

Introduction to GAM and Its Comparison to GLM

The University of Western Ontario

27-2-2015

Outline

- ▶ Motivation and History of Regression
- ▶ Generalized Linear Models (GLM)
- ▶ Generalized Additive Models (GAM): Basic Theory
- ▶ Generalized Additive Model: An Example

Motivation

$$\text{Assume } \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \text{ and } \mathbf{x} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}.$$

Find an “appropriate” function f such that:

$$\mathbf{y} = f(\mathbf{x}) + \varepsilon$$

where ε is a “noise”. Observe that:

$$\mathbb{E}(\mathbf{y} | \mathbf{x}_1, \dots, \mathbf{x}_p) = f(\mathbf{x}_1, \dots, \mathbf{x}_p).$$

Generalized Linear Models

John Nelder and Robert Wedderburn (1972): If we choose a link function g so that $g(\mu) = \alpha + \mathbf{x}_1\beta_1 + \cdots + \mathbf{x}_p\beta_p$, we get generalized linear model which has three elements:

- (i) ε belongs to the exponential family.
- (ii) A linear predictor $\eta = \mathbf{x}\beta$.
- (iii) A link function g such that:

$$g(\mathbb{E}(\mathbf{y}|\mathbf{x})) = \eta.$$

example: If $f(\mathbf{x}_1, \dots, \mathbf{x}_p) = \alpha + \mathbf{x}_1\beta_1 + \cdots + \mathbf{x}_p\beta_p$ and $g(\mu) = \mu$ we get multiple linear regression. Here

$$\varepsilon \sim \text{I.I.N}(0, \sigma^2)$$

Additive Model (AM)

J. Friedman and W. Stuetzle (1981): If

$$f(\mathbf{x}_1, \dots, \mathbf{x}_p) = \alpha + f_1(\mathbf{x}_1) + \dots + f_p(\mathbf{x}_p)$$

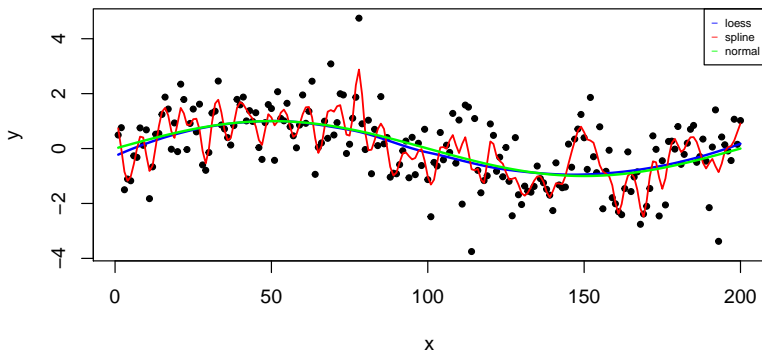
we get additive model, which is a nonparametric regression method.

1. Relaxes parameteric framework (more flexibility).
2. How can we find the appropriate functions?
3. How much complexity is good enough?

Smoothing (Flexibility)

Two smoothing techniques:

- ▶ LOESS(local polynomial regression fitting)
- ▶ Spline Smoothing



$$y = \sin(2\pi x/n) + \epsilon \quad \epsilon \sim N(0, 1) \quad n = 200$$

Backfitting Algorithm for AM

- (i) Initialize: $\hat{\alpha} = \frac{1}{N} \sum_{i=1}^N y_i$, $\hat{f}_j \equiv 0$ for all j .
- (ii) Cycle: $j : 1 \dots, p$:
 - ▶ $\hat{f}_j \leftarrow \mathcal{S}_j \left[\{y_i - \hat{\alpha} - \sum_{k \neq j} \hat{f}_k(x_{ik})\}_{i=1}^N \right]$,
 - ▶ $\hat{f}_j \leftarrow \hat{f}_j - \frac{1}{N} \sum_{i=1}^N \hat{f}_j(x_{ij})$
- (iii) Repeat (ii) until it “converges”.

