Module 2: Spline and Kernel Methods

Spline and Kernel Methods for GLMs

STAT/BIOSTAT 527, University of Washington

Emily Fox

May 7th, 2013

Review of GLMs

- Mean parameters are a linear combination of inputs, passed through a possibly nonlinear function

- Assume a distribution in the exponential family

$$p(y \mid x) = \exp\left[\frac{y\theta(x) - b(\theta(x))}{\sigma^2} + c(y, \sigma^2)\right]$$

  - Using theory of exponential families,

$$E[Y \mid x] =$$
$$\mathrm{var}(Y \mid x) =$$

Review of GLMs

$p(y \mid x) = \exp\left[\frac{y\theta(x) - b(\theta(x))}{\sigma^2} + c(y, \sigma^2)\right]$

- Mean parameters are a linear combination of inputs, passed through a possibly nonlinear function
- A parametric GLM assumes

$$g(\mu(x)) = \beta^T x$$

  - With a canonical link function,

$$\theta(x) = g(\mu(x))$$

  - The link function is assumed to be invertible

Examples

$p(y \mid x) = \exp\left[\frac{y\theta(x) - b(\theta(x))}{\sigma^2} + c(y, \sigma^2)\right]$

- Linear regression

$$\log p(y_i \mid x_i, \beta, \sigma^2) = \frac{y_i \mu_i - \frac{\mu_i^2}{2}}{\sigma^2} - \frac{1}{2}\left(\frac{y_i^2}{\sigma^2} + \log(2\pi\sigma^2)\right)$$

Examples

$$p(y \mid x) = \exp\left[\frac{y\theta(x) - b(\theta(x))}{\sigma^2} + c(y, \sigma^2)\right]$$

- Binomial regression

$$\log p(y_i \mid x_i, \beta, \sigma^2) = y_i \log\left(\frac{\pi_i}{1 - \pi_i}\right) + m \log(1 - \pi_i) + \log\binom{m}{y_i}$$

Examples

$$p(y \mid x) = \exp\left[\frac{y\theta(x) - b(\theta(x))}{\sigma^2} + c(y, \sigma^2)\right]$$

- Poisson regression

$$\log p(y_i \mid x_i, \beta, \sigma^2) = y_i \log \mu_i - \mu_i - \log(y_i!)$$

ML Estimation

$p(y \mid x) = \exp\left[\frac{y\theta(x) - b(\theta(x))}{\sigma^2} + c(y, \sigma^2)\right]$

- Maximize the log-likelihood

$$\log p(y_1, \ldots, y_n \mid \beta) =$$

$$\frac{d\ell_i}{d\beta_j} = \frac{d\ell_i}{d\theta_i} \frac{d\theta_i}{d\beta_j} =$$

- No closed-form solution, so use iterative methods
  - 2nd order methods like IRLS require Hessian

$$H = -\frac{1}{\sigma^2} X^T S X \qquad S = \mathrm{diag}(\frac{d\mu_1}{d\theta_1}, \ldots, \frac{d\mu_n}{d\theta_n})$$

ML Estimation

$p(y \mid x) = \exp\left[\frac{y\theta(x) - b(\theta(x))}{\sigma^2} + c(y, \sigma^2)\right]$

- IRLS Newton updates:

$$\beta_{t+1} = (X^T S_t X)^{-1} X^T S_t z_t$$

$$z_t = \theta_t + S_t^{-1}(y - \mu_t)$$

$$\theta_t = X\beta_t \qquad \mu_t = g^{-1}(X\beta_t)$$

Nonparametrics + GLMs

$$p(y \mid x) = \exp\left[\frac{y\theta(x) - b(\theta(x))}{\sigma^2} + c(y, \sigma^2)\right]$$

- Consider a more general form

$$g(\mu(x)) = f(x) \qquad \theta(x) = g(\mu(x))$$

- Can consider many forms for *f*(x) that we have studied in this course, e.g.
  - Smoothing splines
  - Penalized regression splines
  - Local regression (kernel methods)
  - …

Smoothing Splines + GLMs

- For the standard L_2 loss we considered a penalized RSS:

$$\min_f \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \int f''(x)^2 dx$$

  - With normal, additive errors, this is equivalent to penalized log-likelihood

$$\min_f \sum_{i=1}^n \log p(y_i \mid x_i, f) - \frac{1}{2}\lambda \int f''(x)^2 dx$$

  - For GLMs, we just use the specified exponential family distribution instead of a normal likelihood

