LOGISTIC REGRESSION

IS DISCRIMINATIVE CLASSIFIER, ONLY FITS THE MODEL, DIRECTLY FIT $P(Y|X)$, NO JOINT $P(X,Y)$ MODEL

MODEL: $P(y|x,w) = \mathrm{BER}(y|\mathrm{SIGM}(w^T x))$

MLE ESTIMATE: $NLL(w) = -\sum \log\left[\mu_i^{I(y_i=1)} \times (1-\mu_i)^{I(y_i=0)}\right] = -\sum^N \left[y_i \log \mu_i + (1-y_i)\log(1-\mu_i)\right]$

CROSS ENTROPY ERROR FCN

$\tilde{y}(n) = \hat{y} - y$

- $NLL = \sum^N \log(1+\exp(-\tilde{y}_i w^T x_i))$ — NO CAN WRITE IN CLOSED FORM $\rightarrow$ OPTIMIZATION ALGORITHM, WE NEED GRADIENT AND HESSIAN

$g(w) = \frac{d}{dw} f(w) = \sum_i (\mu_i - y_i) x_i = X^T(\mu - y)$

$H(w) = \frac{d}{dw} g(w)^T = \sum_i (\nabla_w \mu_i) x_i^T = \sum_i \mu_i(1-\mu_i) x_i x_i^T = X^T S X$; $S = \mathrm{DIAG}(\mu_i(1-\mu_i))$, H IS POSITIVE DEFINITE OFC SO NLL IS CONVEX, YAY!

GRADIENT DESCENT (STEEPEST DESCENT)

$\theta_{k+1} = \theta_k - \eta_k g_k$, η_k LEARNING RATE ; HOW DO WE SET IT?

— CONSTANT $\longrightarrow$ TOO LOW, TAKES LONG TIME
$\longrightarrow$ TOO LARGE, FAILS TO CONVERGE, OSCILLATES.

— **LINE SEARCH**

GUARANTEED TO CONVERGE (GLOBAL CONVERGENCE) NO MATTER WHERE IT STARTS

$f(\theta + \eta d) \approx f(\theta) + \eta g^T d$, TAYLOR ; d IS DESCENT DIRECTION. WE WANT IT SMALL SO THAT $f(\theta+\eta d) < f(\theta)$ BUT NOT TOO SMALL. η TO MINIMIZE: $\phi(\eta) = f(\theta_k + \eta d_k)$. $\phi'(\eta) = 0$ || $\phi'(\eta) = d^T g$ BY CHAIN RULE

- ZIG-ZAGS

— $g = 0$: WE ARRIVED

— $g \perp d$: STEP STOPS WHERE GRADIENT IS $\perp$ TO SEARCH DIRECTION

TO MINIMIZE ZIGZAGGING

- MOMENTUM $\theta_{k+1} = \theta_k - \eta_k g_k + \mu_k(\theta_k - \theta_{k-1})$, $0 < \mu_k < 1$
(HEAVY BALL METHOD)

- CONJUGATE GRADIENTS QUADRATIC OBJECTIVES $f(\theta) = \theta^T A \theta$

REGULARIZATION

ACTUAL MLE IS WHEN $\|w\| \to \infty$, INFINITELY STEEP SIGMOID $I(w^T x > w_0)$, BRITTLE SOLUTION, POOR GENERALIZATION

L_2 REGULARIZE: $f'(w) = NLL(w) + \lambda w^T w$ AND PLUG IN GRADIENT OPTIMIZER

$g'(w) = g(w) + \lambda w$

$H'(w) = H(w) + \lambda I$

NEWTON'S METHOD

SECOND ORDER METHOD, TAKES CURVATURE OF SPACE INTO ACCOUNT (HESSIAN), FASTER $\theta_{n+1} = \theta_n - \eta_n H_n^{-1} g_n$. STEP $d = -H^{-1} g_n$ MINIMIZES 2^{ND} ORDER APPROXIMATION OF F

