KERNELS

MEASURE SIMILARITY BETWEEN OBJECTS **KERNEL FUNCTION**: $K(x,x') \in \mathbb{R}$, TYPICALLY SYMMETRIC AND NON-NEGATIVE ; USED WHEN NO CLEAR HOW TO MAKE FEATURE VECTORS. IE STRINGS

- **GAUSSIAN / RBF KERNELS**

$$K = \exp\left(-\tfrac{1}{2}(x-x')^T \Sigma^{-1}(x-x')\right)$$ **SQUARED EXPONENTIAL** • σ^2 SCALE DIMENSION / BANDWIDTH

$$K = \exp\left(-\tfrac{1}{2}(x-x')\sum^{D} \tfrac{1}{\sigma_j^2}(x-x')\right)$$ **ARD IF $\sigma^2 \to \infty$**, Σ DIAGONAL

$$K = \exp\left(-\frac{\|x-x'\|}{2\sigma^2}\right)$$ **ISOTROPIC / SPHERICAL** • **RBF** BECAUSE ONLY FUNCTION OF $\|x-x'\|$

- **DOCUMENT COMPARISON**

$$K = \frac{x_i^T x_i'}{\|x_i\|_2 \|x'\|_2}$$ **COSINE SIMILARITY** ; ANGLE BETWEEN VECTORS $(0,1)$; SENSITIVE TO STOP-WORDS

$\longrightarrow$ **USE WITH TF-IDF TRANSFORM** $TF_{ij} = \log(1 + x_{ij})$ TERM FREQUENCY ; $IDF = \dfrac{N}{1 + \sum \mathbb{I}(x_{ij} > 0)}$ INVERSE DOCUMENT FREQUENCY ; $TF\text{-}IDF = [TF(x_{ij}) \times IDF(j)]_j^V$

- **MERCER / POSITIVE DEFINITE KERNELS**

$$\text{GRAM MATRIX} = K = \begin{bmatrix} K(x_1,x_1) & \cdots & K(x_1,x_M) \\ \vdots & & \vdots \\ K(x_N,x_1) & \cdots & K(x_N,x_M) \end{bmatrix}$$

- A KERNEL IS MERCER IFF GRAM MATRIX IS POSITIVE DEFINITE $\forall$ INPUT
- IF KERNEL MERCER $\exists\, \phi\; x \to \mathbb{R}^D$ $K(x,x') = \phi(x)^T\phi(x')$, ϕ DEPENDS ON EIGENFUNCTIONS OF K
- **POLYNOMIAL KERNEL** $K = (\gamma x^T x' + R)$, $R > 0$
- RBF AND COSIM ARE MERCER
- **SIGMOID** $K = \tanh(\gamma x^T x' + R)$ IS NOT MERCER
- IF K_1, K_2 MERCER $\to K_1 + K_2$ MERCER
- CAN GO KERNEL $\to$ FEATURE VECTOR

- **LINEAR KERNELS**

IF $\phi(x) = x$ $\quad K = x^T x'$. GOOD WHEN HIGH DIM AND ORIGINAL FEATURES ARE INFORMATIVE , LINEAR DECISION BOUNDARY WORKS, NO NEED TO TRANSFORM SPACE

- **MATERN KERNEL**

$R = \|x - x'\|$, $\nu > 0$, $R > 0$, K_ν BESSEL FUNCTION ; $\nu \to \infty$ SE KERNEL

$$K = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}R}{\ell}\right)^{\nu} K_\nu\left(\frac{\sqrt{2\nu}R}{\ell}\right)$$

- $\nu = 1/2 \longrightarrow K = \exp(-R/\ell)$. IF USED IN GAUSSIAN PROCESS : ORNSTEIN-UHLENBECK PROCESS DESCRIBES VELOCITY OF BROWNIAN MOTION PARTICLE

- **STRING KERNELS**

$W_s \geq 0$; A^* SET OF ALL POSSIBLE STRINGS IN ALPHABET (KLEENE STAR) , $\phi_s(x)$ NO OF TIMES SUBSTRING s APPEARS IN x

- $K = \sum_{s \in A^*} W_s \phi(x)\phi(x')$ • IS MERCER • $O(|x| + |x'|)$
- $W_s = 0$ FOR $|s| > 1 \longrightarrow$ BAG OF CHARACTERS
- s BORDERED W/ WHITESPACE $\longrightarrow$ BAG OF WORDS
- FIXED LENGTH $k \to$ **K-SPECTRUM**
- CAN GENERALIZE TO ~~PARSE~~ / COMMAND TREES

