MATH 829: Introduction to Data Mining and Analysis
Lab 1: phoneme dataset

Dominique Guillot

Departments of Mathematical Sciences
University of Delaware

March 18, 2016

Recall: Basis for cubic splines

Cubic splines basis: With 2 knots ξ_1, ξ_2:

$$h_1(X) = 1, \qquad h_3(X) = X^2, \qquad h_5(X) = (X-\xi_1)_+^3,$$
$$h_2(X) = X, \qquad h_4(X) = X^3, \qquad h_6(X) = (X-\xi_2)_+^3.$$

More generally, with M knots, add $(X-\xi_3)_+^3, \ldots, (X-\xi_M)_+^3$.

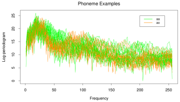
Natural cubic splines basis: With M knots

$$N_1(X) = 1, \quad N_2(X) = X, \quad N_{k+2}(X) = d_k(X) - d_{M-1}(x),$$

where
$$d_k(X) = \frac{(X-\xi_k)_+^3 - (X-\xi_M)_+^3}{\xi_M - \xi_k}.$$

Example: Phoneme recognition

Example: Phoneme Recognition (ESL, Example 5.2.3)

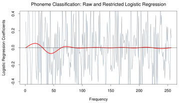


$X = X(f)$
$f =$ frequency.

$$\log \frac{P(aa|X)}{P(ao|X)} = \sum_{i=1}^{256} X(f_i)\beta_i = X^T\beta.$$

15 examples each of the phonemes "aa" and "ao" sampled from a total of 695 "aa"s and 1022 "ao"s.

Phoneme recognition (cont.)



	Raw	Regularized
Training error	0.080	0.185
Test error	0.255	0.158

Logistic regression coefficients, and smoothed version with natural cubic splines.

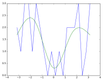
$$\beta(f) = \sum_{i=1}^{M} h_m(f)\theta_m = \mathbf{H}\theta,$$

where $\mathbf{H}$ is a $p \times M$ matrix of spline functions.
Now, note that
$$X^T\beta = X^T\mathbf{H}\theta.$$

Letting $x^* = \mathbf{H}^T x$, we can therefore fit the logistic regression on the *smoothed* inputs.

Work to do

- Write a function to construct natural cubic splines (can use a `class` if you want).
- Test your function:



- Construct the matrix $\mathbf{H} \in \mathbb{R}^{p \times M}$ where $\mathbf{H}_{ij} = h_j(f_i)$ as in the previous slide.
- Load the phoneme data. $X \in \mathbb{R}^{n \times p}$, $y \in \{0, 1\}^n$.
- Use a logistic regression on the transformed data $X\mathbf{H}$ to predict the phonemes. Compute your prediction error.