STEPS

- UNTIL CONVERGENCE, STEP = K
- $g_n = \nabla f(\theta_n)$, $H_n = \nabla^2 f(\theta_n)$
- $H_n d_n = -g_n$, SOLVE FOR d_n
- LINE SEARCH TO MINIMIZE η_n ON d_n
- $\theta_{n+1} = \theta_n + \eta_n d_n$

• IF H_n IS NOT POSITIVE DEFINITE, FCN NOT CONVEX, GO BACK TO STEEPEST DESCENT

LEVENBERG - MARQUARDT ADAPTS BETWEEN NEWTON AND S.D. STEPS

- $\theta_{n+1} = \theta_n - (H_n + \lambda \, \text{DIAG}(H_n))^{-1} \cdot g_n$ $\begin{cases} \lambda \gg \to \text{VANILLA GRAD DESCENT} \\ \lambda \ll \to \text{NEWTON'S} \end{cases}$

• ERROR INCREASE → REJECT UPDATE STEP AND INCREASE λ

ERROR DECREASE → ACCEPT UPDATE AND DECREASE λ

IRLS - ITERATIVELY REWEIGHTED LEAST SQUARES

NEWTON'S TO FIND MLE FOR BINARY LOGISTIC REGRESSION $w_{n+1} = w_n - H^{-1} g_n = (X^T S X)^{-1} X^T S_n z_n$ HESSIAN IS EXACT SO $\eta_n = 1$, COOL.

$z_n = X w_n + S^{-1}(y - \mu_n)$ **WORKING RESPONSE**

IS WEIGHTED LEAST SQUARE MINIMIZING $\sum S_{ni} (z_{ni} - w^T x)^2$ **S IS DIAGONAL** → $z_{ni} = w_n^T x_i + \dfrac{y_i - \mu_{ni}}{\mu_{ni}(1 - \mu_{ni})}$

STEPS

- $w_* = 0_D$
- $w_0 = \log(\bar{y}/(1-\bar{y}))$

=

$\eta_i = w_0 + w^T x_i$ $\quad z_i = \eta_i + \dfrac{y_i - \mu_i}{s_i}$

$\mu_i = \text{SIGM}(\eta_i)$ $\quad S = \text{DIAG}(s_{1:N})$

$s_i = \mu_i(1 - \mu_i)$ $\quad w = (X^T S X)^{-1} X^T S z$

UNTIL CONVERGENCE

BFGS

IS QUASI-NEWTON METHODS, H CAN BE EXPENSIVE TO COMPUTE EXPLICITLY. APPROXIMATE H USING GRADIENT AT EACH STEP

$B_n \approx H_n$

$$B_{n+1} = B_n + \frac{y_n y_n^T}{y_n^T s_n} - \frac{(B_n s_n)(B_n s_n)^T}{s_n^T B_n s_n} \quad \text{BFGS FORMULA}$$

$s_n = \theta_n - \theta_{n-1} \qquad y_n = g_n - g_{n-1}$

• $B_0 = I$

ENSURES MATRIX REMAINS POSITIVE DEF.

'DIAGONAL + LOW RANK APPROXIMATION'

• $O(D^2)$ SPACE

• **ALTERNATIVE:** BFGS APPROXIMATES $C_n \approx H_n^{-1}$ INVERSE HESSIAN RIGHT AWAY

L-BFGS

LIMITED MEMORY VERSION

IS ACTUALLY DIAGONAL + LOW RANK, USES ONLY m MOST RECENT (s_n, y_n) • $O(mD)$ • **m ~ 20**

OBS: BROYDEN FAMILY FORMULAS: $H_{n+1} = (1-\phi) H_{n+1}^{DFP} + \phi H_{n+1}^{BFGS}$ **HDFP IS** SIMILAR BUT LESS ROBUST THAN BFGS

MULTI-CLASS LOGISTIC REGRESSION

OR MAX-ENTROPY CLASSIFIER $\quad P(y=c|x,W) = \dfrac{\exp(w_c^T x)}{\sum_c \exp(w^T x)}$

