1D Regression

- ☛ $$y = f(x) + \epsilon$$

  ϵ i.i.d. with mean 0.

- ☛ Univariate Linear Regression:

  $$f(x) = a + bx$$

  fit by least squares. Minimize:

  $$\sum_i (y_i - f(x_i))^2 = \sum_i (y_i - a - bx_i)^2$$

  to get a, b.

- ☛ The set of all possible functions is .....

Non-linear problems

- What if the underlying function is not linear?
- Try: fit non-linear function from a bag of functions
- Problem: which bag? The space of all functions is HUGE
- Another problem: We only have SOME data: want to find the underlying function but avoid noise
- Need to be selective in choosing possible non-linear functions

Basis expansion: polynomial terms

- Univariate LS has two basis functions: $\{h_0(x) = 1, h_1(x) = x\}$
- The resulting fit is a linear combination of h_0, h_1:

$$\widehat{f}(x) = a * h_0(x) + b * h_1(x) = a + b * x$$

- One way: add non-linear functions of x to the bag. Polynomial terms seem as good as any:

$$h_j(x) = x^j, \quad j = 0, \ldots, p$$

- Construct matrix X, with:

$$(X)_{ij} = h_j(x_i)$$

- and fit linear regression with $p + 1$ terms

Global vs Local fits

- One problem with polynomial regression: global fit
- Must find very good global basis for global fit: unlikely to find the “true” one
- Other way: fit locally with “simple” functions
- Why it works: It is easier to find a suitable basis for a part of a function.
- Tradeoff: in each part we only have a fraction of data to work with: must be extra-careful not to overfit.

Polynomial Splines

- Flebility: fit low-order polynomials in small windows of the support of x.
- Most popular are order 4 (cubic) splines
- Must join the pieces somehow: with M-order splines we make sure derivatives up-to M-2 order match at knots
- "Naive" basis for cubic splines:

$$h_w(x) = 1_w . x_w, x_w^2, x_w^3$$

  but many coefficents constrained by macthing derivatives
- Truncated-power basis set:

$$1, x, x^2, x^3, \{(x - \zeta_j)_+^3\}_{j=1}^K$$

  equivalent to "naive" set plus constraints
- Procedure:

- choose knots, $\zeta_j, j = 1, \ldots, K$
- populate matrix X using truncated power basis set (in columns) each evaluated at all data points, x_i (rows)
- Run linear regression with **??** terms.

➠ Natural Cubic splines: $f(x)$ linear beyond data: extra two constraints on each side

➠ The number of parameters (degrees of freedom) is now **?**

Regularization

- Avoid knot-selection problem. Use all possible knots (unique x_i's)
- But have over-parameterized regression (N+2 parameters, N data points)
- Need to regularize (shrink) coeficients:

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_i \left(y_i - \sum_j \beta_j h_j(x_i) \right)^2$$

subject to: $\boldsymbol{\beta}^T \Omega \boldsymbol{\beta} < c$

- Without constraint we get usual least squares fit: here we get infinite number of them
- The constraint on β only allows those fits with certain β.
- Ω controls over-all smoothness of the final fit:

$$\Omega_{jk} = \int h_j''(x) h_k''(x) dx$$

- This remarkably solves a general variational problem:

$$\underset{f}{\operatorname{argmin}} \sum_i (y_i - f(x_i))^2 + \lambda \int_a^b \{f''(t)\}^2 dt$$

λ is in one-to-one correspondance with c above.

- Solution: Natural Cubic Spline with knots at each x_i.
- Benefit: Can get all fits $\hat{f}(x_i)$ in O(N).

B-spline Basis

- Most smoothing splines computationally fitted using B-spline basis
- B-spline are a basis for polynomial splines on a closed interval. Each cubic B-spline spans at most 5 knots.
- Computationally, one sets up an $N \times (N+2)$ matrix X of ordered, evaluated B-spline basis. Each column, j, is a j^{th} B-spline, and its center moves from left-most to right-most point.
- X has banded structure and so does $(X'X + \lambda\Omega)$, where:

$$\Omega(i,j) = \int B_i''(t) B_j''(t) dt$$

- One then solves a penalized regression problem:

$$\widehat{\boldsymbol{f}} = X(X'X + \lambda\Omega)^{-1} X' \boldsymbol{y}$$

- This is actually done using Choleski and back-substitution to get O(N) running time.

- Conceptually, the function f to be fitted is expanded into a B-spline basis set:

$$f(x) = \sum_j \gamma_j B_j(x)$$

and fit obtained by constrained least-squares:

$$\widehat{f} = \underset{f}{\operatorname{argmin}} \, \|\boldsymbol{y} - \boldsymbol{f}\|^2$$

subject to penalty:

$$J(f) < c$$

Here $J(f)$ is the familiar squared second-derivative functional:

$$J(f) = \int \{f''(t)\}^2 dt$$

and c is one-to-one with λ

Equivalent DF

- Smoothing spline (and many other smoothing procedures) are usually called *semi-parametric* models
- Once you expand x into basis set, it looks like any other regression
- BUT individual terms, $h_j(x)$ have no real meaning
- With penalties, one cannot count number of terms to get degrees of freedom
- Equivalent expression is needed for guidance and (approx) inference
- In regular regression:

  $$\mathrm{df} = \mathrm{tr}(X(X^T X)^{-1} X^T)$$

  this is the trace of a *hat*, or projection matrix.
- All penalized regressions (including cubic

smoothing splines) are obtained by:

$$\widehat{\boldsymbol{y}} = H(H^T H + \lambda\Omega)^{-1} H^T \boldsymbol{y} = \boldsymbol{S}\boldsymbol{y}$$

- while $\boldsymbol{S}$ is not a projection matrix, it has similar properties (it is a *shrunk* projection operator)
- Define:

$$\text{EDF} = \text{tr}\boldsymbol{S}$$