Here \mathcal{S}_j is a (weighted) smoother operator chosen for the covariate j . Usually, it is chosen to be cubic spline smoother.

GAM in R

gam is implemented in several R packages such as

- ▶ **mgcv**: penalized smoothing spline approach
- ▶ **gam**: more smoother options (loess, spline, etc)
- ▶ **gss**: spline-based approach

LM vs AM

ozone dataset

From “faraway” R-package

(Breiman and Friedman - 1985)

- ▶ Output: O3 (atmospheric ozone concentration)
- ▶ covariates:
 - ▶ temp (temperature measured at El Monte)
 - ▶ ibh (inversion base height at LAX)
 - ▶ ibt (inversion top temperature at LAX)

Comparison of LM and AM

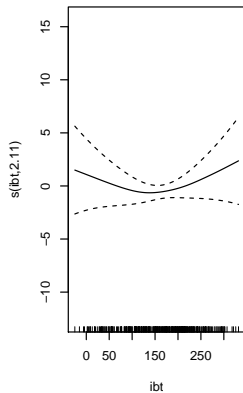
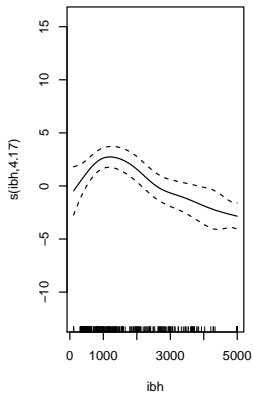
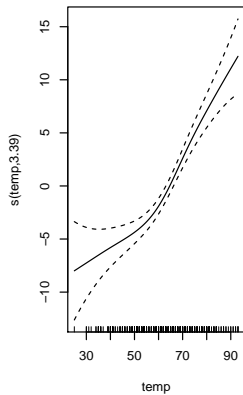
Linear model:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-7.7280	1.6217	-4.77	0.0000
temp	0.3804	0.0402	9.47	0.0000
ibh	-0.0012	0.0003	-4.62	0.0000
ibt	-0.0058	0.0102	-0.57	0.5678

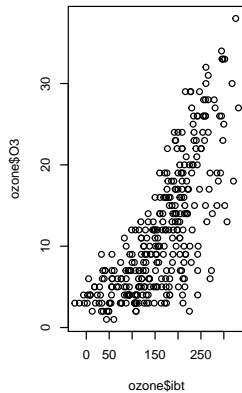
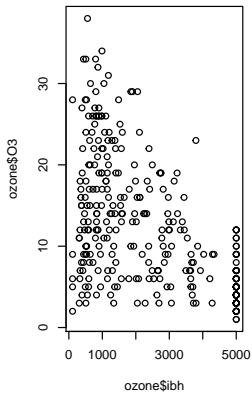
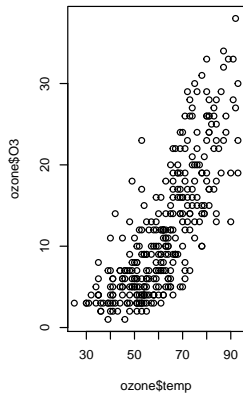
Additive model:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	11.78	0.24	49.44	0.00
s(temp)	3.39	4.26	20.68	0.00
s(ibh)	4.17	5.08	7.34	0.00
s(ibt)	2.11	2.73	1.61	0.19

Comparison of LM and AM



Plot of ozone data



Generalized Additive Model(GAM)

Trevor Hastie and Robert Tibshirani in 1990: Combine GLM and AM to get generalized additive model:

$$g\left(\mathbb{E}(\mathbf{y}|\mathbf{x})\right) = \alpha + f_1(\mathbf{x}_1) + \cdots + f_p(\mathbf{x}_p).$$

Example:

$$\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1\mathbf{x}_1 + \cdots + \beta_p\mathbf{x}_p.$$

$$\log\left(\frac{\pi}{1-\pi}\right) = \alpha + f_1(\mathbf{x}_1) + \cdots + f_p(\mathbf{x}_p).$$

Penalized Least Square

- Minimize:

$$\sum_{i=1}^n \left\{ y_i - \sum_{j=1}^p f_j(x_{ij}) \right\}^2 + \sum_{j=1}^p \lambda_j \int \left(f_j''(t) \right)^2 dt$$

where $\lambda_j \geq 0$ for every j .