- CONDITIONAL LOGIT MODEL NORMALIZES OVER DIFFERENT CLASSES FOREACH DATACASE

- $\mu_{ic} = P(y_i=c|x_i,W_i) = S(\eta_i)_c$, $\eta_i = W^T x_i$ IS $C\times 1$ VECTOR , $y_{ic} = \mathbb{I}(y_i=c)$ ONE OF C ENCODING , BIT = 1 IFF $y_i = c$

- $\ell(w) = \log \prod^N \prod^C \mu_{ic}^{y_{ic}} = \sum^N \sum^C y_{ic} \log \mu_{ic} = \sum^N \left[\left(\sum^C y_{ic} w_c^T x_i \right) - \log \left(\sum^C \exp(w_c^T x_i) \right) \right]$

- NLL $= -\ell(w)$

GRADIENT: $G(w) = \nabla f(w) = \sum^N (\mu_i - y_i) \otimes x_i$

$$A \otimes B = \begin{bmatrix} a_{11}B & \dots & a_{1n}B \\ \vdots & & \vdots \\ a_{m1}B & \dots & a_{mn}B \end{bmatrix}$$

KRONECKER PRODUCT

$y_i = (\mathbb{I}(y=1) \dots \mathbb{I}(y=C-1))$ $\quad \mu_i(w) = [P(y_i=1|x_i,W_i) \dots P(y_i=C-1|X,W)]$

$$G(w) = \sum_i \begin{bmatrix} (\mu_{i1}-y_{i1})x_{i1} \\ (\mu_{i1}-y_{i1})x_{i2} \\ \vdots \\ \vdots \end{bmatrix}$$

$\nabla_{w_c} f(w) = \sum_i (\mu_{ic} - y_{ic}) x_i$

SAME FORM AS BINARY LOGISTIC REGRESSION ; ERROR TIMES x_i $\longrightarrow$ BECAUSE IS GLM

HESSIAN: $H(w) = \nabla f(w) = \sum^N (\mathrm{DIAG}(\mu_i) - \mu_i \mu_i^T) \otimes (x_i x_i^T)$, ALSO BLOCK, $H_{cc'}(w) = \sum_i \mu_{ic}(\delta_{c,c'} - \mu_{ic'}) x_i x_i^T$

IS POSITIVE DEFINITE $\longrightarrow$ UNIQUE MLE

MINIMIZING $f'(w) = -\log P(D|W) - \log P(W)$, $P(W) = \prod_c \mathcal{N}(w_c|0, V_0)$

- $f'(w) = f(w) + \frac{1}{2} \sum_c w_c V_0^{-1} w_c$

- $g'(w) = g(w) + V_0^{-1} \left(\sum^c w_c \right)$

- $H'(w) = H(w) + I_c \otimes V_0^{-1}$

HESSIAN IS $O((CD) \times (CD))$ SO HERE BFGS IS THUMBS UP!

BAYESIAN LOGISTIC REGRESSION

WE WANT FULL POSTERIOR OVER PARAMS $p(w|D)$, NOT CONVENIENTE BECAUSE L.R. HAS NO CONJUGATE PRIOR. APPROXIMATIONS I.E. MCMC, VARIATIONAL INFERENCE, EXPECTATION PROPAGATION, BUT LATER.

LAPLACE APPROXIMATION

LET'S APPROXIMATE THE POSTERIOR TO A GAUSSIAN $p(\theta|D)=\frac{1}{Z}e^{-E(\theta)}$ **E(θ) ENERGY FUNCTION** $E(\theta)=-\log p(\theta,D)$; Z NORM CONSTANT

$$E(\theta)\approx E(\theta^*)+(\theta-\theta^*)^T g+\frac{1}{2}(\theta-\theta^*)^T H(\theta-\theta^*)$$

TAYLOR EXPANSION AROUND THE MODE

$g=\nabla E(\theta)|_{\theta^*}$ $H=\frac{\partial^2 E(\theta)}{\partial\theta\,\partial\theta^T}\Big|_{\theta^*}$ MODE → GRADIENT IS 0

$$p(\theta|D)=\mathcal{N}(\theta|\theta^*,H^{-1})$$

$$Z\approx\int\hat{p}(\theta|D)\,d\theta=e^{-E(\theta^*)}(2\pi)^{D/2}|H|^{-1/2}=p(D)$$

