GAUSSIAN PROCESSES

FROM DISTRIBUTION OVER FCN — ARE SMOOTHERS

OPTIMAL SUPERVISED LEARNING: INFER DISTRIBUTION OF FUNCTIONS OVER DATA, USE IT TO PREDICTION; $P(f|D)$ VS $P(\theta|D)$

$$P(y|x_*, X, y) = \int P(y_*|f, x_*) P(f|X, y)\, df$$

- MAIN GP DRAWBACK: THEY TAKE $O(N^3)$ DUE TO K INVERSION/DECOMPOSITION

GP: DEFINES PRIOR OVER FCN → INTO POSTERIOR AFTER SEEING SOME DATA. DEFINE DISTRIBUTION OVER FCN VALUES AT FINITE SET OF POINTS $x_1 \dots x_N$: $P(f(x_1) \dots f(x_N))$ IS JOINTLY GAUSSIAN $\mu(x), \Sigma(x) = \sum_{ij} K(x_i, x_j)$ KERNEL FUNCTION. KERNEL SIMILAR → FCN OUTPUT SIMILAR. K IS COV

IN REGRESSION CLOSED FORM IN $O(N^3)$, IN CLASSIFICATION APPROXIMATIONS. BAYESIAN ALTERNATIVE TO KERNEL METHODS.

GRAPHICALLY f VALUES ARE HIDDEN NODES

GP FOR REGRESSION AND KRIGING

$$f(x) \sim GP(m(x), K(x,x'))$$, MEAN FCN $m(x) = E[f(x)]$, COV FCN/KERNEL $K(x,x') = E[(f(x) - m(x))(f'(x) - m'(x))^T]$ → JOINT GAUSSIAN → $P(f|X) = N(f|\mu, K)$

— NOISE-FREE

WE WANT SAME VALUES FOR ALREADY SEEN X_s, INTERPOLATION. $X_* \to f_*$

- JOINT: $\begin{pmatrix} f \\ f_* \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \\ \mu_* \end{pmatrix}, \begin{pmatrix} K & K_* \\ K_*^T & K_{**} \end{pmatrix}\right)$
- POSTERIOR $P(f_* | X_*, X, f) = N(f_* | \mu_*, \Sigma_*)$
  - $\mu_* = m(X_*) + K_*^T K^{-1}(f - m(X))$
  - $\Sigma_* = K_{**} - K_*^T K^{-1} K_*$

- IS SAMPLING FROM MVN OF DIM OF TRAINING SET
- MAXIMIZE MARGINAL LIKELIHOOD
- SUM OF SET OF BASIS FCN POSITIONED ON TRAINING DATA W/ PRIOR NET
- EM HAS 2 COV FCN

- SQUARE EXPONENTIAL KERNEL:

$$K(x,x') = \sigma^2 \exp\left(-\frac{1}{2\ell^2}(x - x')^2\right)$$

ℓ = HORIZONTAL SCALE, σ^2 = VERTICAL SCALE

— NOISY OBSERVATIONS

WE OBSERVE $y = f(x) + \varepsilon$, $\varepsilon \sim N(0, \sigma_y^2)$ • $COV[y|X] = K + \sigma_y^2 I_N = K_y$, DIAGONAL BECAUSE WE ASSUME INDEPENDENT NOISE.

- POSTERIOR

$$\mu_* = K_*^T K_y^{-1} y, \quad \Sigma_* = K_{**} - K^T K_y^{-1} K_*$$

- POSTERIOR MEAN: $\bar{f}_* = K_*^T K^{-1} y = \sum a_i K(x_i, x_*)$, $\alpha = K_y^{-1} \cdot y$

- PERFORMANCE DEPENDS EXCLUSIVELY ON KERNEL AND ITS PARAMS. IN RBF W NOISY ℓ, σ_f^2 ARE H-V SCALE AND σ_y^2 NOISE VARIANCE
- LARGE SCALE → IRRELEVANT DIMENSION

— IN MULTI-D SE IS: $K(x_p, x_q) = \sigma_f^2 \exp\left(-\frac{1}{2}(x_p - x_q)^T M (x_p - x_q)\right) + \sigma_y^2 \delta_{pq}$ • M IS ISOTROPIC $\ell^{-2} I$, $DIAG(\ell^{-2})$

OR $M = \Lambda\Lambda^T + DIAG(\ell^{-2})$ ← FACTOR ANALYSIS DISTANCE — EACH DIM W/ OWN SCALE

LOW RANK + DIAG APPROX ≈ FACTOR ANALYSIS — CHOLESKY

— KERNEL PARAMETER ESTIMATION

CAN GRID SEARCH BUT SLOW. EMPIRICAL BAYES, MAXIMIZE MARGINAL LIKELIHOOD; USE STANDARD GRADIENT OPTIMIZERS

- $P(y|x) = \int P(y|f,x) P(f|x)\, df \to \log P(y|x) = \log N(y|0, K_y) = -\frac{1}{2} y^T K_y^{-1} y - \frac{1}{2}\log|K_y| - \frac{N}{2}\log(2\pi)$

