MACHINE LEARNING

- **SUPERVISED** → CLASSIFICATION / REGRESSION | CLASS: Y IS CATEGORICAL, REGR: Y IS REAL, CONTINUOUS
- **UNSUPERVISED** → CLUSTERING, DENSITY ESTIMATION, DIMENSIONALITY REDUCTION, GRAPH STRUCTURE DISCOVERY, IMPUTATION, LOCAL FEATURES
- **REINFORCEMENT** → REINFORCEMENT

KNN ALGORITHM

CLASSIFIER

$$P(Y=C \mid X, D, K) = \frac{1}{K} \sum_{i \in N_K(x, D)} I(y_i = C)$$

I IS INDICATOR

- LOOKS AT K NEAREST POINTS (TRAINING) CLOSEST TO TEST INPUT. RETURNS EMPIRICAL FRACTION AS ESTIMATE
- IF $K=1$ → VORONOI TESSELATION

- **CURSE OF DIMENSIONALITY** → DIMENSIONS GROW LINEARLY, ~~SIZE~~ VOLUME OF SAMPLE SPACE GROWS EXPONENTIALLY
- **SOLUTION!** LET'S MAKE ASSUMPTIONS ABOUT THE MODEL, INDUCTIVE BIAS (PARAMETERS IN PARAMETRIC MODEL)
- **CROSS VALIDATION**
  - K-FOLDS: $K-1$ AS TRAINING, 1 AS TEST, AVG CLASSIF. ERROR ON FOLDS, • $K=1$: LEAVE-ONE-OUT CV

INFORMATION THEORY 4 DUMMIES

ENTROPY: $H(x) = -\sum_{k}^{k} P(X=k) \cdot \log_2 P(X=k)$ BINARY/BERNOULLI ENTROPY: $H(x) = -[p \log p + (1-p) \log_2 (1-p)]$

- **KL DIVERGENCE:** MEASURES DISSIMILARITY OF PROBABILITY DISTRIBUTIONS, N° OF EXTRA BITS WE TO USING q INSTEAD OF p

$$KL(p \| q) = \sum_{n}^{n} p_n \log \frac{p_n}{q_n} = \sum p_n \log p_n - \sum p_n \log q_n = -H(p) + H(p,q)$$

$H(p,q)$ = CROSS-ENTROPY $= -\sum p_n \log q_n$ → • N° BITS TO ENCODE MSG FROM p IF WE USE MODEL q

$KL(p \| q) \geq 0$

- **UNIFORM** IS DISCRETE DISTRIBUTION WITH MAX ENTROPY → LAPLACE'S PRINCIPLE OF INSUFFICIENT REASON CHOOSE UNIFORM WHEN NO PRIORS
- **MUTUAL INFORMATION** HOW SIMILAR JOINT DISTRIBUTION IS TO FACTORED DISTRIBUTION
  - $I(x,y) = KL(P(x,y) \| P(x)P(y)) = \sum^{x} \sum^{y} P(x,y) \cdot \log \frac{P(x,y)}{P(x)P(y)}$ • $I(x,y) = 0$ IFF $P(x,y) = P(x)P(y)$ = INDEPENDENCE
  - $I(x,y) = H(x) - H(x|y) = H(y) - H(y|x)$, $H(x|y)$ CONDITIONAL ENTROPY $= \sum_x P(x) H(y \mid X=x)$
  - MI IS REDUCTION IN UNCERTAINTY FOR X AFTER HAVING OBSERVED Y, OR VICEVERSA
  - POINTWISE MI (PMI) = FOR EVENTS, NOT DISTRIBUTIONS $= \log \frac{P(x|y)}{P(x)} = \log \frac{P(y|x)}{P(y)}$
  - ~~MUTUAL~~ MAXIMAL INFORMATION COEFFICIENT (MIC): FOR CONTINUOUS DISTRIBUTIONS, DISCRETIZE 2D GRIDS; BIN THE DATA, TRY DIFF. BIN SIZES

$$m(x,y) = \frac{\max_G I(x(G), y(G))}{\log \min(x,y)}$$ MIC: $\max [m(x,y)]$ FOR ALL SIZES | 0 = NO RELATIONSHIP; 1 = NOISE-FREE RELATIONSHIP, NOT JUST LINEAR