EXPONENTIAL FAMILY

FAMILY OF PROB. DISTRIBUTIONS WITH VERY COOL PROPERTIES. GAUSSIAN, BERNOULLI, STUDENT'S T

- ONLY FAMILY WITH FINITE SUFFICIENT STATISTICS $\xrightarrow{\text{UNDER REGULARITY}}$ MAKES LOSSLESS COMPRESSION FEASIBLE → REQUIRES SUPPORT NOT BE DEPENDANT FROM PARAMETERS
- ONLY FAMILY WITH CONJUGATE PRIORS
- LEAST SET OF ASSUMPTIONS UNDER CONSTRAINTS
- USED FOR GLM, VARIATIONAL INFERENCE

DEFINITION

A PDF / PMF IS IN EXP FAMILY IF

$$P(x|\theta) = \frac{1}{Z(\theta)} h(x) \exp\left[\theta^T \phi(x)\right] = h(x) \exp\left[\theta^T \phi(x) - A(\theta)\right]$$

$$Z(\theta) = \int h(x) \exp\left[\theta^T \phi(x)\right] dx \qquad A(\theta) = \log(Z(\theta))$$

$\phi(x)$ = VECTOR OF SUFFICIENT STATISTICS

$h(x)$ = SCALING, OFTEN = 1

- IF $\phi(x) = x$ → NATURAL EXPONENTIAL FAMILY

$Z(\theta)$ = PARTITION FUNCTION

$A(\theta)$ = LOG PARTITION FUNCTION / CUMULANT GENERATING FUNCTION

$\eta(\theta)$ = PARAMETERS → CANONICAL PARAMETERS

- IF $\dim(\theta) < \dim(\eta(\theta))$ CURVED EXPONENTIAL FAMILY
- IF $\eta(\theta) = \theta$ → CANONICAL FORM

BERNOULLI

$$\mathrm{Ber}(x|\mu) = \mu^x (1-\mu)^{1-x} = \exp\left[x \log(\mu) + (1-x)\log(1-\mu)\right] = \exp\left[\phi(x)^T \theta\right], \quad \phi(x) = [\mathbb{I}(x=0), \mathbb{I}(x=1)], \quad \theta = [\log(\mu), \log(1-\mu)]$$

IS OVERCOMPLETE, LINEAR DEPENDENCE

MINIMAL REPRESENTATION → UNIQUE θ ASSOCIATED

$$\mathrm{Ber}(x|\mu) = (1-\mu) \exp\left[x \cdot \log\left(\frac{\mu}{1-\mu}\right)\right] \qquad \phi(x) = x, \quad \theta = \log\left(\frac{\mu}{1-\mu}\right) \text{ LOG ODDS RATIO}, \quad Z = \frac{1}{1-\mu} \Big\} \text{ CANONICAL FORM}$$

$$\mu = \mathrm{sigm}(\theta) = \frac{1}{1+e^{-\theta}}$$

UNIVARIATE GAUSSIAN

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{Z(\theta)} \exp\left(\theta^T \phi(x)\right) \qquad \theta = \begin{pmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{pmatrix} \quad \phi(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix} \quad Z(\mu,\sigma^2) = \sqrt{2\pi}\sigma \exp\left[\frac{\mu^2}{2\sigma^2}\right]$$

CUMULANTS

DERIVATIVES OF $A(\theta)$ → CUMULANTS

$$\frac{dA}{d\theta} = \frac{d}{d\theta}\left(\log \int \exp(\theta\phi(x)) h(x) dx\right) = \int \phi(x) p(x) dx = E[\phi(x)]$$

$$\frac{d^2A}{d\theta} = \dots = E[\phi^2(x)] - E[\phi(x)]^2 = \mathrm{var}(x)$$

IF MULTIV: $\nabla^2 A(\theta) = \mathrm{cov}(\phi(x))$

→ $A(\theta)$ IS CONVEX

FOR BERNOULLI: $A(\theta) = \log(1+e^\theta)$, $\frac{dA}{d\theta} = \frac{e^\theta}{1+e^\theta} = \frac{1}{1+e^{-\theta}} = \mathrm{sigm}(\theta) = \mu$, $\frac{d^2A}{d\theta} = (1-\mu)\mu$

