Newton's Method and logistic regression, more least squares

least-squares polynomial regression

replace each X_i with feature vector $\Phi(X_i)$

ex. $\Phi(X_i) = [X_{i1}^2 \quad X_{i1}X_{i2} \quad X_{i2}^2 \quad X_{i1} \quad X_{i2} \quad 1]^T$

fictitious dimension

otherwise just like linear / logistic regression

very easy to overfit data

Weighted least squares regression

$$\sum_{i=1}^{n} \omega_i (\hat{Y}_i - Y_i)^2 \qquad \omega_i \to \Omega = \begin{bmatrix} \omega_1 & & 0 \\ & \omega_2 & \\ 0 & & \ddots \\ & & & \omega_n \end{bmatrix}$$

$$\min_w (X^T w - y)^T \Omega (X^T w - y)$$

w/ calculus: $X^T \Omega X w = X^T \Omega y$

Newtons Method

- smooth fn $J(w)$
- faster

at v, approx $J(w)$ by quadratic, jump to critical point

↳ repeat about convergence

Taylor series about v:

$$\nabla J(w) = \nabla J(v) + \underbrace{(\nabla^2 J(v))}_{\text{Hessian of } J}(w-v) + O(\|w-v\|^2)$$

find critical point $\nabla J(w) = 0$:

$$w = v - (\nabla^2 J(v))^{-1} \nabla J(v)$$

Newton's Method:

$w \leftarrow$ init

repeat until convergence:

$e \leftarrow$ sol $(\nabla^2 J(w)) e = -\nabla J(v)$

$w \leftarrow w + e$

- Doesn't know difference of min / max / saddle points
- Depends on starting point

Logistic regression

recall: $s'(\gamma) = s(\gamma)(1 - s(\gamma))$, $s_i = s(X_i^T w)$ $\quad s = \begin{bmatrix} s_1 \\ \vdots \\ s_n \end{bmatrix}$

$$\nabla_w J = -\sum_{i=1}^{n} (y_i - s_i) X_i = -X^T(y - s)$$

$$\nabla_w^2 J = \sum_{i=1}^{n} s_i(1-s_i) X_i X_i^T = X^T \Omega X \quad \Omega := \begin{bmatrix} s_1(1-s_1) & & 0 \\ & \ddots & \\ 0 & & s_n(1-s_n) \end{bmatrix} \succ 0$$

$\Rightarrow X^T \Omega X \succeq 0$

$\Rightarrow J(w)$ is convex

Newton's method:

$w \leftarrow 0$

repeat until convergence:

fns of w

$e \leftarrow$ sol to $(X^T \Omega X)e = X^T(y-s)$

$w \leftarrow w+e$

iteratively weighted least squares

LDA vs. Logistic Regression

advantage of LDA:

- for well seperated classes, stable; log reg. unstable
- >2 classes; log reg better for 2 class (softmax handler)
- LDA slightly more accurate

advantages of log reg:

- more emphasis on decision boundary (close points in LDA not given diff. priority)
- always seperates linearly seperable pts
- more robust to non-gaussian dists. (larger skew)
- naturally fits in $(0,1)$

ROC Curves (for test sets)