_{ip}),$$

     as the response.

  3. ...

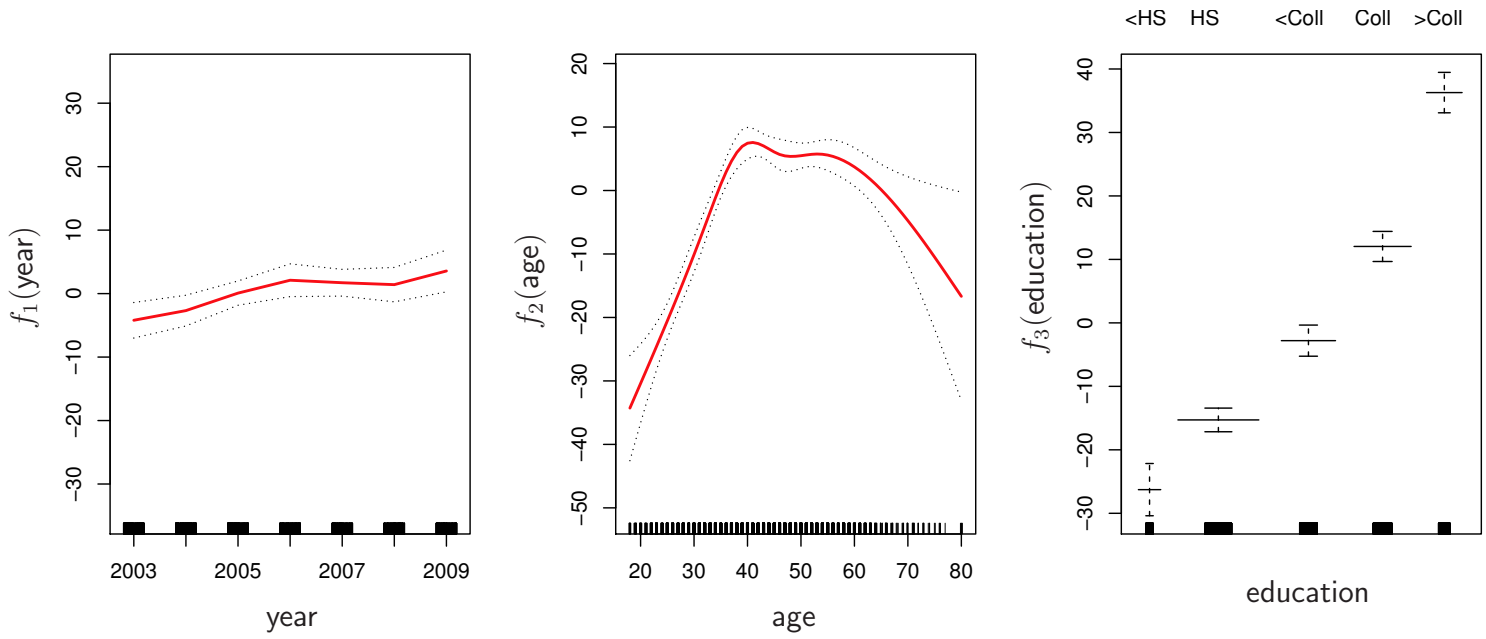
  4. Iterate

- This works for smoothing splines and local regression.

Properties of GAMs

- GAMs are a step from linear regression toward a fully nonparametric method.
- The only constraint is additivity. This can be partially addressed by adding key interaction variables $X_i X_j$.
- We can summarize flexibility of each component function f_j in terms of degrees of freedom.
- Under assumptions analogous to those in linear regression, we can often examine the significance of each of the variables, e.g. by running a F-test on all basis functions associated with the variable.