- The solution is given by cubic splines and we need to minimize:

$$\left(\mathbf{y} - \sum_{j=1}^p \mathbf{f}_j \right)^\top \left(\mathbf{y} - \sum_{j=1}^p \mathbf{f}_j \right) + \sum_{j=1}^p \lambda_j \mathbf{f}_j^\top \mathbf{K}_j \mathbf{f}_j$$

where \mathbf{K}_j 's are penalty matrices.

- If we differentiate with respect to \mathbf{f}_k we get:

$$\hat{\mathbf{f}}_k = \left(\mathbf{I} + \lambda_k \mathbf{K}_k \right)^{-1} \left(\mathbf{y} - \sum_{j \neq k} \hat{\mathbf{f}}_j \right)$$

Local Scoring Algorithm for Additive Logistic Regression

- (i) Initialize: $\hat{\alpha} = \log \left(\frac{\bar{y}}{1-\bar{y}} \right)$, where $\bar{y} = \text{ave}(y_i)$, and $\hat{f}_j \equiv 0$ for all j .
- (ii) Define $\hat{\eta}_i = \hat{\alpha} + \sum_j \hat{f}_j(x_{ij})$ and $\hat{p}_i = \frac{1}{1+\exp(-\hat{\eta}_i)}$:
 - Put the working target variable:

$$z_i = \hat{\eta}_i + \frac{y_i - \hat{p}_i}{\hat{p}_i(1 - \hat{p}_i)}.$$

- Put weights:

$$w_i = \hat{p}_i(1 - \hat{p}_i).$$

- Use weighted backfitting to find an additive model for the target z_i . This produces new estimates for $\hat{\alpha}$ and \hat{f}_j for every j .

- (iii) Repeat (ii) until it “converges”.

Spam Data Set

From “nutshell” R-package.

```
##      all   our   int    fr busi cred    yo semic
## 1 0.64 0.32 0.00 0.32 0.00 0.00 0.96 0.00
## 2 0.50 0.14 0.07 0.14 0.07 0.00 1.59 0.00
## 3 0.71 1.23 0.12 0.06 0.06 0.32 0.51 0.01
##      par bra    exc dollar pound capav capmax
## 1 0.000    0 0.778 0.000 0.000 3.756      61
## 2 0.132    0 0.372 0.180 0.048 5.114     101
## 3 0.143    0 0.276 0.184 0.010 9.821     485
##      captot   res
## 1      278 spam
## 2     1028 spam
## 3     2259 spam
```


Spam Data-GLM Logistic Regression

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.4350	0.1093	-22.27	0.0000
all	0.2750	0.1014	2.71	0.0067
our	0.5307	0.0744	7.13	0.0000
int	0.9766	0.1845	5.29	0.0000
fr	0.8110	0.1014	8.00	0.0000
busi	0.9430	0.1758	5.36	0.0000
cred	1.4401	0.4027	3.58	0.0003
yo	0.4572	0.0465	9.82	0.0000
semic	-0.6275	0.4553	-1.38	0.1682
par	-1.5458	0.3628	-4.26	0.0000
bra	-3.3787	1.3829	-2.44	0.0146
exc	0.7470	0.1097	6.81	0.0000
dollar	7.7664	0.6891	11.27	0.0000
pound	0.4541	0.1886	2.41	0.0161
capav	0.0061	0.0226	0.27	0.7858
capmax	0.0119	0.0021	5.66	0.0000
captot	0.0005	0.0001	3.37	0.0008

Cross Validation-GLM Logistic Regression

We have divided the data set into two pieces. One as the training set and one as the test set.

	email	spam
1	56.8%	4.5%
2	9.5%	29.1%

We can compute the prediction error here.