← LAPLACE APPROXIMATION OF ML

ALSO GAUSSIAN APPROXIMATION, SADDLE POINT APPROXIMATION

$$\log p(D)\approx\log p(D|\theta^*)+\underbrace{\log p(\theta^*)-\frac{1}{2}\log|H|}_{\text{OCCAM FACTOR}}$$

OCCAM FACTOR, MEASURES MODEL COMPLEXITY. IF UNIFORM PRIOR $p(\theta)\propto 1$ NO 2nd TERM (NO CURVATURE) AND $\theta^*\rightarrow\hat{\theta}_{MLE}$

$$\log p(D)\approx\log p(D|\hat{\theta})-\frac{D}{2}\log N$$ ← BIC SCORE

GAUSSIAN APPROXIMATION OF LOGREG

GAUSSIAN PRIOR ↓ $p(w)=\mathcal{N}(w|0,V_0)$

$p(w|D)\approx\mathcal{N}(w|\hat{w},H^{-1})$ $\hat{w}=\arg\min_w E(w)$, $E(w)=-(\log p(D|w)+\log p(w))$, $H=\nabla^2E(w)|_{\hat{w}}$

ISSUES WHEN DATA IS LINEARLY SEPARABLE → MLE NOT WELL DEFINED, GOES TO INFINITY, SIGMOID VERY STEEP

REGULARIZE W/ SPHERICAL PRIOR $\mathcal{N}(w|0,100I)$. BETTER THAN MAP ALONE

POSTERIOR PREDICTIVE

$p(y|x,D)=\int p(y|x,w)\,p(w|D)\,dw$ INTRACTABLE IN THIS CASE. PLUG-IN APPROX $p(y=1|x,D)\approx p(y=1|x,E[w])$ $E[w]$ POSTERIOR MEAN

BAYES POINT, UNDERESTIMATES UNCERTAINTY

— **MONTE CARLO APPROXIMATION** $p(y=1|x,D)\approx\frac{1}{S}\sum sigm((w^s)^Tx)$, $w^s\sim p(w|D)$ SAMPLES FROM THE POSTERIOR

IF MC POSTERIOR → SAMPLES 4 PREDICTION. IF GAUSSIAN POSTERIOR → NO SAMPLES FROM IT

— **PROBIT APPROXIMATION**

$p(w|D)\approx\mathcal{N}(w|m_N,V_N)$ GAUSSIAN APPROX POSTERIOR | $p(y=1|x,D)=\int sigm(a)\,\mathcal{N}(a|\mu_a,\sigma_a^2)\,da$ POSTERIOR PREDICTIVE

$=\cancel{x^T V_N x}$ ~~[MURPHY P. 259]~~

APPROX SIGM WITH PROBIT = CDF OF NORMAL DISTRIBUTION

ANALYTICALLY CONVOLUABLE WITH GAUSSIANS

$sigm(a)\approx\Phi(\lambda a)$

$$\int sigm(a)\,\mathcal{N}(a|\mu,\sigma^2)\,da\approx sigm(\kappa(\sigma^2)\mu),\quad \kappa(\sigma^2)=(1+\pi\sigma^2/8)^{-1/2}$$

• $p(y=1|D,x)\approx sigm(\kappa(\sigma_a^2)\mu_a)$ ← MODERATED OUTPUT: LESS XTREME THAN PLUG-IN, BUT SAME DECISION BOUNDARY

BAYESIAN LOGISTIC REGRESSION - OUTLIER DETECTION

- USUALLY: WITH RESIDUALS $R: y_i - \hat{y}_i$ SHOULD FOLLOW $N(0, \sigma^2)$, QQ-PLOT THEORETICAL VS EMPIRICAL QUANTILES
- BINARY DATA: STATISTICS NOT ASYMPTOTICALLY NORMAL. BAYESIAN → POINTS FOR WHICH $P(y|\hat{y})$ IS SMALL