MLE

LIKELIHOOD: $P(D|\theta) = \left[\prod_i^N h(x_i)\right] g(\theta)^N \exp\left(\eta(\theta)^T \cdot \left|\sum^N \phi(x_i)\right|\right)$; $\phi_D = \left[\sum \phi_1(x_i) \dots \sum_N \phi_u(x_i)\right]$

LOG-LIKELIHOOD: $\log(P(D|\theta)) = \theta^T \phi(D) - NA(\theta)$. • $-A(\theta)$ IS CONCAVE, $\theta^T\phi(x)$ LINEAR $\rightarrow$ LOG LIKELIHOOD IS CONCAVE, UNIQUE GLOBAL M

$$\nabla_\theta \log p(D|\theta) = \phi(D) - N\,E[\phi(X)]$$

$E[\phi(X)] = \frac{1}{N}\sum \phi(x_i)$ • EMPIRICAL AVERAGE OF SUFF. STATS. MUST EQUAL THEORETICAL XPECTED SUFF. STATS,

METHODS OF MOMENTS (2)

IE FOR BERNOULLI: $E(\phi(x)) = P(X=1) = \hat{\mu} = \frac{1}{N}\sum^N \mathbb{I}(X=1)$

BAYESIAN FORMULATIONS

FOR CONJUGATE PRIOR TO MAKE SENSE $\rightarrow$ LIKELIHOOD MUST HAVE FINITE STATISTICS $\rightarrow$ ONLY EXP FAMILY

LIKELIHOOD: $P(D|\theta) \propto g(\theta)^N \exp(\eta(\theta)^T s_N)$, $s_N = \sum s_i$ | $P(D|\eta) \propto \exp(N\eta^T\bar{s} - NA(\theta))$

PRIOR: $P(\theta|\nu_0,\tau_0) \propto g(\theta)^{\nu_0} \exp(\eta(\theta)^T\tau_0)$ | $P(\eta|\nu_0,\tau_0) = \exp(\nu_0\eta_0{}^T\bar{\tau}_0 - \nu_0 A(\eta))$

POSTERIOR: $P(\theta|D) = P(\theta|\nu_N,\tau_N) = P(\theta|\nu_0+N, \tau_0+s_N)$ | $P(\eta|D) = P\left(\eta \,\middle|\, \nu_0+N, \frac{\nu_0\bar{\tau}_0 + N\bar{s}}{\nu_0+N}\right)$ CONVEX COMBINATION OF PRIOR MEAN HYPERPARAMS AND AVG OF SUFF. STATS

JUST UPDATES HYPER.PARAMS BY ADDITION

POSTERIOR PREDICTIVE: $P(D'|D) = \int P(D'|\theta)\,P(\theta|D)\,d\theta = \left[\prod^N h(\tilde{x}_i)\right] \frac{Z(\bar{\tau}_0 + \bar{s}(D) + \bar{s}(D'))}{Z(\bar{\tau}_0 + \bar{s}(D))}$

MAXIMUM ENTROPY DERIVATION

MAXIMUM ENTROPY PRINCIPLE: WE SHOULD PICK DISTRIBUTION WITH MAX ENTROPY (CLOSEST TO UNIFORM), IMPOSING THAT MOMENTS MATCH WITH EMPIRICAL MOMENTS OF FUNCTION

<SOME CALCULATIONS LATER> $P(x) = \frac{1}{Z}\exp\left(-\sum \lambda_u f_u(x)\right)$ **HAS FORM OF EXP FAMILY, GIBBS DISTRIBUTION**

GENERALIZED LINEAR MODELS

ANY MODEL WHERE OUTPUT DENSITY IS IN EXPONENTIAL FAMILY AND WHERE MEAN PARAMS ARE LINEAR COMBINATION OF INPUTS, POSSIBLY THROUGH A NONLINEARITY. LINEAR REGRESSION AND LOGISTIC REGRESSION ARE GLM.