Smoothing Splines + GLMs

- Penalized log-likelihood with a roughness penalty

$$\min_f \sum_{i=1}^{n} \log p(y_i \mid x_i, f) - \frac{1}{2}\lambda \int f^{''}(x)^2 dx$$

- Example = ***logistic regression***

Bernoulli observations

$$\log p(y_i \mid x_i) = y_i \log\left(\frac{p(x_i)}{1-p(x_i)}\right) + \log(1-p(x_i))$$

modeled as

$$\log p(y_i \mid x_i, f) = y_i f(x_i) - \log(1+e^{f(x_i)})$$

Smoothing Splines + GLMs

- Penalized log-likelihood with a roughness penalty

$$\min_f \sum_{i=1}^{n} \frac{y_i f(x_i) - b(f(x_i))}{\sigma^2} - \frac{1}{2}\lambda \int f^{''}(x)^2 dx$$

- Result is a finite-dimensional natural spline with knots at the unique values of *x*, just as before

$$f(x) = \sum_{j=1}^{n} N_j(x)\beta_j$$

Penalized Reg. Splines + GLMs

- Penalized log-likelihood with a roughness penalty

$$\min_f \sum_{i=1}^{n} \frac{y_i f(x_i) - b(f(x_i))}{\sigma^2} - \frac{\lambda}{2}\beta^T D\beta$$

- Recall that *f* is assumed to be some spline basis expansion
- Derivative with respect to β_j

Penalized Reg. Splines + GLMs

- Penalized log-likelihood with a roughness penalty

$$\frac{d\ell_p}{d\beta_j} = \sum_{i=1}^{n} \frac{d\mu_i}{d\beta_j}\frac{y_i - \mu_i}{\sigma^2 V_i} - \lambda D\beta_j = 0$$

- Again, no closed-form solution as with parametric GLMs
- Use "penalized" IRLS

$$\beta_{t+1} = (X^T S_t X + \lambda D)^{-1} X^T S_t z_t$$

$$z_t = \theta_t + S_t^{-1}(y - \mu_t)$$

$$S_t = \mathrm{diag}\left(\frac{d\mu_1/d\theta_1}{\sigma^2 V_1}, \ldots, \frac{d\mu_n/d\theta_n}{\sigma^2 V_n}\right)$$

- Return: $L^\lambda = X(X^T S X + \lambda D)^{-1} X^T S$

Local Linear Regression

- Consider locally weighted linear regression instead
- Local linear model around fixed target x_0 :

$$\beta_{0x_0} + \beta_{1x_0}(x - x_0)$$

- Minimize:

$$\min_{\underline{\beta}_{x_0}} \sum_i K_\lambda(x_0, x_i)\left(y_i - \beta_{0x_0} - \beta_{1x_0}(x_i - x_0)\right)^2$$

- Return: $\hat{f}(x_0) = \hat{\beta}_{0x_0}$ ← fit at x_0

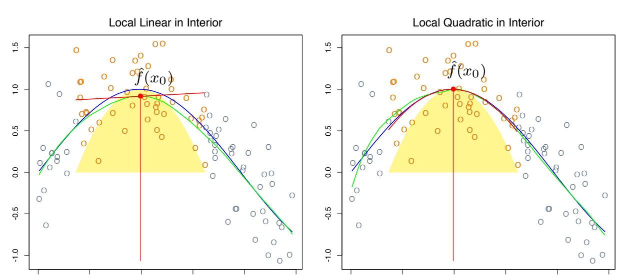
Note: not equivalent to fitting a local constant!

- Fit a new local polynomial for *every* target x_0

©Emily Fox 2013 15

Local Polynomial Regression

- Local linear regression is biased in regions of curvature
  - "Trimming the hills" and "filling the valleys"
- Local quadratics tend to eliminate this bias, but at the cost of increased variance



From Hastie, Tibshirani, Friedman book

©Emily Fox 2013 16

Local Polynomial Regression

- Consider local polynomial of degree *d* centered about x_0

$$P_{x_0}(x;\beta_{x_0}) = \beta_{0x_0} + \beta_{1x_0}(x-x_0) + \frac{\beta_{2x_0}}{2!}(x-x_0)^2 + \dots + \frac{\beta_{dx_0}}{d!}(x-x_0)^d$$