  OUTLIERS: POINTS WITH LOW PROBABILITY UNDER X-VALIDATED POSTERIOR PREDICTIVE

$$P(y_i | x_i, x_{-i}, y_{-i}) = \int p(y_i|x_i, w) \prod_{i' \neq i} p(y_{i'}|x_{i'}, w) \, p(w) \, dw$$

ONLINE LEARNING AND OPTIMIZATION

- OFFLINE: $f(\theta) = \frac{1}{N} \sum^{N} f(\theta, z_i)$, $\underset{SUP}{z_i = (x_i, y_i)}$ OR $\underset{UNSUP}{z_i = x_i}$, $f(z_i, \theta) = $ LOSS, I.E. $-\log p(y_i|x_i, \theta)$ OR $L(y, h(x_i, \theta))$

- ONLINE: REGRET MINIMIZATION AVG LOSS RELATIVE TO THE BEST THAT COULD'VE GOTTEN IN HINDSIGHT WITH FIXED PARAMETERS

  REGRET: $\frac{1}{k} \sum_1^k f(\theta_t, z_t) - \min_{\theta \in \Theta} \frac{1}{k} \sum f(\theta_*, z_t)$ IS OBJECTIVE / LOSS

ONLINE GRADIENT DESCENT $\theta_{k+1} = \text{proj}_\Theta(\theta_k - \eta_k g_k)$ | $\text{proj}_V = \underset{w \in V}{\text{argmin}} \|w - v\|_2$ IS PROJECTION OF v ON V, $g_k = \nabla f(\theta_k z_k)$, η_k STEPSIZE

- ONLINE: MINIMIZE EXPECTED LOSS IN THE FUTURE <STOCHASTIC OPTIMIZATION> SOME VARS IN OBJECTIVE ARE RANDOM

$$f(\theta) = E[f(\theta, z)]$$

SGD: $\bar{\theta}_k = \frac{1}{k} \sum_1^k \theta_t$ (RUNNING AVERAGE) $\xrightarrow{\text{ONLINE}}$ $\bar{\theta}_k = \bar{\theta}_{k-1} - \frac{1}{k}(\bar{\theta}_{k-1} - \bar{\theta}_k)$ POLYAK-RUPPERT AVERAGING

HOW TO SET STEP SIZE: ROBBINS-MONRO CONDITIONS: $\sum \eta_k = \infty$, $\sum \eta_k^2 < \infty$ FOR CONVERGENCE

$$\eta_k = 1/k \quad \text{OR} \quad \eta_k = (\underset{>0}{\tau_0} + k)^{-\underset{(0.5, 1]}{\kappa}}$$

FORGET OLD VALUES CONTROL

HEURISTIC:
- TRY A RANGE OF η VALUES
- CHOOSE ONE W/ FASTEST DECREASE IN OBJECTIVE
- APPLY TO REST OF DATA

DRAWBACKS: MANUAL TUNING OF PARAMETERS

SAME η SIZE FOR ALL STEPS, SAME STEP FOR ALL PARAMS

[EARLY STOPPING: STOP AT PLATEAU, NOT NEC.LY CONVERGENCE]

ADAGRAD

ALIKE TO DIAGONAL HESSIAN APPROXIMATION, PER PARAMETER STEP VALUE ADAPTING TO CURVATURE OF LOSS FCN

$$\theta_i(k+1) = \theta_i(k) - \eta \frac{g_i(k)}{\tau_0 + \sqrt{s_i(k)}} \qquad s_i(k) = s_i(k-1) + g_i(k)^2$$

[SGD STEPS
- INIT θ, η
- REPEAT
  - PERMUTE DATA
  - $i = 1:N$
    - $g = \nabla f(\theta, z)$
    - $\theta \leftarrow \text{proj}_\Theta(\theta - \eta g)$
  - UPDATE η
- CONVERGENCE]

[EFFICIENT TO COMPUTE GRADIENT ON MINIBATCHES. $B=1 \to$ SGD, $B=N \to$ STEEPEST DESCENT]

