7 PERCEPTRON

- McCulloch-Pitts Neuron $y = \Theta(\underline{w}\cdot\underline{x} - T)$

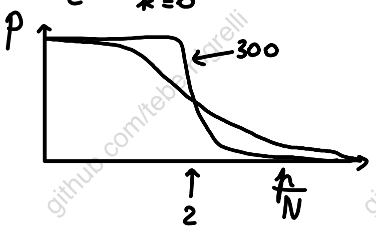
- FEEDFORWARD/RECURRENT
  (PATTERN RECOGNITION) (SHORT-TERM MEMORY)

- PERCEPTRON LEARNING
  ALGORITHM:
  FOR WRONG CLASSIFICATIONS OF $\vec{x}^\mu$
  $\vec{w}_{t+1} = \vec{w}_t + \eta \vec{x}^\mu$
  → CONVERGENCE IN $t < \frac{K^2}{\varepsilon^2}$ (PROOF)

COVER FORMULA (P FINDING w 100% CORRECT)

$$P = \frac{1}{2^{p-1}} \sum_{k=0}^{N-1} \binom{p-1}{k}, \quad \binom{p-1}{k} = 0 \text{ IF } k > p-1$$

(PROOF)



COVER GENERALISATIONS
→ SIGN CONSTRAINED ⇒ CAPACITY $p = N$

- COVER LIMITATIONS
  → SLOWER LEARNING IF DATA IS UNBIASED
  → NO WAY TO QUANTIFY ROBUSTNESS
  → LEARNING SHOULD BE ROBUST AGAINST NOISE

8 SUPPORT VECTOR MACHINES

LINEAR SEPARABILITY ASSUMPTION

MAXIMISE DISTANCE BETWEEN POINTS:
PICK w ST.
$$\begin{cases} w^T x_+ + b = 1 \\ w^T x_- + b = -1 \end{cases} \quad \text{FOR } x_+, x_- \text{ SUPPORT VECTORS}$$

OPTIMISATION PROBLEM:
$$\max_w \frac{2}{\|w\|} \quad \text{(MARGIN)}$$
$$\text{ST. } \begin{cases} w^T x_n \geq 1 & \text{IF } y_n = +1 \\ w^T x_n \leq -1 & \text{IF } y_n = -1 \end{cases}$$

OR
$$\min \|w\|^2 \text{ ST. } y_n(w^T x_n + b) \geq 1 \quad \text{(USE GRAD. DESCENT)}$$

TRADEOFF: MARGIN/# MISTAKES

SOFT MARGIN (ALLOWS FOR SOME MISTAKES)
$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{n=1}^{N} \xi_n \leftarrow \text{SLACK PER VARIABLE}$$
$$\text{ST. } y_n(\underline{w}\cdot\underline{x}_n + b) \geq 1 - \xi_n$$
$$\xi_n \geq 0$$

C REG. PARAMETER: LARGER C MEANS MORE SEVERE PENALISATION, SO NARROWER MARGIN

OR $f(x) = \underline{w}\cdot\underline{x} + b$
$$\begin{cases} y_n f(x_n) \geq 1 - \xi_n \\ \xi_n \geq 0 \end{cases}$$

$\xi_n = \max(0, 1 - y_n f(x_n))$
JUST MINIMISE:
$$\min_w \underbrace{\tfrac{1}{2}\|w\|^2}_{\text{REGULARISATION TERM}} + C \underbrace{\sum_{n=1}^{N} \max(0, 1 - y_n f(x_n))}_{\text{HINGE LOSS}}$$

THE PROBLEM IS CONVEX ($\sum$ CONV = CONV)

HINGE LOSS
$$\ell(t) = \max\{0, \overset{\text{ONLY IF WRONG}}{1 - t}\}$$
$$t = y f(\underline{x}) = y(\underline{w}\cdot\underline{x} + b)$$

ZERO-ONE LOSS → CANNOT BE DIFFERENTIATED, DIFFICULT TO MINIMISE

GRADIENT DESCENT:
$$w_{t+1} = w_t - \eta_t \nabla_w \underbrace{C(w_t)}_{g(w) \text{ TO BE MINIMISED}}$$
$$= w_t - \eta \frac{1}{N} \sum_{n=1}^{N} \left(\lambda w_t + \nabla_w \mathcal{L}(x_n, y_n, \underline{w}_t)\right)$$

SGD USES MINI BATCHES INSTEAD OF ALL THE DATA
SINGLE LOSS OVER DATA, MODEL

DUE TO HINGE LOSS, ONLY POINTS STRICTLY VIOLATING THE MARGIN CONTRIBUTE TO THE GRADIENT

9 Support Vector Machines II

PRIMAL SVM ⟺ DUAL SVM

PRIMAL SVM: PARAMETERS INCREASE WITH # FEATURES
↓
USE IF MORE DATA THAN FEATURES

DUAL SVM: PARAMETERS INCREASE WITH # DATA
↓
USE IF MORE FEATURES THAN DATA

TYPICAL APPLICATION OF SVM IN REAL LIFE

NONLINEAR SVM → MORE FLEXIBLE
↓
DISCRIMINANT FUNCTION: "LINEARITY" IS KEPT

$$f(x) = \begin{cases} 1 & \hat{w}\cdot\phi(x)+b>0 \\ -1 & \text{ELSE} \end{cases}$$