DATA FIT — MODEL COMPLEXITY — CONSTANT

↳ TRADEOFF

USED TO SOLVE LINEAR SYSTEMS

- VIA MARGINAL LIKELIHOOD AND GRADIENT DESCENT

- $\frac{\partial}{\partial \theta_j} \log P(y|X) = \frac{1}{2} TR\left[\left(\alpha\alpha^T - K_y^{-1}\right)\frac{\partial K_y}{\partial \theta_j}\right]$; $\alpha = K_y^{-1} y$ · $O(N^3)$ FOR K_y^{-1} AND THEN $O(n^2)$ PER HYPERPARAM

— ALTERNATIVE: BAYESIAN INFERENCE: COMPUTE POSTERIOR AND NOT POINT ESTIMATES $P(f|D) \propto \sum_s^S P(f|D, \theta_s) P(\theta_s|D)\, \delta_s$ BUT CURSE OF DIMENSIONS IF GRID AND HIGH D/M → MONTECARLO

→ CENTRAL COMPOSITE DESIGN PUT GRID POINT AT MODE AND AT ± 1SD ON EACH DIMENSION

— ALTERNATIVE: MULTIPLE KERNEL LEARNING · $K(x,x') = \sum_i w_i K_i(x,x')$, OPTIMIZE w_iS → DATA FUSION

- COMPUTATIONAL ISSUES: $\bar{f}_x$, STUPID TO DIRECTLY INVERT K_y. USE CHOLESKY DECOMPOSITION (1) $K_y = LL^T$ OR SOLVE $K_y \alpha = y$ WITH CONJUGATE GRADIENTS (2)
  - 1 $O(N^3)$ CHO, $O(N^2)$ FOR α, $O(N)$ MEAN, $O(N^2)$ VARIANCE
  - 2 $O(UN^2)$ FOR U ITERATIONS. $O(N^3)$ FOR FULL.

- SEMI-PARAMETRIC GP

  ADDS LINEAR MODEL FOR PROCESS MEAN $f(x) = \beta^T \phi(x) + R(x)$. $R(x) \sim GP(0, k, k')$ MODELS THE RESIDUALS

  IF $\beta \sim N(b, B) \rightarrow f(x) \sim GP(\phi(x)^T b, k(x,x') + \phi(x)^T B \phi(x))$ (INTEGRATE OUT PARAMS)

  - INTEGRATE β OUT FOR PREDICTIVE DISTRIBUTION FOR X
  - PREDICTIVE MEAN IS LINEAR MODEL OUTPUT + CORRECTION DUE TO GP
  - + COVARIANCE IS GP COV + UNCERTAINTY IN β

GP + GLM FOR CLASSIFICATION

MAIN PROBLEM: GAUSSIAN PRIOR NOT CONJUGATE TO BERNOULLI / MULTINOULLI LIKELIHOOD, APPROXIMATIONS: XPECT-PROPAGATION, VARIATIONAL, MCMC, GAUSSIAN

- BINARY CLASSIFICATION: $P(y_i | x_i) = \sigma(y_i f(x_i))$, $\sigma(z) = \text{SIGM}(z)$ LOGISTIC, OR $\Phi(z)$ PROBIT. $f \sim GP(0, k)$
  - POSTERIOR: $P(f | X, y) \sim N(\hat{f}, (k^{-1} + W)^{-1})$ - POSTERIOR PREDICTIVE: $\pi_x = P(y=1 | x_x, X, y) = \int \sigma(f_x) P(f_x | x_x, X, y) df_x$
  - MARGINAL LIKELIHOOD: $\log p(y|x) = \ell(\hat{f}) - \frac{1}{2} \log |H| + \text{const}$
  - DO NOT DIRECTLY INVERT K, L. USE CHOLESKY AN OTHER TRICKS. FITTING IS $O(N^3)$, PREDICTION IS $O(N^2 N_x)$ ALGO 15.2 P 524 MURPHY
- MULTI-CLASS: $P(y_i | x_i) = \text{CAT}(y_i | S(f_i))$, $f_i = (f_{i1}, f_{ic})$, $f_c = GP(0, K_c)$ ONE LATENT FCN PER CLASS. APRIORI INDEPENDENT, MAY DIFFERENT KERNELS
  - POSTERIOR: $O(CN^3)$
  - POST PREDICTIVE: $P(y | x_*, X, y) \approx \int \text{CAT}(y | S(f_*)) N(f_* | E[f_*], \text{cov}[f_*]) df_*$
  - MARGINAL LIKELIHOOD:
  - COMPLEXITY: FIT IN $O(TCN^3)$ TIME AND $O(CN^2)$ SPACE. PREDICTION IN $O(CN^3 + CN^2 N_*)$

- IS FINDING THE GP PRIOR OVER $f(x)$. (LATENT FCN) → PUT IT THROUGH LOGISTIC FUNCTION TO FIND PROB ON PREDICTED CLASS

GAUSSIAN PROCESSES VS THE WORLD

• ARMA, HMMS, MRF NETS, CRF ARE ALL GPS!!