Example: Regression for `Wage`

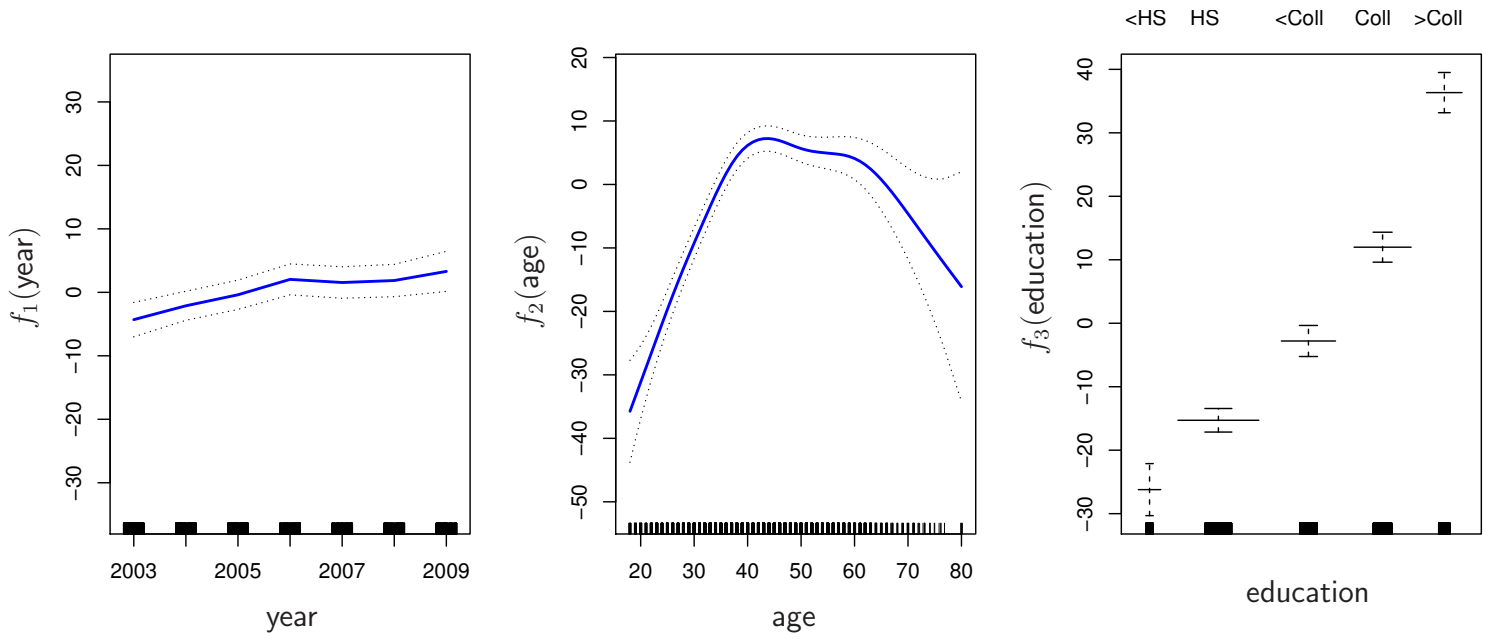


`year`: natural spline with df=4.

`age`: natural spline with df=5.

`education`: step function.

Example: Regression for `Wage`



`year`: smoothing spline with df=4.

`age`: smoothing spline with df=5.

`education`: step function.