- **PYRAMID MATCH KERNELS**

USED FOR IMAGES, **SIFT**, FEATURES VECTOR XTRACTED FROM IMAGE POIS , VECTORIZED , BAG OF SYMBOLS . FEATURES MAPPED TO HISTOGRAM @ MULTI RESOLUTION COMPARED W/ INTERSECTION . APPROXIMATES OPTIMAL MATCH OF SIMILARITIES BETWEEN MATCHING POINTS . IS MERCER

PROBABILISTIC KERNELS

PROBABILISTIC GENERATIVE MODEL OF FEATURE VECTORS $p(x|\theta)$

– PROBABILITY PRODUCT KERNEL

$$k = \int p(x|x_i)^{\rho} p(x|x_j)^{\rho} dx \quad , \rho > 0 \, , \; p(x|x_i) \to p(x|\hat{\theta}(x_i)) \, , \; \hat{\theta} \text{ PARAM COMPUTED FROM SINGLE VECTOR}$$

- FITS MODEL TO SINGLE DATAPOINT ⟶ USES IT TO MEASURE SIMILARITY
- IF $p(x|\theta) = N(\mu, \sigma^2 I)$, σ FIXED, $\rho = 1$, ⟶ OBTAIN RBF

– FISHER KERNELS

$g(x) = \nabla_\theta \log p(x|\theta)|_{\hat{\theta}}$ ↳ LL GRADIENT, SCORE

$F = \nabla\nabla \log p(x|\theta)|_{\hat{\theta}}$ ↳ FISHER INFO MATRIX, HESSIAN

$\hat{\theta}$ ALL DATA.

- $k = g(x)^T F^{-1} g(x')$
- $g(x)$ DIRECTIONAL GRADIENT TO MAXIMIZE LIKELIHOOD
- KERNEL MEASURES SIMILARITY IF $g(x)$s ARE SIMILAR WRT GEOMETRY OF LIKELIHOOD FCN CURVATURE

KERNEL MACHINES

GLMs WHERE INPUT FEATURE VECTOR IS $\phi(x) = [k(x, \mu_1) \dots k(x, \mu_k)]$ μ_k = CENTROIDS

↳ KERNELIZED FEATURE VECTOR

- THEN USE KFV FOR LOGISTIC $p(y|x,\theta) = \text{Ber}(w^T \phi(x))$ OR LINEAR $p(y|x,\theta) = N(w^T\phi(x))$ REGRESSION.
- OK FOR NON LINEARLY SEPARABLE BOUNDARIES
- BANDWIDTH AFFECTS FREQUENCY, RANGE
- HOW TO PICK μ_k?

  IF FEW DIM → TILE THE SPACE UNIFORMLY

  IF DIM HIGH → NUMERICAL OPTIMIZATION / MCMC

IDEA: FIND CLUSTERS IN DATA AND ASSIGN ONE PROTOTYPE PER CLUSTER

BUT MOST DENSITY OF POINTS $\neq$ MOST USEFUL

BUT I STILL HAVE TO PICK NO. OF CLUSTERS!

IDEA #2: MAKE EACH EXAMPLE A PROTOTYPE!

$\phi(x) = [k(x, x_1), \dots, k(x, x_N)] \to D = N$ ⟶

- SPARSITY PRIOR ON W ⟶ SPARSITY VECTOR MACHINE, I.E. ℓ_1VM, OR GROUP LASSO FOR MULTICLASS
- ℓ_2VM → OFC NOT SPARSE
- ARD/SBL → RELEVANT VECTOR MACHINE
- SUPPORT VECTOR MACHINE ⟶ MODIFIES LIKELIHOOD TERM! NO SPARSITY PRIOR

TEH KERNEL TRICK!

NOT DEFINE KERNELIZED FEATURE VECTOR ⟶ WORK WITH ORIGINAL BUT REPLACE INNER PRODUCTS $\langle x \, x' \rangle$ WITH $k(x, x')$. ONLY IF K IS MERCER

$\langle x, y \rangle = x^T y = \sum_i^N x_i y_i$

- KERNELIZED NN CLASSIFICATION $||x - x_i||_2^2 = \langle x_i, x_i \rangle + \langle x_i', x_i' \rangle - 2\langle x_i, x_i' \rangle$
- KERNELIZED K-MEDOIDS CLUSTERING: EACH CENTROID IS ONE OF DATA VECTORS, NOT ARBITRARY POINTS