BASICS

σ^2 IS DISPERSION PARAMETER, θ IS CANONICAL (NATURAL) PARAMETER, μ IS MEAN PARAMETER. TO GO FROM θ TO μ WE USE $\psi \rightarrow \theta = \psi(\mu)$, ψ IS INVERTIBLE

$\mu = \psi^{-1}(\theta) = A'(\theta)$

MEAN FUNCTION: $\eta_i = w^T x_i \longrightarrow \mu_i = g^{-1}(\eta_i) = g^{-1}(w^T x)$

LINK FUNCTION: $g()$, CAN PICK ANY AS LONG AS IT'S INVERTIBLE. IN LOGISTIC g^{-1} = SIGM

CANONICAL LINK FUNCTION: IF WE PICK $g = \psi$. I.E. IN BERNOULLI $g(\mu) = \log\left(\frac{\eta}{1-\eta}\right)$ LOGIT FCN, $\mu = \text{sigm}(\eta)$

- NORMAL $\longrightarrow$ IDENTITY $\longrightarrow \theta = \mu$
- POISSON $\longrightarrow$ LOG $\longrightarrow \theta = \log(\mu)$

ALSO, VERY VERY GENERALLY, $E[y|x_i w, \sigma^2] = \mu_i = A'(\theta)$

$\text{var}[y|x, w, \sigma^2] = \sigma_i^2 = A''(\theta)\sigma^2$

ML AND MAP

ALL GLM CAN BE FIT W/ SAME PROCEDURE AS LOGISTIC REGRESSION $\ell(w) = \log p(D|w) = \frac{1}{\sigma^2}\sum \ell_i$, $\ell_i = \theta_i y_i - A(\theta_i)$

GRADIENT WITH CHAIN RULE $\frac{d\ell_i}{dw_j} = \frac{d\ell}{d\theta} \cdot \frac{d\theta}{d\mu_i} \cdot \frac{d\mu_i}{d\eta} \cdot \frac{d\eta}{dw_j} =$ IF CANONICAL LINK $\theta_i = \eta_i$ $\nabla_w \ell(w) = \frac{1}{\sigma^2}\left[\sum^N (y - \mu_i) x_i\right]$

SUM OF INPUT VECTOR WEIGHED BY ERRORS.

- THEN WE CAN USE GRADIENT DESCENT ALGORITHM
- OR BETTER, A 2ND ORDER METHOD $H = -\frac{1}{\sigma^2} X^T S X$, S IS DIAGONAL WEIGHING MATRIX (CANONICAL LINK)
  - IF NON-CANONICAL LINK, USE EXPECTED HESSIAN, OR **FISHER INFORMATION MATRIX**, HAS SAME FORM OF H UNDER CANONICAL LINK
- **IF MAP:** INTRODUCE A GAUSSIAN PRIOR, LIKE ℓ^2 REGULARIZATION IN LOGISTIC REGRESSION

BAYESIAN INFERENCE: MCMC, VARIATIONAL INFERENCE, OR GAUSSIAN APPROXIMATIONS

PROBIT REGRESSION

$g^{-1}(\eta) = \Phi(\eta)$, ERF, CDF OF GAUSSIAN

- FIND GRADIENT, HESSIAN, STILL CLOSED FORM
- PLUG IN GRADIENT-BASED OPTIMIZER
- CAN BE INTERPRETED AS **RANDOM UTILITY MODEL (RUM)**

- ~~CA~~ SUITED TO **ORDINAL REGRESSION**, WHERE RESPONSE IS DISCRETE-VALUED WITH ORDER. MULTIPLE THRESHOLDS.
- **MULTINOMIAL PROBIT:** UNORDERED CATEGORICAL VALUES; MODELS C CORRELATED BINARY OUTCOMES