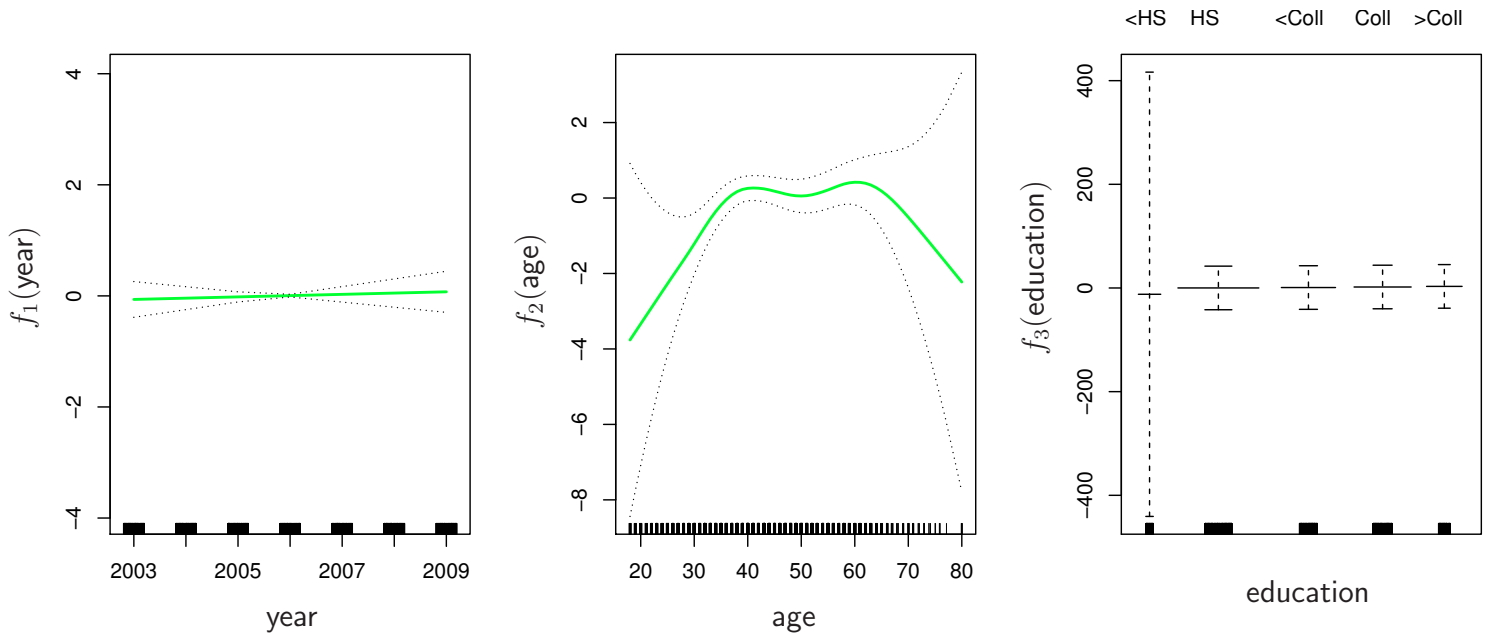
GAMs for classification

We can model the log-odds in a classification problem using a GAM:

$$\log \frac{P(Y = 1 \mid X)}{P(Y = 0 \mid X)} = \beta_0 + f_1(X_1) + \cdots + f_p(X_p).$$

We fit the model either via standard logistic regression (with an augmented set of derived features / basis functions) of via backfitting.

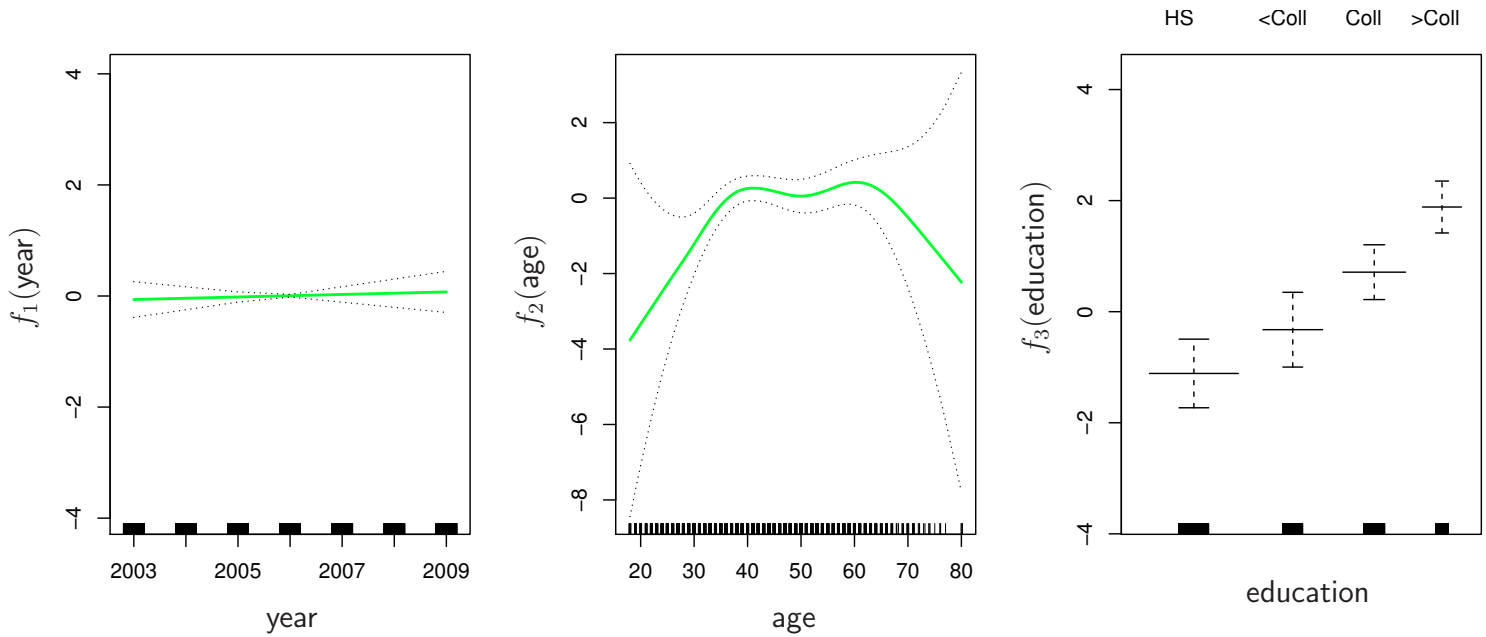
Example: Classification for `Wage>250`



`year`: linear.
age: smoothing spline with df=5.
`education`: step function.