  WHEN UPDATE MEASURES SUM DISTANCE OF ALL CLUSTER OBJECTS TO ALL OTHER IN CLUSTER; PICK ONE WITH LOWEST $\mu_k = \text{argmin} \sum d(i, i')$

  – $O(n_k^2)$ PER CLUSTER VS $O(n_k D)$ OF K-MEANS – CAN TURN INTO CLASSIFIER – CAN KERNELIZE $d(i, i') = ||x - x_i||_2^2$

- KERNELIZED RIDGE REGRESSION

  – PRIMAL PROBLEM: $w = (\sum x_i x_i^T + \lambda I_D)^{-1} X^T y \xrightarrow{\text{MATRIX INVERSION LEMMA}} w = X^T(XX^T - \lambda I_N)^{-1} y$

  – DUAL PROBLEM $\alpha = (K + \lambda I_N)^{-1} y$ ⟶ $w = X^T \alpha = \sum \alpha_i x_i$

  K ↳ XX^T = GRAM MATRIX

  - SOLUTION IS LINEAR COMBINATION/SUM OF TRAINING VECTORS
  - PREDICTION: $\hat{f}(x) = w^T x \sum \alpha_i x x^T = \sum \alpha_i k(x, x')$

  – COMPUTING α IS $O(N^3)$ → w IS $O(D^3)$ SO USEFUL IN HIGH DIM JUST BECAUSE SPEED; PREDICTION WITH α IS $O(ND)$, WITH w IS $O(D)$

  ↳ SPEED UP WITH SPARSITY

KERNEL PCA

PCA BY FINDING EIGENVECTORS OF XX^T AND KERNEL TRICK.

USEFUL EVEN W/O KERNEL BECAUSE IS $D \times D$ VS $N \times N$

$V = \sum_1^N \alpha_i \phi(x_i)$ — V EVECS NORMAL PROBLEM, α_i EVECS KERNEL PROBLEM

- U, Λ EIGENVECTOR, VALUES OF XX^T $\longrightarrow$ EIGENVECTORS OF $X^TX = V = X^TU$ $\xrightarrow{\text{NORMALIZE}}$ $V_{PCA} = X^T U \Lambda^{-1/2}$
- $K = XX^T$, Φ = NOTIONAL DESIGN MATRIX IN FEATURE SPACE $\longrightarrow$ $V_{KPCA}: \Phi^T U \Lambda^{-1/2}$ CANNOT COMPUTE! POSSIBLY ∞ DIM $\longrightarrow$
- CAN PROJECT x_* IN FEATURE SPACE: $\phi_*^T V_{KPCA} = \phi_*^T \Phi^T U \Lambda^{-1/2} = k_*^T U \Lambda^{-1/2}$, $k_* = [k(x_*, x_1) \dots k(x_*, x_N)]$
- PROJECTED DATA NOT CENTERED, CANNOT SIMPLY SUBTRACT MEAN $\longrightarrow$ $\tilde{\phi}_i = \phi(x_i) - \frac{1}{N}\sum \phi(x_i)$ CENTERED FEATURE VECTOR $\longrightarrow$ $\tilde{K} = HKH$

  $H = 1 - \frac{1}{N} 1_N 1_N^T$ CENTERING MATRIX
- VIA $K(x, x')$ WE IMPLICITLY REPLACE x WITH $\phi(x)$
- KPCA CAN DO UP TO N COMPONENTS
- USEFUL FOR CLASSIFICATION, NOT SO MUCH FOR VISUALIZATION

ALGO:
- PRODUCE $K = K_{ij}(k(i,j))$, NORMALIZE + CENTER
- COMPUTE EVECS EVALS OF K
- RETAIN TOP EVECS; NORMALIZE BY SQRT OF EVALS
- USE THEM TO PROJECT BACK DATA

SUPPORT VECTOR MACHINES

$J(w, \lambda) = \sum L(y_i, \hat{y}) + \lambda \|w\|^2$ IS OF EMPIRICAL RISK

KMMS: PROBABILISTICALLY UNNATURAL,
SPARSITY IN LOSS AND NOT IN PRIOR
ENCODE KERNEL VIA ALGORITHMIC TRICK
DO NOT RESULT IN PROBABILISTIC OUTPUTS

IDEA: WE CAN REPLACE LOSS TO SOMETHING THAT WILL ENFORCE SPARSITY SO THAT PREDICTIONS ONLY DEPEND ON SUBSET OF TRAINING DATA
$\longrightarrow$ SUPPORT VECTORS: POINTS FOR WHICH ERROR IS OUTSIDE TUBE
- KERNEL TRICK + MODIFIED LOSS
- MINIMIZE CLASSIF. ERROR, MAXIMIZE GEOMETRIC MARGIN

