LINEAR REGRESSION

GENERAL MODEL: $P(y|x,\theta) = N(y|w^T x, \sigma^2)$

BASIS FUNCTION EXPANSION: $P(y|x,\theta) = N(y|w^T \phi(x), \sigma^2)$

STILL LINEAR WRT PARAMETERS

HIGHER DEGREES OF $\phi(x) \longrightarrow$ MORE COMPLEX FUNCTIONS

MLE ESTIMATION

$\hat{\theta} = \text{ARGMAX} \log p(D|\theta)$. SAMPLES ARE ASSUMED IID, $\ell(\theta) = \sum^{N} \log p(y_i|x_i,\theta)$, BUT WE MINIMIZE NLL BECAUSE EQUIVALENT AND EASIER

- PLUG-IN GAUSSIAN FORM

$$\ell(\theta) = -\frac{1}{2\sigma^2} \underbrace{\sum^{N} (y_i - w^T x_i)^2}_{\text{RSS (SSE)}} - \frac{N}{2} \log(2\pi\sigma^2), \quad \frac{SSE}{N} = MSE, \quad RSS(w) = \|\epsilon\|_2^2 = \sum_1^N \epsilon_i^2$$

- **LEAST SQUARES** BECAUSE MLE FOR w MINIMIZES RSS — **NLL SURFACE** IS QUADRATIC BOWL, CONVEX, UNIQUE MINIMUM, WE DERIVE

- **DERIVATION** $NLL(w) = \frac{1}{2} w^T (X^T X) w - w^T (X^T y)$, $X^T X = \sum^N x_i x_i^T$ = SUM-OF-SQUARES MATRIX

$$X^T y = \sum^N x_i y_i$$

GRADIENT: $G(w) = [X^T X w - X^T y] = \sum^N x_i (w^T x_i - y_i) = 0 \longrightarrow X^T X w = X^T y$

$$\hat{w}_{OLS} = (X^T X)^{-1} X^T y$$

- **GEOMETRIC INTERPRETATION**

N EXAMPLES, D FEATURES, $N > D$. COLS OF X ARE LINEAR SUBSPACE OF $|D|$ EMBEDDED IN N DIMENSIONS.

WE SEEK $\hat{y} \in \mathbb{R}^n$ WHICH LIES IN COLUMN LINEAR SUBSPACE AND IS CLOSEST AS POSSIBLE TO y = ARGMIN $\|y - \hat{y}\|_2$

IS OK BECAUSE $\hat{y} \in \text{SPAN}(X)$, w EXISTS. $\longrightarrow$ **RESIDUAL VECTOR** TO BE ORTHOGONAL TO EVERY COLUMN IN X

$$\tilde{x}_j^T (y - \hat{y}) = 0 \longrightarrow X^T (y - Xw) = 0 \longrightarrow \hat{w} = (X^T X)^{-1} X^T y$$

$$\hat{y} = X\hat{w} = \underbrace{X(X^T X)^{-1} X^T}_{\text{PROJECTION / HAT MATRIX}} y \longrightarrow$$ **ORTHOGONAL PROJECTION** OF y ON COLSPACE OF X

- **CONVEXITY**

- A SET IS CONVEX IF WE DRAW LINE BETWEEN TWO POINTS AND LINE ALWAYS LIES INSIDE THE SET
- A **FUNCTION** IS CONVEX IF **EPIGRAPH** (SET OF PTS ABOVE FCN) IS CONVEX • **CONCAVE** $\longrightarrow -f(\theta)$ IS CONVEX
- UNIQUE GLOBAL MINIMUM $\longrightarrow$ SECOND DERIVATIVE ALWAYS POSITIVE $\longrightarrow$ A TWICE-DIFF.LE, CONTINUOUS, MULTIVAR FCN IS CONVEX IFF HESSIAN IS POSITIVE DEFINITE FOR ALL θ

- **ROBUST LINEAR REGRESSION**

OLS IS VERY SENSITIVE TO OUTLIERS BECAUSE QUADRATIC LOSS $\longrightarrow$ HIGHER IMPACT. REPLACE GAUSSIAN RESPONSE W/ STH MORE HEAVY TAIL

I.E. LAPLACE DISTRIBUTION $P(y|x,w,b) = \text{LAP}(y|x^T w, b) \propto \exp\left(-\frac{1}{b} \underbrace{|y - w^T x|}_{R}\right)$ $\ell(w) = \sum |R_i(w)|$ RESIDUALS

NONLINEAR!