Important fact: no one in the dataset with `education` $<$ `HS` made more than 250 per year.

Example: Classification for Wage>250



year: linear.
age: smoothing spline with df=5.
education: step function.

Exclude samples with education $<$ HS.

Text mining

Problem:

- We have a corpus of documents in English, e.g. the New York Times archive of 13 million documents from 1851-Present.
- How can we gain an unsupervised understanding of the content or and relationship amongst these documents?
  - How do different documents relate to one another?
  - Can the variation in the corpus be described by a few latent or hidden variables?
  - Are there prominent clusters of documents?
  - Can we discover the recurring topics underlying these documents?

Step 1: Turn each document into a "bag of words"

Batista and Regime Flee Cuba; Castro Moving to Take Power; Mobs Riot and Loot in Havana

ARMY HALTS FIRE

Rebels Seize Santiago and Santa Clara-- March on Capital

By R. HART PHILLIPS

Havana, Friday, Jan. 2 -- Fulgencio Batista resigned as President of rebellion-torn Cuba yesterday and fled to exile in the Dominican Republic. The rebel forces of Fidel Castro moved swiftly to seize power throughout the island.

Dr. Manuel Urrutia, Senor Castro's own choice, appeared likely early this morning to become the provisional President. Col. Ramon Barquin, who had been imprisoned for conspiring against the Batista Government, was brought here by military plane from the Isle Pines penitentiary and named chief of the joint staffs.

RELATED HEADLINES

Castro Superior in Arms, Batista Declares in Exile

U.S. Aides Wary On Cuba's Future: Rebel-Conservative Conflict Envisaged -- Castro Men Take Over Embassy

Casinos Wrecked:

↓

{Batista, Cuba, Cuba, Power, Riot, ...}

Apple Introduces Innovative Cellphone

By JOHN MARKOFF
Published: January 10, 2007

SAN FRANCISCO, Jan. 9 — With characteristic showmanship, Steven P. Jobs introduced Apple's long-awaited entry into the cellphone world Tuesday, pronouncing it an achievement on a par with the Macintosh and the iPod.

The creation, the iPhone, priced at $499 or $599, will not be for everyone. It will be available with a single carrier, Cingular Wireless, at midyear. Its essential functions — music player, camera, Web browser and e-mail tool as well as phone — have become commonplace in hand-held devices.

↓

{Apple, phone, innovative, Jobs, ...}

Step 2: Compute a vector of word counts for each document

{Batista, Cuba, Cuba, Power, Riot, ...}

↓

{Batista: 1, Cuba: 2, Castro: 3, ...}

{Apple, phone, innovative, Jobs, ...}

↓

{Apple: 4, phone: 3, innovative: 1, ...}

Latent semantic analysis

- Arrange all the word count vectors in an n by p matrix, where n is the number of documents and p is the number of words in the vocabulary.
- Apply PCA to this matrix.
- The principal components give a weight to each word. They contrast words which explain the most variation in the corpus. For example, if documents are US political blogs, one component might principally explain differences in Democrat vs. Republican vocabulary.

Latent Dirichlet Allocation

Model:

- We have K topics. Each topic is a distribution over words, e.g.

  Topic 1: 0.2 wall street + 0.5 financial + 0.2 fed + 0.1 occupy

  Topic 2: 0.3 parenting + 0.3 tutors + 0.1 ivy + 0.3 admissions

- Each document has a distribution over topics, e.g.

  Document 1: 0.5 Topic 1 + 0.4 Topic 2 + 0.1 Topic 9

  Document 2: 0.7 Topic 5 + 0.3 Topic 3

- To generate a document, first pick its topic distribution and then sample each word independently by (1) picking a random topic, and (2) picking a random word from said topic.

Latent Dirichlet Allocation

Learning algorithm:

- We need to estimate:
  1. The distribution of words in each topic.
  2. The distribution of topics in each document.
- The learning algorithm is Bayesian.
- There are parameters which tune:
  - How concentrated the distribution of each topic is (how many words per topic).
  - How concentrated the distribution of each document is (how many topics per document).