GLM 4 MULTI-TASK LEARNING

FIT MANY RELATED CLASSIFICATION MODELS. BETTER PERFORMANCE IF WE FIT ALL PARAMS AT SAME TIME. ALSO: TRANSFER LEARNING
VIA HIERARCHICAL BAYESIAN METHODS LEARNING TO LEARN

TIPS COLLABORATIVE FILTERING ; MANY GROUPS, MANY FEATURES. MAJORITY OF GROUPS HAVE LITTLE DATA, LONG TAILS.
CAN'T FIT SAME MODEL FOR ALL GROUPS, BUT CAN'T FIT EACH ONE SEPARATELY RELIABLY. WE ENCOURAGE PARAMS TO BE SIMILAR

- $E[y_{ij}|x_{ij}] = g(x_{ij}^T, \beta_j)$, $\beta_j \sim N(\beta_*, \sigma_j^2 I)$, $\beta_* \sim N(\mu, \sigma_*^2 I)$
- CAN MAP WITH STD GRADIENT METHODS

GROUPS WITH SMALLER SAMPLE SIZE BORROW PREDICTIVE STRENGTH FROM LARGER ONES VIA COMMON PARENT. β_*

- σ_j^2 CONTROLS HOW MUCH DEPENDANCE FROM C.P.
- σ_*^2 OVERALL PRIOR STRENGTH

- EXAMPLE: PERSONALIZED MAIL FILTERING $\rightarrow$ β_* FROM EVERYONE'S MAIL, β_j FROM SINGLE USER MAIL
- OTHER PRIORS (THAN GAUSSIAN) POSSIBLE, $\rightarrow$ SPARSITY INDUCING PRIOR ON β_j FOR MULTI-TASK FEATURE SELECTION (CONJOINT ANALYSIS)
- NEGATIVE TRANSFER: MULTITASK LEARNING FUCKS UP (WORSE) WHEN PARAMS ARE QUALITATIVELY DIFFERENT, WRONG INDUCTIVE BIAS ON PRIOR

GENERALIZED LINEAR MIXED MODELS

INFORMATION AT BOTH GROUP LEVEL AND ITEM LEVEL, STILL MULTITASK ; LIKE ANOVA A BIT.

$E[y_{ij}|x_{ij}, x_j] = g(\phi_1(x_{ij})^T \beta_j + \phi_2(x_{ij})^T \alpha)$ β_j RANDOM FX α FIXED FX • IF $P(y|x)$ GLM $\rightarrow$ GLMM

VARY RANDOMLY OR GROUPS

- CAN BE DIFFICULT TO FIT $\rightarrow$ $P(y_{ij}|\theta)$ MAY NOT BE CONJUGATE TO $P(\theta)$, TWO LEVELS OF UNKNOWNS θ AND $\eta = (\mu, \sigma)$ FOR PRIOR.
- FULL BAYESIAN INFERENCE: MCMC, VARIATIONAL BAYES OR EMPIRICAL BAYES, OR EXPECTATION - MINIMIZATION

RANKING

LEARNING TO RANK, LETOR PROBLEM. QUERY, DOCUMENTS, RELEVANCE, ETC...

- PROBABILISTIC LANGUAGE MODEL, BAG OF WORDS $sim(q, d) \triangleq p(q|d) = \prod p(q_i|d)$. q IS WORD, d IS DOCUMENT
- HAS TO BE SMOOTHED $p(t|d) = (1-\lambda)\frac{TF(t,d)}{LEN(d)} + \lambda p(t|\text{background})$ TF IS TERM FREQUENCY

— POINTWISE

FOR EACH QUERY - DOCUMENT PAIR DEFINE FEATURE VECTOR $x(q,d)$; LABELS (Y/N, OR ~~NO~~ ORDERED CATEGORICAL)

THEN WE $p(y=1|x(q,d))$ OR $p(y=r|x(q,d))$

- SIMPLE BUT NO CONSIDER LOCATION OF DOCS IN RESULT LIST.
- ERRORS AT TOP = ERRORS AT BEGINNING.

— PAIRWISE

- $p(y_{jk}|x(q,d_j), x(q,d_k))$ $y=1 \rightarrow REL(d_j,q) > REL(d_k,q)$ ELSE 0 BINARY CLASSIFIER
- $p(y_{jk}=1|x_j, x_k) = sigm(f(x_j) - f(x_k))$; F IS OFTEN LINEAR SCORING FUNCTION $f = w^T x$ RANKNET
- MLE OF w VIA MAX LL OR MIN CROSS ENTROPY LOSS, OPTIMIZE W/ GRADIENT DESCENT

— LISTWISE

- FULL CONTEXT FOR RELEVANCY
- DEFINE TOTAL ORDER ON LIST WITH PERMUTATION OF INDICES π . PLACKETT - LUCE DISTR: $p(\pi|s) = \prod_{1}^{M} \frac{s_j}{\sum_{M}^{u} s_u}$ $s_j = (\pi^{-1}(j))$ SCORE OF DOC AT j POSITION
- $\pi = (A,B,C)$ $\longrightarrow$ $p(\pi) = p(A=1)\,p(B=2|A=1)\,p(C=3|B=2,A=1)$
- INCORPORATE FEATURES VIA $s(d) = f(x(q,d))$ LINEAR $f = w^T x$ LISTNET
- MINIMIZE CROSS-ENTROPY $-\sum_i \sum_\pi p(\pi|y_i) \log p(\pi|s_i)$, INTRACTABLE $\rightarrow$ CONSIDER ONLY TOP K POSITIONS
- $K=1$ CROSSENTROPY TAKES $O(m)$ TIME
- IF ONLY 1 DOCUMENT IS RELEVANT $\longrightarrow$ CAN USE MULTINOMIAL LOGISTIC / SOFTMAX $p(y=c|x) = \frac{\exp(s_c)}{\sum_{1}^{M} \exp(s_{c'})}$ USED IN COLLAB FILTERING

ASSORTED LOSS FUNCTIONS FOR RANKING

- MEAN RECIPROCAL RANK (MRR) QUERY Q RANK OF FIRST $R(q)$ $MRR = 1/R(q)$
- MEAN AVG PRECISION (MAP) PRECISION $= P@K(\pi) =$ NUM RELEVANTS IN TOP K POS OF π / K , $AP(\pi) = \frac{\sum_K P@K(\pi)}{\text{NUM RELEVANTS}}$

  $MAP = \sum_q AP(\pi)/N_q$

- NORMALIZED DISCOUNTED CUMULATIVE GAIN (NDCG) RELEVANCE LABELS, MULTIPLE LEVELS

  $DCG@K(r) = r_1 + \sum_{2}^{K} \frac{r_i}{\log_2 i}$, r_i RELEVANCE

  - DCG VARIES WITH LENGTH OF LIST $\longrightarrow$ NORMALIZED WITH OPTIMAL ORDERING $IDCG@K = \arg\max DCG@K$
  - NDCG = DCG / IDCG

- RANK CORRELATION BETWEEN RANKED LIST π AND RELEVANCE JUDGEMENT π^* VIA I.E. WEIGHTED KENDALL τ STATISTIC
- WARP LOSS WEIGHTED APPROXIMATE PAIRWISE. BETTER THAN PRECISION @K TRANSFORMS INTEGER RANK TO REAL PENALTY
- LOSSES USED BAYESIANLY. FIT MODEL W/ POSTERIOR INFERENCE $\rightarrow$ THEN CHOSE ACTIONS TO MINIMIZE EXPECTED FUTURE LOSS. SAMPLE FROM POSTERIOR THEN AVG OVER θs FOR DIFFERENT THRESHOLDS
- FREQUENTISTLY MINIMIZE EMPIRICAL LOSS ON TRAINING SET, BUT NOT DIFFERENTIABLE $\longrightarrow$ USE GRADIENT FREE OPTIMIZATION OR SURROGATE LOSSES, I.E CROSSENTROPY