FOR REGRESSION:

- EPSILON INSENSITIVE LOSS FUNCTION: $L_\epsilon(y, \hat{y}) = \begin{cases} 0 & \text{IF } |y - \hat{y}| < \epsilon \\ |y - \hat{y}| - \epsilon & \text{ELSE} \end{cases}$
  - ANY POINT OUTSIDE ϵ-TUBE IS PENALIZED L ON PREDICTION
  - NOT DIFFERENTIABLE: USE SLACK VARIABLES TO REPRESENT DEGREE OF POINT OUTSIDE TUBE. ALLOW MISCLASSIFICATION — SOFT MARGIN

    $y_i \leq f(x) + \epsilon + \xi^+$

    $y_i \geq f(x_i) + \epsilon - \xi^-$
  - $J = C\sum^N (\xi^+ + \xi^-) + \frac{1}{2}\|w\|$ LOSS / OBJECTIVE
  - C IS INVERSE REG. COEFFICIENT
- OPTIMAL SOLUTION $\hat{w} = \sum \alpha_i x_i$, α IS SPARSE BECAUSE DON'T CARE ABOUT ERRORS $< \epsilon$
- X FOR WHICH $\alpha_i > 0$ $\rightarrow$ SUPPORT VECTORS
- PREDICTIONS: $\hat{y}(x) = \hat{w}_0 + \sum_i \alpha_i k(x_i, x)$

FOR CLASSIFICATION:

- HINGE LOSS: $L_{HINGE}(y, \eta) = \max(0, 1 - y\eta) = (1 - y\eta)_+$ $\quad \eta = f(x)$ = 'CONFIDENCE' OF PREDICTION. NO PROBABILISTIC. QUADRATIC PROBLEM / ALGO
- OBJECTIVE: $\min \frac{1}{2}\|w\|^2 + C\sum^N \xi_i \; : \; \xi_i \geq 0$ SOLVED IN $O(N^3)$, OPTIM TO $O(N^2)$, C REGULARIZERS
- X INCORRECTLY CLASSIFIED $\rightarrow$ SUPPORT VECTORS
- PREDICTION: $\hat{y}(x) = \text{SGN}\left(\hat{w}_0 + \sum \alpha_i k(x_i, x)\right)$ $O(SD)$ TO COMPUTE

MARGIN PRINCIPLE: SVM MAXIMIZES PERPENDICULAR DISTANCE OF BOUNDARY TO CLOSEST POINT, MARGIN. LARGE MARGIN CLASSIFIER. SOFT MARGIN CONSTRAINTS WITH ξ
$\sum \xi_i$ IS UPPER BOUND ON MISCLASSIFICATIONS AT TRAINING TIME. $C = 1/\nu N$ CONTROLS THIS FRACTION $\rightarrow$ ν-SVM

PROBABILISTIC OUTPUT: NO BUILT-IN SENSE, CAN INTERPRET VIA LOG-ODDS RATIO. BUT RESULTS NOT WELL CALIBRATED

MULTI-CLASS CLASSIFICATION: OUTPUT NOT CALIBRATED $\rightarrow$ CANNOT INTO SOFTMAX. ONE-VS-REST: C CLASSIFIERS, EACH ON $C^i, (1 - C^i)$: PROBLEMATIC CLASS IMBALANCE
ONE VS ONE: $C(C-1)/2$ CLASSIFIERS, ON EVERY PAIR, CLASSIFY ON HIGHEST VOTE COUNT.

PICKING C: RECOMMENDED PICK IT WITH CV OVER 2d GRID VALUES. ALTERNATIVELY = LASSO / LARS ALGO. START WITH LARGE λ

PROBABILISTIC INTERPRETATION: WITH A LOT OF RELAXATIONS WE CAN INTERPRET HINGE-LOSS AS GAUSSIAN SCALE MIXTURE $\rightarrow$ CAN USE BAYESIAN TO SET HYPERPARAMS

PERFORMANCE: ALL KERNEL METHODS SIMILAR ACCURACY OVER RANGE OF PROBLEMS AND SAME KERNEL. GP L2VM ARE $O(N^3)$ SVM IS $O(N)$ WHEN LINEAR KERNEL
RVM USUALLY SLOWER THAN L1VM BUT GREEDY TRAINED IS FASTER. IF SPEED MATTERS: USE RVM
IF CALIBRATED PROBABILITY MATTERS: USE GP (GAUSSIAN PROCESS)