SHORTCUT TO APPLY THE NONLINEARITY WHILE SKIPPING SOME COMPUTATIONS AND ONLY OBTAINING THE $f(x)\cdot f(y) = ?$ INSTEAD OF EVAL. f TWICE → • MORE FLEXIBLE • LESS COMPUTE

$\phi(x)$ IS QUICKER TO COMPUTE BECAUSE IT SKIPS INTERMEDIATE HIGH-DIM STEPS

$$k(x_1,x_2) = \phi(x_1)\cdot\phi(x_2)$$

WANT TO DESCRIBE — REARRANGE IN TERMS OF ϕ

USING CONVEX DUALITY,

$$w^* = \sum_{n=1}^{N} \alpha_n y_n x_n$$

→ REPRESENTER THEOREM: SOLUTION TO OPTIMISATION IS IN SPAN OF DATA

KERNEL FUNCTION:
$k: X\times X \to \mathbb{R}$
• $k(x,y) = k(y,x)$
• POS-DEFINIT. ⟹ ENSURES CONVEXITY

KERNEL TRICK: APPLY NON-LIN TO DATA, THEN APPLY LINEAR SVM

KERNEL PROPERTY IS KEPT.
• $\lambda \geq 0 \to \lambda k_1$
• $k_1 + k_2$
• k_1^n, $\prod k_i$

SAMPLE KERNELS
• POLYN, DEG $= d \to K(x,y) = (x\cdot y)^d$
DEG $\leq d \to (1+x\cdot y)^d$
• RADIAL BASIS FUNCTIONS $\to \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$

10 Recurrent NN

ALLOW FOR MEMORY OPERATIONS

TYPES OF NEURONS: ± 1; $\{0,1\}$; $1\ldots N$

$$S_i(t+1) = \phi\left(\sum_j w_{ij} S_j(t) + I_i(t)\right)$$

OFFSET DEPENDS ON i, t
W MATRIX OF WEIGHTS FOR CROSS-BEHAVIOUR

PARALLEL / ASYNCHRONOUS UPDATES

DISCRETE / CONTINUOUS TIME → $r_i(t+1) = \ldots$

$$\tau \frac{dr_i}{dt} = -r_i + \ldots$$

ϕ ACTIVATION FUNCTION (SIGMOID, RELU)

SYMMETRIC NETWORK $(J_{ij} = J_{ji})$ → ENERGY FUNCTION (LYAPUNOV) : $-\frac{1}{2}\sum_{j\neq i} J_{ij} S_i S_j = E(S_1 \ldots S_N)$

⇓ AS "TEMP. DECREASES" AKA CONVERGENCE

$$P(S_1 \ldots S_h) = \frac{1}{Z}\exp(-\beta E(S_1 \ldots S_N))$$

11 Hopfield Model

DESCRIPTION: N BINARY (± 1) NEURONS,

$$S_i(t+1) = \text{SGN}\left(\sum_j J_{ij} S_j(t)\right)$$

↓
NETWORK CONNECTS INPUTS AND OUTPUTS, MUST EMBED SOME LOGIC IN THE LIMIT STATE

P PATTERNS TO MEMORISE $(\xi_i^\mu = \pm 1)$

HEBBIAN MATRIX : $J_{ij} = \frac{1}{N}\sum_\mu \xi_i^\mu \xi_j^\mu$

NOISE, CROSS-INTERACTIONS

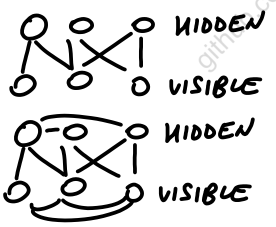
PATTERN COMPLETION : IF THERE IS AN OVERLAP WITH A STORED PATTERN, THE NETWORK CONVERGES TO IT IN ONE STEP

12 Restricted Boltzmann Machines

• TRAIN MACHINES TO LEARN THE DISTRIBUTION OF THE DATA FEATURES

• RESTRICTED → CONNECTIONS ONLY BETWEEN VISIBLE ⟷ NON VISIBLE

• BOLTZMANN MACHINE → SAME TYPE NEURONS ARE CONNECTED



PHASES:
- LEARN : VISIBLE $= x$, MUST LEARN TO GENERATE x WITH HIGH PROB.
- SAMPLE: FIXED WEIGHTS, DEFINE ENERGY f, NATURAL CONVERGENCE TO LOWEST VALUE