- **LINEAR MODELS** EQUIVALENT TO GP WITH $k(x,x') = x^T \Sigma x'$. DEGENERATE COV FCN BECAUSE AT MOST D NON ZERO EIGENVALS
  UNDERFITTING (NOT FLEXIBLE ENOUGH) OVERCONFIDENT { PRIOR TO POOR → POSTERIOR TOO CONCENTRATED }

- **LINEAR SMOOTHERS**
  ARE STUFF OF THE FORM $f(x_*) = \sum_i w_i(x_*) y_i$. LINEAR FUNCTIONS OF TRAINING OUTPUTS. KERNEL REGRESSION, LOCALLY WEIGHTED REGRESSION ARE GPS BECAUSE POSTERIOR PREDICTIVE MEAN CAN BE EXPRESSED AS SUCH. $w_i(x) = [(K+\sigma^2 I_N)^{-1} k_x]_i$
  IN GPS KERNEL BANDWIDTH IS AUTO DECREASING WITH INCREASING N **DOF:** $\mathrm{TR}(K(K+\sigma^2 I)^{-1}) = \sum \frac{\lambda_i}{\lambda_i + \sigma_y^2}$ IF $\lambda_i \ll 1$ CORRESPONDING BASIS FUNCTION HAS LITTLE INFLUENCE

- **SVM** CAN REWRITE SVM OBJECTIVE IN FORM ANALOGOUS TO MAP ESTIMATES FOR GP. NO CAN FULLY CONVERT BECAUSE THERE IS NO LIKELIHOOD MATCHING HINGE-LOSS.

- **L1VM, RVM** EQUIVALENT TO GP WITH $k(x,x') = \sum_i^D \frac{1}{\alpha_i} \phi(x)\phi(x')$ DEGENERATE AND DEPENDS ON TRAINING DATA → OVERCONFIDENT

- **NEURAL NETS:** ARE A NONLINEAR GENERALIZATION OF GLM. IF NO. OF HIDDEN UNITS → ∞ AND ACTIVATION FCN IS BOUNDED WE GET GP CAMERIVE TRUE KERNEL FORMS ACCORDING TO GP.

- **SMOOTHING SPLINES:** NONPARAMETRIC MODELS USED TO INTERPOLATE 1D OR 2D DATA FIT f MINIMIZING DISCREPANCY OF DATA PLUS SMOOTHING PENALIZATION FOR WIGGLYNESS $J(f) = \sum (f(x_i) - y_i)^2 + \lambda \int \frac{d^m}{dx^m} f(x)^2 dx$ PIECEWISE POLYNOMIAL
  CAN BE FIT WITH RIDGE REGRESSION **REGRESSION SPLINE:** PLACES POLYNOMIALS AT FIXED SET OF K LOCATIONS. KNOTS.
  CHOOSING NO AND PLACE OF KNOTS IS LIKE PICKING SUPPORT VECTORS
  CUBIC SPLINE IS MAP OF $f(x) = \beta_0 + \beta_1 x + R(x)$ $R(x) \sim GP(0, \sigma_f^2 K(x,x'))$, $K(x,x') = \int_0^1 (x-u)_+ (x'-u)_+ du$ GENERALIZES TO HIGHER DIMS

- **RKHS:**
  REPRODUCING-KERNEL HILBERT SPACES. GENERALIZES NOTION OF SMOOTHNESS. CALC FCNS WRT KERNEL → LARGE NORMS BECAUSE MANY EIGENFUNCTIONS
  - **PROBLEM:** $J(f) = \frac{1}{2\sigma_y^2} \sum (y_i - f(x_i))^2 + \frac{1}{2}\|f\|_H^2$ → FCN HAS TO BE $f(x) = \sum \alpha_i k(x, x_i)$ → $\hat{f}(x_*)$ IS SAME AS POSTERIOR MEAN OF GP PREDICTIVE DISTRIBUTION

GP LATENT VARIABLE MODEL

COMBINES KERNELS WITH PROBABILISTIC PCA → GP-LVM • CAN FORMULATE PPCA IN DUALLY. MAXIMIZE Z AND INTEGRATE W OUT.

- LIKELIHOOD COMES OUT DEPENDING ON yy^T → CAN SOLVE WITH EIGENVALUE METHOD.
- NO MLE ANYMORE → GRADIENT BASED OPTIMIZERS
- LEARNS MAPPING FROM LATENT SPACE TO OBSERVED SPACE
- CAN MAKE SENSE TO USE FOR CLASSIFICATION OR VIZ.
- INPUTS TREATED AS LATENTS AND OPTIMIZED, PARAMS ARE INTEGRATED OUT
- IF LINEAR KERNEL → PCA BUT ALSO $K_2 = K + \sigma^2 I$ K IS GRAM MATRIX GENERAL K
- PRIOR IS ON W
- PCA IS GPLVM W/LINEAR KERNEL
- LOCATIONS OF POINTS IN LATENT SPACE IS DONE BY MAXIMIZING LIKELIHOOD WRT X

SPLINES