- Minimize: $\min_{\beta_{x_0}} \sum_{i=1}^{n} K_\lambda(x_0, x_i)(y_i - P_{x_0}(x;\beta_{x_0}))^2$
- Equivalently:

$$\min_{\beta_{x_0}} (y - X_{x_0}\beta_{x_0})^T W_{x_0} (y - X_{x_0}\beta)$$

$$X_{x_0} = \begin{bmatrix} 1 & x_1 - x_0 & \cdots & \frac{(x_1-x_0)^d}{d!} \\ \vdots & & & \\ 1 & x_n - x_0 & \cdots & \frac{(x_n-x_0)^d}{d!} \end{bmatrix}$$

- Return: $\hat{f}(x_0) = \hat{\beta}_{0x_0}$
- Bias only has components of degree *d+1* and higher

Local Likelihood Methods

- Just as with spline methods, replace RSS with log-likelihood
- For $\theta_i = x_i^T\beta$

$$\ell(\beta) = \sum_{i=1}^{n} \ell(y_i, x_i^T\beta)$$

- Under a local polynomial model,

$$\ell(\beta) = \sum_{i=1}^{n} K_\lambda(x_0, x_i)\ell(y_i, P_{x_0}(x_i;\beta))$$

Local Likelihood Methods

$$\ell(\beta) = \sum_{i=1}^{n} K_\lambda(x_0, x_i)\ell(y_i, P_{x_0}(x_i;\beta))$$

- Example: ***multiclass logistic regression***
- For

$$Pr(G = j \mid X = x) = \frac{e^{\beta_{j0}+\beta_j^T x}}{1 + \sum_{k=1}^{J-1} e^{\beta_{k0}+\beta_k^T x}}$$

- Under a local polynomial model,

$$\sum_{i=1}^{n} K_\lambda(x_0, x_i)\Bigg\{\beta_{g_i 0}(x_0) + \beta_{g_i}(x_0)^T(x_i - x_0) - \log\left[1 + \sum_{k=1}^{J-1}\exp(\beta_{k0}(x_0) + \beta_k(x_0)^T(x_i - x_0)\right]\Bigg\}$$

Local Likelihood Methods

$$\ell(\beta) = \sum_{i=1}^{n} K_\lambda(x_0, x_i)\ell(y_i, P_{x_0}(x_i;\beta))$$

- Example: ***multiclass logistic regression***
- For

$$Pr(G = j \mid X = x) = \frac{e^{\beta_{j0}+\beta_j^T x}}{1 + \sum_{k=1}^{J-1} e^{\beta_{k0}+\beta_k^T x}}$$

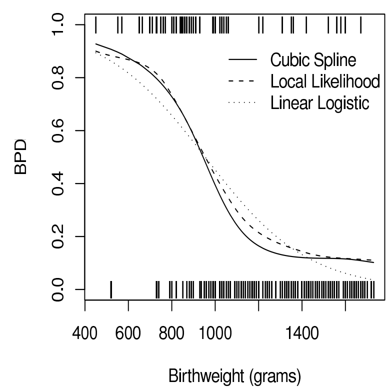
- Return:

Example

- Bronchopulmonary dysplasia (BPD) and birthweight data
- Logistic regression model with binomial observations

$$\log p(y_i \mid x_i, f) = y_i f^\lambda(x_i) - n_i \log(1 + e^{f^\lambda(x_i)})$$

- Choose λ by AIC
- Notice that behavior for high birthweights is quite different from that of linear logistic model



From Wakefield textbook

Example

- Bronchopulmonary dysplasia (BPD) and birthweight data
- Logistic regression model with binomial observations

$$\log p(y_i \mid x_i, f) = y_i f^\lambda(x_i) - n_i \log(1 + e^{f^\lambda(x_i)})$$

- For the local likelihood fit, we have

$$\ell(\beta) = \sum_{i=1}^{n} K_\lambda(x_0, x_i) n_i \left[\frac{y_i}{n_i} P_{x_0}(x_i; \beta) - \log(1 + e^{P_{x_0}(x_i;\beta)}) \right]$$

- Fit uses tri-cube kernel

Local Fits of Autoregressions

$$\ell(\beta) = \sum_{i=1}^{n} K_\lambda(x_0, x_i)\ell(y_i, P_{x_0}(x_i; \beta))$$

- An autoregressive time series model of order *k*

$$\begin{aligned} y_t &= \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_k y_{t-k} + \epsilon_t \\ &= z_t^T \beta + \epsilon_t \end{aligned}$$

- Using a local likelihood approach, can consider kernel

$$K_\lambda(z_0, z_t)$$

- Allows for fit of the autoregressive coefficients to vary in time by considering only a short history