- OPTIMIZE NLL (SPLIT VARIABLE TRICK)
- OPTIMIZE **HUBER LOSS**

$$L_H(R,\delta) \begin{cases} R^2/2 & |R| \le \delta \quad \ell_2 \\ \delta|R| - \delta^2/2 & |R| > \delta \quad \ell_1 \end{cases}$$

EVERYWHERE DIFF, C^1, FASTER 2 OPTIMIZE BC. SMOOTH OPT. METHODS

VARIANTS OF LINEAR REGRESSION

	LIKELIHOOD	PRIOR
LEAST SQUARES	GAUSSIAN	UNIFORM
RIDGE	GAUSSIAN	GAUSSIAN
LASSO	GAUSSIAN	LAPLACE
ROBUST	LAPLACE	UNIFORM
ROBUST	STUDENT	UNIFORM

RIDGE REGRESSION

RESILIENT TO OVERFITTING. GAUSSIAN PRIOR. $P(w) = \prod_j N(w_j | 0, \tau^2)$ ENCOURAGES PARAMS TO BE SMALL

- ARGMAX $\sum^{N} \log N(y_i | w_0 + w^T x_i, \sigma^2) + \sum^{D} \log N(w_j | 0, \tau^2)$ $\xrightarrow[\text{MINIMIZE}]{}$ $J(w) = \underbrace{\frac{1}{N} \sum (y_i - (w_0 + w^T x_i))^2}_{\text{MLE/NLL}} + \underbrace{\lambda \|w\|_2^2}_{\text{PENALTY}}$

$\hat{w}_{\text{RIDGE}} = (\lambda I_D + X^T X)^{-1} X^T y$

$\lambda = \sigma^2 / \tau^2$, $\|w\|_2^2 = w^T w$ (SQUARED TWO NORM)

- λ PICKED WITH CV

- GAUSSIAN PRIOR $\longrightarrow$ ℓ_2 REGULARIZATION / WEIGHT DECAY. PENALIZES SUM OF MAGNITUDE OF ws

– COMPUTATIONAL EFFICIENCY

$(\lambda I_D + X^T X)$ IS BETTER CONDITIONED / MORE LIKELY TO BE INVERTED BUT 4 NUMERICAL STABILITY IT'S BETTER NOT TO INVERT MATRICES ALTOGETHER

AUGMENT X WITH DATA FROM PRIOR $\tilde{X} = \begin{pmatrix} X/\sigma \\ \sqrt{\Lambda} \end{pmatrix}$ $\tilde{y} = \begin{pmatrix} y/\sigma \\ 0_{D\times 1} \end{pmatrix}$ $\tilde{X} = QR$, Q IS ORTHONORMAL $(Q^T Q = QQ^T = I)$ AND R IS UPPER TRIANGULAR (EASY TO INVERT)

- QR DECOMP.

  $w_{\text{RIDGE}} = R^{-1} R^{-T} R^T Q^T \tilde{y} = R^{-1} Q \tilde{y}$

  OK EVEN IF X, y (IMPLEMENTATION ALL USE QR DECOMPOSITION) $\underline{O(ND^2)}$

- $D \gg N$ DO SVD DECOMPOSITION FIRST. $w_{\text{RIDGE}} = V(Z^T Z + \lambda I_N)^{-1} Z^T y$ REPLACE x_i (D-DIM) WITH z_i (N-DIM), THEN RETRANSFORM TO |D| WITH V, $O(DN^2)$

– SHRINKAGE

RELATION BETWEEN RIDGE PREDICTIONS AND SINGULAR VALUES OF X (VIA SVD)