SMOOTHING KERNELS

USED FOR NONPARAMETRIC DENSITY ESTIMATES. UNSUPERVISED DENSITY ESTIMATION OR GENERATIVE CLASSIF/REGRESSION MODELS

- $\int k(x)\,dx = 1$, $\int x\,k(x)\,dx = 0$, $\int x^2 k(x)\,dx > 0$

GAUSSIAN/RBF: $k_h(x) = \frac{1}{h^D (2\pi)^{D/2}} \prod \exp\left(-\frac{1}{2h^2} x_j^2\right)$ H IS BANDWIDTH, 'WIDTH' OF KERNEL, σ^2

- UNBOUNDED SUPPORT

EPANECHNIKOV: $k(x) = \frac{3}{4}(1-x^2)\,\mathbb{I}(|x| \le 1)$

- COMPACT SUPPORT
- NONDIFF AT ITS BOUNDARY

TRI-CUBE: $k(x) = \frac{70}{81}(1-|x|^3)\,\mathbb{I}(|x| \le 1)$

- COMPACT
- TWICE DIFFERENTIABLE AT ITS BOUNDARY

BOXCAR: UNIFORM DISTRIBUTION $k(x) = \mathbb{I}(|x| \le 1)$

KERNEL DENSITY ESTIMATION (KDE)

ALTERNATIVE TO PARAMETRIC DENSITY ESTIMATION WITH GMM $\rightarrow$ NEED TO SPECIFY K, MUS

HERE WE JUST SET A CLUSTER CENTER PER DATAPOINT $\mu_i = x_i$ KERNEL DENSITY/PARZEN WINDOW ESTIMATOR: $\hat{p}(x) = \frac{1}{N}\sum k_h(x - x_i)$

- NO MODEL FITTING $\rightarrow$ JUST SET H WITH CV, NO NEED TO PICK K + (CV OR DIRICHLET PROCESS MIXTURE, BAYESIAN)
- LOT OF MEMORY TO STORE, USELESS FOR CLUSTERING TASKS −
- CV MINIMIZE FREQUENTIST RISK

EXAMPLES: BOXCAR $\rightarrow$ HISTOGRAM COUNT
(1D) RBF $\rightarrow$ SMOOTHER HISTO

KNN/CLASSIFIER

WE GROW VOLUME AROUND X UNTIL K DATAPOINTS REGARDLESS OF CLASS ~~VOLUME~~ LABEL. $V(x)$, $N_c(x)$ SAMPLES OF CLASS C

- CLASS-COND. DENSITY $p(x|y=c, D) = \frac{N_c(x)}{N_c V(x)}$, N_c TOTAL COUNT FOR C
- CLASS PRIOR: $p(y=c|D) = \frac{N_c}{N}$
- CLASS POSTERIOR: $p(y=c|D) = \frac{N_c(x)}{K}$

KERNEL REGRESSION

COMPUTE CONDITIONAL EXPECTATION $f(x) = E(y|x) = \int y \cdot p(y|x)\,dy$, KDE APPROX $p(x,y) = \frac{1}{N}\sum^N k_h(x-x_i)\,k_h(y-y_i)$. ... ... SUM TO ONE ZERO MEAN ...

$$f(x) = \sum^N w_i(x)\,y_i$$

$$w_i(x) = \frac{k_h(x - x_i)}{\sum_N k_h(x - x_i)}$$

- PREDICTION IS WEIGHTED SUM OF OUTPUTS AT TRAINING POINTS, WEIGHT DEPEND ON SIMILARITY TO TRAINING POINTS
- ONLY 1 FREE PARAMETER
- OPTIMAL $H = \left(\frac{4}{3N}\right)^{1/5}\hat{\sigma}$

LOCALLY-WEIGHTED REGRESSION:

KERNEL REGRESSION FITS A CONSTANT FUNCTION LOCALLY. WE CAN IMPROVE BY FITTING A LINEAR REGRESSION MODEL FOR EACH POINT X

$\min \sum^N k(x_i, x_*)[y_i - \beta(x_*)^T \phi(x_i)]^2$, $\phi(x) = [1, x]$, $\beta(x_*) = (\Phi^T D(x_*)\Phi)^{-1}\Phi^T D(x_*)\,y$, Φ DESIGN MATRIX, $D = \mathrm{DIAG}(k(x_*, x))$

- $\hat{f}(x_*) = \phi(x_*)^T \beta(x_*) = \sum w_i(x_*)\,y_i$ PREDICTION, $w_i(x_*)$ IS EQUIVALENT KERNEL!