Spam Data-GAM Logisitic Regression

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.60	0.24	-2.49	0.01
s(all)	0.00	0.00	0.00	1.00
s(our)	2.22	2.64	70.55	0.00
s(int)	2.91	2.99	40.35	0.00
s(fr)	2.52	2.85	95.83	0.00
s(busi)	1.02	1.24	10.88	0.00
s(cred)	2.74	2.93	7.69	0.05
s(yo)	2.13	2.58	60.01	0.00
s(semic)	2.82	2.97	9.05	0.03
s(par)	1.09	1.24	29.65	0.00
s(bra)	2.27	2.61	7.95	0.03
s(exc)	2.86	2.98	226.12	0.00
s(dollar)	2.97	3.00	143.92	0.00
s(pound)	0.72	0.97	2.84	0.09
s(capav)	2.96	3.00	43.16	0.00
s(capmax)	1.56	1.81	1.24	0.48
s(captot)	2.69	2.92	9.03	0.03

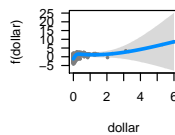
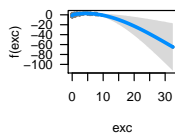
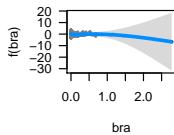
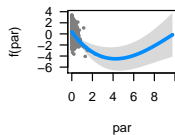
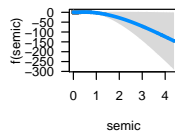
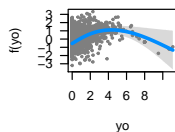
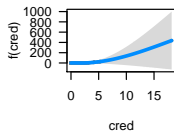
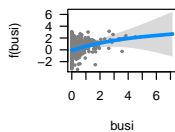
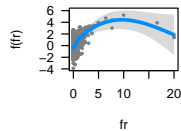
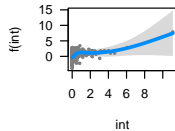
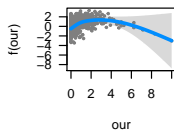
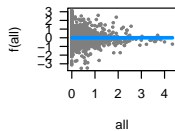
Cross Validation-GAM Logisitic Regression

For the same test set we have the following results.

	email	spam
1	57.6%	3.8%
2	6.0%	32.6%

Again we can compute the prediction error here.

Smoothers Graphs



Application

Probability of default or bankruptcy π .

$$\begin{aligned}\log\left(\frac{\pi}{1-\pi}\right) &= \beta_0 + \beta_1\mathbf{x}_1 + \dots + \beta_p\mathbf{x}_p + \varepsilon \\ \varepsilon &= N(0, \sigma^2)\end{aligned}$$

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1\mathbf{x}_1 + \dots + \beta_p\mathbf{x}_p$$

$$\log\left(\frac{\pi}{1-\pi}\right) = \alpha + f_1(\mathbf{x}_1) + \dots + f_p(\mathbf{x}_p)$$






Conclusion and Comparison

GLM	GAM
Parametric	Non-Parametric
Rigid	Felexible
Easier to Understand/Interpret	Harder to Understand/Interpret
Iterative Weighted Least Square	Backfitting and Local Scoring
Easier to Predict	Higher Accuracy Prediction

Warning: GAM has two main problems:

- (i) Overfitting is an issue.
- (ii) Computationally intense, specially for big data.

References

-  D. Berg, Bankruptcy Prediction by Generalized Additive Models, Department of Mathematics, University of Oslo, Norway, Statistical Research Report No. 1, January 2005
-  J. Faraway: Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models, Chapman and Hall/CRC, 2005
-  T. Hastie, R. Tibshirani: Generalized Additive Models, Chapman and Hall/CRC, 1990.
-  T. Hastie, R. Tibshirani, J. Friedman: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, 2009.
-  S. Wood: Generalized Additive Models : An Introduction with R, Chapman and Hall/CRC, 2006.

THANK YOU!