SMALLER SINGULAR VALUES ARE DIRECTIONS WITH HIGHER POSTERIOR VARIANCE $\longrightarrow$ MOST SHRINKED

$\text{DOF}(\lambda) = \sum^{D} \frac{\sigma^2}{\sigma_j^2 + \lambda}$ $\quad \lambda = 0 \to D$ $\quad \lambda = \infty \to 0$

(THESE ARE THE ONES WE $\overset{\text{RIDGE}}{\text{AXE}}$) SINGULAR VALUES ARE EIGENVECTORS OF $X^T X$. OBS PCA

- CHOLESKY DECOMPOSITION $\Lambda = \sqrt{\Lambda} \cdot \sqrt{\Lambda}^T$
- QR DECOMPOSITION $X = QR$, Q ORTHONORMAL, R UPPER TRIANGULAR
- SVD DECOMPOSITION $X = U \Sigma V^T$, $V^T V = I$, $UU^T = U^T U = I$, Σ DIAG, $Z = UD$

LMS ALGORITHM (ONLINE LINEAR REGRESSION) ALSO DELTA RULE / WIDROW - HOFF RULE

$g_u = x_1(\theta_u^T x_1 - y_1)$ → GRADIENT ACTS AS ERROR SIGNAL

$\theta_{u+1} = \theta_u - \eta_u(\hat{y}_u - y_u)x_u$

NO PROJECTION STEP BECAUSE UNCONSTRAINED.

USUALLY $0{,}1 < \eta < 0{,}4$

LINEAR SEPARABILITY

$w^T \perp$ TO DECISION BOUNDARY **MARGIN:** DISTANCE BETWEEN (OPTIMAL) HYPERPLANE AND ANY DATAPOINT

BAYESIAN LIN REG

FULL POSTERIOR OVER W AND σ^2

- σ^2 KNOWN
  - POSTERIOR: $P(w \mid X, y, \sigma^2) = N(w \mid w_N, V_N)$ $\Big[$ LIKELIHOOD: $P(y \mid X, w, \mu, \sigma^2) = N(y \mid \mu + Xw, \sigma^2 I)$

    $w_N = V_N V_0^{-1} w_0 + \frac{1}{\sigma^2} V_N X^T y$ CONJ. PRIOR: $P(w) = N(w \mid w_0, V_0)$ $\Big]$

    $V_N = \sigma^2 (\sigma^2 V_0^{-1} + X^T X)^{-1}$ IF $w_0 = 0$ AND $V_0 = \tau^2 I \longrightarrow$ RIDGE ESTIMATE

  - POSTERIOR PREDICTIVE

$$P(y \mid x, D, \sigma^2) = \int N(y \mid x^T w, \sigma) N(w \mid w_N, V_N)\, dw = N(y \mid w_N^T x, \sigma_N^2(x)) \quad , \sigma_N^2 = \sigma^2 + \underbrace{x^T V_N x}_{\text{DEPENDS ON HOW CLOSE } x \text{ IS TO TRAINING DATA}}$$

- σ^2 UNKNOWN

~~[illegible]~~ MURPHY PP. 234

EMPIRICAL BAYES (EVIDENCE PROCEDURE)

PICK THE HYPERPARAMETERS OF THE PRIOR $\eta = (\alpha, \beta)$ TO MAXIMIZE MARGINAL LIKELIHOOD $\lambda = 1/\sigma^2$ PRECISION OF NOISE, α PRECISION OF PRIOR $P(w) = N(w \mid 0, \alpha^{-1} I)$. ALTERNATIVE TO CROSS-VALIDATION. • BETTER BECAUSE ED ENABLES COMPUTING OF DIFFERENT α_i FOR EVERY FEATURE $\longrightarrow$ FEATURE SELECTION VIA ARD (AUTOMATIC RELEVANCY DETERMINATION). IMPOSSIBRU WITH CV

- USEFUL TO COMPARE DIFFERENT KINDS OF MODELS

$$P(D \mid m) = \iint P(D \mid w, m) P(w \mid m, \eta) P(\eta \mid m)\, dw\, d\eta \approx \max_{\eta} \int P(D \mid w, m) P(w \mid m, \eta) P(\eta \mid m)\, dw$$

ARD: