BAYESIAN STATISTIC

POSTERIOR $p(\theta|D)$

MAP $\operatorname{ARGMAX} p(\theta|D)$ IS POSTERIOR MODE, POPULAR BECAUSE REDUCES TO OPTIMIZATION PROBLEM. HAS DRAWBACKS

- IS POINT ESTIMATE, NO UNCERTAINTY
- NO REPARAMETRIZATION INVARIANT, MODE IS NOT CONSERVED, $\hat{\theta} = \operatorname{ARGMAX} \cdot p(D|\theta)p(\theta)|I(\theta)|^{-1/2}$ ← FISHER INFO OF $p(x|\theta)$
- MEASURE 0, NO TAKES SPACE INTO ACCOUNT ⟶ USE **LOSS FUNCTIONS**

$$0\text{-}1 \text{ LOSS} = I(\theta \neq \hat{\theta}) \rightarrow \text{MAP} = \text{POSTERIOR MODE}$$

$$\text{SQUARED LOSS} = (\theta - \hat{\theta})^2 \rightarrow \text{POSTERIOR MEAN}$$

$$|\theta - \hat{\theta}| \rightarrow \text{POSTERIOR MEDIAN}$$

- CREDIBLE INTERVALS → LIKE CONFIDENCE INTERVALS BUT BAYESIAN, CONTINUOUS REGION W/ $(1-\alpha)$ OF POSTERIOR PROBABILITY MASS **CENTRAL INTERVAL**

HPD / HDI : SET OF MOST PROBABLE POINTS $100 \cdot (1-\alpha)$ OF PROBABILITY MASS

BAYESIAN MODEL SELECTION

- CROSSVALIDATION REQUIRES FITTING EACH MODEL K TIMES, WAY INEFFICIENT, **LET'S COMPUTE POSTERIOR OVER MODELS**

$$p(m|D) = \frac{p(D|m)p(m)}{\sum_m p(m,D)} \rightarrow \hat{m} = \operatorname{ARGMAX} p(m|D) \xrightarrow[\text{PICK } m \text{ MAXIMIZING}]{\text{UNIF. PRIOR}} p(D|m) = \int p(D|\theta)p(\theta|m)\,d\theta$$

MARGINAL LIKELIHOOD, INTEGRATED LIKELIHOOD, EVIDENCE

- WE AREN'T MAXIMIZING PARAMS OR MLE, BUT INTEGRATING THEM AWAY / MARGINALIZING THEM, NO OVERFITTING B/C MOAR PARAMS $\neq$ BETTER MARGINAL LIKELIHOOD **'OCCAM'S RAZOR'**, SIMPLEST MODEL XPLAINING DATA
- NO RESTRICTED TO GRID OF VALUES, NUMERICAL OPTIMIZATION, **EMPIRICAL BAYES / TYPE II MAX. LIKELIHOOD** $\lambda^* = \operatorname{ARGMAX}_\lambda p(D|\lambda)$ ↳ POINT ESTIMATE OF POSTERIOR HYPERPARAMS
- WE NEED TO COMPUTE MARGINAL LIKELIHOOD $p(D|m)$ ⟶ EASY WHEN WE HAVE CONJUGATE PRIOR

PRIOR $p(\theta) = \frac{q(\theta)}{Z_0}$ ← NORM. CONSTANT **LIKELIHOOD** $p(D|\theta) = \frac{q(D|\theta)}{Z_L}$ **POSTERIOR** $p(\theta|D) = \frac{q(\theta|D)}{Z_N}$, $q(\theta|D) = q(D|\theta)q(\theta)$ | $\frac{q(\theta|D)}{Z_N} = \frac{q(D|\theta)q(\theta)}{Z_L Z_0 p(D)}$, $p(D) = \frac{Z_N}{Z_0 Z_L}$

- MARGINAL LIKELIHOOD CAN BE TOUGH ⟶ **BAYESIAN INFORMATION CRITERION**

$$\text{BIC} = \log p(D|\hat{\theta}) - \frac{\text{DOF}(\hat{\theta})}{2}\log N \propto \log p(D), \quad \hat{\theta} \text{ IS MLE}$$

AS PENALIZED LOG LIKELIHOOD

$$\text{BIC-COST} = -2\log p(D|\hat{\theta}) + \text{DOF}(\hat{\theta})\log N \approx -2\log p(D) \rightarrow \text{MINIMIZATION}$$

'MDL' PRINCIPLE, KOLMOGOROV COMPLEXITY.

- AKAIKE INFO CRITERION, ALTERNATIVE TO BIC → $\text{AIC} = \log p(D|\hat{\theta}_{MLE}) - \text{DOF}(m)$

- WHEN DOING MARGINAL LIKELIHOODS PRIOR IS IMPORTANT AS WE WEIGHT LIKELIHOODS W/ IT, IF WE DONT' KNOW PRIOR, WE PUT A PRIOR ON IT, HYPERPRIOR,

$$p(D|m) = \iint p(D|w)p(w|\alpha,m)p(\alpha,m)\,dw\,d\alpha$$

USUALLY HYPERPRIOR IS UNINFORMATIVE

USUALLY α IS OPTIMIZED AND NOT INTEGRATED OUT $p(D|m) = \int p(D|w)p(w|\hat{\alpha},m)\,dw$, $\hat{\alpha} = \operatorname{ARGMAX} p(D|\alpha,m) = \operatorname{ARGMAX}\int p(D|w)p(w|\alpha)\,dw$

EMPIRICAL BAYES!

BAYES FACTORS

IF PRIOR IS UNIFORM $\longrightarrow$ MODEL SELECTION IS PICKING MODEL WITH HIGHEST MARG. LIKELIHOOD

BAYES FACTOR IS RATIO OF ML $BF = \dfrac{P(D|M_1)}{P(D|M_0)} = \dfrac{P(M_1|D)}{P(M_0|D)} \Big/ \dfrac{P(M_1)}{P(M_0)}$

- FOR COMPARING TWO MODELS, NULL/ALTERNATIVE

- IT'S A LIKELIHOOD RATIO SO WE CAN COMPARE MODELS OF DIFFERENT COMPLEXITY
- BAYESIAN ALTERNATIVE TO P-VALUE
- JEFFREY-LINDLEY PARADOX! USE PROPER PRIORS (INTEGRATING TO 1) WHEN DOING MODEL SELECTION

PRIORS

- NO STRONG BELIEFS ABOUT $\theta \longrightarrow$ UNINFORMATIVE PRIORS, CAN BE NON-INTUITIVE TO FIGURE OUT; VALIDATE CONCLUSIONS VIA SENSITIVITY ANALYSIS

- JEFFREY'S PRIOR: TECHNIQUE FOR CREATING NON-INFORMATIVE PRIORS. ANY REPARAMETRIZATION OF $P(\phi)$ SHALL ALSO BE UNINFORMATIVE.

GENERALLY: $P(\theta) = P(\phi)\left|\dfrac{d\phi}{d\theta}\right|$, IF $P(\phi) \propto I(\phi)^{1/2}$ FISHER INFO. $I(\phi) = -E\left[\left(\dfrac{d \log P(X|\phi)}{d\phi}\right)^2\right]$ CURVATURE OF XPECTED LOG-LIKELIHOOD, MEASURE OF STABILITY OF MLE

$$I(\theta) = -E\left[\left(\frac{d \log P(X|\theta)}{d\theta}\right)^2\right] = I(\phi)\left|\frac{d\phi}{d\theta}\right|^2 \longrightarrow I(\theta)^{1/2} = I(\phi)^{1/2}\left|\frac{d\phi}{d\theta}\right|$$

TRANSFORMED PRIOR IS THE SAME

- JEFFREYS FOR MULTIPLE PARAMS: $P(\theta) \propto \sqrt{\det I(\theta)}$

- CAN ALSO BE MADE TRANSLATION INVARIANT OR SCALE INVARIANT, CAN BE IMPROPER AS LONG AS POSTERIOR IS PROPER

- ROBUST PRIOR! WE AREN'T CONFIDENT, NO MUCH INFLUENCE, HEAVY TAILS SO THINGS AREN'T CLOSE TO MEAN

- MIXTURES OF CONJUGATE PRIORS IS ALSO CONJUGATE, GOOD COMPROMISE BETWEEN COMPUTATIONAL PERF. AND FLEXIBILITY

$$P(\theta) = \sum_k \underbrace{P(z=k)}_{\text{WEIGHT}}\underbrace{P(\theta|z=k)}_{\text{CONJUGATE}}$$

- ALSO POSTERIOR CAN BE WRITTEN AS MIXTURE OF CONJUGATES

HIERARCHICAL BAYES

WHAT IF WE DON'T KNOW PRIOR? PRIOR ON PRIOR! HIERARCHICAL MODEL $\eta \rightarrow \theta \rightarrow D$, COMMON IN GRAPHICAL MODELS PARAMETER TYING; POOLED MLE.

EMPIRICAL BAYES

LIKE EARLIER

- MLE $\longrightarrow$ MAP $\longrightarrow$ MLE-II (EB) $\longrightarrow$ MAP II $\longrightarrow$ FULL BAYES

$\xrightarrow{\hspace{8em}}$ BAYESIANNESS

BAYESIAN DECISION THEORY

ACTION SPACE A, LOSS $L(y, a)$. HOW COMPATIBLE a IS WITH STATE y. **GOAL:** $\delta: X \to A$ DEVISE POLICY, OPTIMAL ACTION FOR EACH POSSIBLE INPUT

OPTIMAL = LOSS-MINIMIZING = $\delta(x) = \arg\min E[L(y,a)]$ || UTILITY: $U = -L$, MAXIMIZING $\delta(x) = \arg\max E[U(y,a)]$

MAX EXPECTED UTILITY, RATIONAL BEHAVIOR

POSTERIOR EXPECTED LOSS: $\rho(a|x) = E_{p(y|x)}[L(y,a)] = \sum L(y,a)\, p(y|x)$ (MINIMIZE) $\longrightarrow$ **BAYES ESTIMATOR** $\delta(x) = \arg\min \rho(a|x)$

— **0-1 LOSS** $L(y,a) = \mathbb{1}(y \neq a) \begin{cases} 0 & a = y \\ 1 & a \neq y \end{cases}$ $\rho(a|x) = p(a \neq y|x) = 1 - p(y|x)$ (MAXIMIZE). $y^* = \arg\max p(y|x)$ POSTERIOR MODE, MAP

- **REJECT OPTION** WHEN $p(y|x)$ IS VERY UNCERTAIN; RISK-AVERSE DOMAINS. $a = C+1$ IS REJECTION

$$L(y=j, a=i) = \begin{cases} 0 & i = j \\ \lambda_R & i = C+1 \\ \lambda_S & \text{else} \end{cases}$$

REJECT IF MOST PROBABLE CLASS HAS $p < 1 - \frac{\lambda_R}{\lambda_S}$

λ_R: COST REJECTION, λ_S = SUBST. ERROR

— **L_2 LOSS, SQUARED ERROR**

$L(y,a) = (y-a)^2$ $\rho(a|x) = E[(y-a)^2|x] = E[y^2|x] - 2a E[y|x] + a^2$

$\frac{\partial}{\partial a}\rho(a|x) \to \hat{y} = E[y|x] = \int y \cdot p(y|x)\, dy$ POSTERIOR MEAN, MIN. MEAN SQ. ERROR, MMSE

— **L_1 LOSS**

$L(y,a) = |y-a| \longrightarrow$ OPT. ESTIMATE IS POSTERIOR MEDIAN $P(y < a|x) = P(y \geq a|x) = 0.5$

— **IN SUPERVISED LEARNING** LOSS OF ACTION δ WHEN STATE IS θ = $L(\theta, \delta) = \sum_x \sum_y L(y, \delta(x))\, p(x,y|\theta)$ **GENERALIZATION ERROR**

$\longrightarrow \rho(\delta|D) = \int p(\theta|D)\, L(\theta,\delta)\, d\theta$

— **IN BINARY DECISION PROCESSES** WITH $F_{POS} | F_{NEG}$ COSTS

$\rho(\hat{y}=0|x) = L_{FN} \cdot p(y=1|x)$

$\rho(\hat{y}=1|x) = L_{FP}\, p(y=0|x)$

PICK 1 IFF $\frac{p(y=1|x)}{p(y=0|x)} > \frac{L_{FP}}{L_{FN}}$

— **CONFUSION MATRIX**

	$y = 1$	$y = 0$
$\hat{y} = 1$	TP	FP
$\hat{y} = 0$	FN	TN

$\hat{N}_+ = TP + FP$

$\hat{N}_- = FN + TN$

$N_+ = TP + FN$

$N_- = FP + TN$

- TP/N_+ = TPR, SENSITIVITY, RECALL
- $\frac{FN}{N_+}$ = FNR, MISS RATE, TYPE II
- FP/N_- = FPR, TYPE I
- TN/N_- = TNR, SPECIFICITY

- ON DIFFERENT THRESHOLDS τ TO PICK OPTIMAL

- ROC CURVES

PLOT TPR vs FPR FOR DIFFERENT τ, MEASURE USING **AUC** OR EQUAL ERROR RATE **EER** FPR = TPR @ τ

- PRECISION | RECALL CURVES

FOR RARE EVENTS ROC IS NOT MUCH INFORMATIVE

PRECISION = $TP/\hat{N}_+$

RECALL = TP/N_+

$$P = \frac{\sum y_i \hat{y}_i}{\sum \hat{y}_i} \qquad R = \frac{\sum y_i \hat{y}}{\sum y_i}$$

F-SCORE: $\frac{2PR}{R+P}$, HARMONIC MEAN

FOR SINGLE THRESHOLD

- FALSE DISCOVERY RATES

MANY BINARY DECISIONS $P(y_i = 1 | D) > \tau$, $D = \{x_i\}^N$, $N \gg$; IS MULTIPLE HYPOTHESIS TESTING

HOW TO SET τ? $\longrightarrow$ MINIMIZE FALSE POSITIVES $FD(\tau, D) = \sum_i \underbrace{(1 - P_i)}_{\text{ERR.}} \cdot \underbrace{I(P_i > \tau)}_{\text{DISCOVERY}}$ $\quad FDR(\tau, D) = \frac{FD(\tau, D)}{N(\tau, D)}$

CHOOSE τ