BECAUSE TAKES FEW STEPS TO DETERMINE DIRECTION. AN UNCERTAINTY HELPS AVOID LOCAL MINIMA

- ADADELTA, RMSPROP, ADAM

PERCEPTRON

ONLINE BINARY LOGISTIC REGRESSION $\theta_n = \theta_{n-1} - \eta_n g_n = \theta_{n-1} - \eta_n(\mu_i - y_i)x_i \mid \mu_i = p(y=1 \mid x_i, \theta_i) = E[y_i \mid x_i, \theta_n]$

HAS SAME FORM AS **LMS** BECAUSE GENERALIZED LINEAR MODELS

$y_i = \text{ARGMAX } p(y \mid x_i, \theta)$, $\mu_i = \text{SIGM}(\theta^T, x)$, REPLACE W $\hat{y}$ $\quad g \approx (\hat{y}_i - y)x_i \rightarrow y \in \{-1, 1\} \rightarrow \hat{y} = \text{SIGN}(\theta^T x_i)$

UPDATE NO CHANGE IF CLASSIFICATION IS RIGHT, UPDATE ONLY IF WRONG $\quad \theta_n = \theta_{n-1} + \eta_n y_i x_i$

- CONVERGES IFF LINEARLY SEPARABLE DATA
- HISTORICALLY IMPORTANT BUT MORE MODERN ALGOS ARE BETTER

BAYESIAN ONLINE LEARNING

RECURSIVE APPLICATION OF BAYES RULE $\quad p(\theta \mid D_{1:n}) \propto p(D_n \mid \theta)\, p(\theta \mid D_{1:n-1})$

- RETURNS A FULL POSTERIOR
  ONLINE FRIENDLY FOR HYPERPARAMS
- 2ND ORDER MODELS ARE TRICKY ONLINE
- CAN BE QUICKER THAN SGD
- DIFFERENT RATE FOREACH PARAM.
  SIMPLE APPROXIMATION OF CURVATURE OF SPACE

EXAMPLES: KALMAN FILTER (ONLINE LINEAR REGRESSION ... CONVERGE TO OFFLINE (OPTIMAL) VALUE IN SINGLE PASS OVER THE DATA)
PARTICLE FILTER
DENSITY FILTER

ADA DELTA

- RESTRICTS ACCUMULATION WINDOW
- EXPONENTIAL MOVING AVERAGE $E[g^2] = \rho E[g^2]_{t-1} + (1-\rho) g^2 t \longrightarrow RMS[gt] = \sqrt{E[g^2]_t + \epsilon}$
- $\Delta x_t = \dfrac{-\eta}{RMS[gt]} \cdot gt$ 'UNIT NORMALIZATION' $\Delta x_t = \dfrac{-RMS[\Delta x]_{t-1}}{RMS[g]t} \cdot gt$

RMSPROP

- KEEP MA OF SQUARED GRADIENT FOR EACH WEIGHT
- $MS = 0.9\, MS(w, t-1) + 0.1 \left(\partial E / \partial w[t]\right)^2$
- $\Delta x_t = -\dfrac{gt}{\sqrt{MS(w,t)}}$

ADAM

- EXPONENTIAL DECAY RATES β_1, β_2. 1ST MOMENT m, 2ND MOMENT v.
- $m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$ EMA MEAN
- $v_t = \beta_2 v_{t-1} + (1-\beta_2) g^2 t$ EMA VAR $\rightarrow$ DIAG FISHER MATRIX APPROX | ADAGRAD IF $\beta_2 \rightarrow 1$, $\beta_1 \rightarrow 0$
- BIAS CORRECT: $\hat{m}_t = m_t/(1-\beta_1^t)$, $\hat{v}_t = v_t/(1-\beta_2^t)$
- $\Delta x = -\dfrac{\alpha\, \hat{m}_t}{(\sqrt{\hat{v}_t} + \epsilon)} \approx$ SNR-ISH
- ADAMAX: SCALES GRADIENTS PROPORTIONALLY TO INFINITY NORM