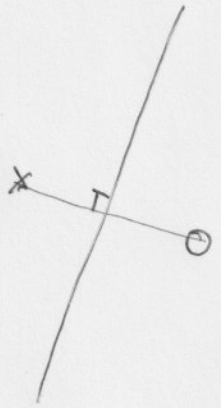
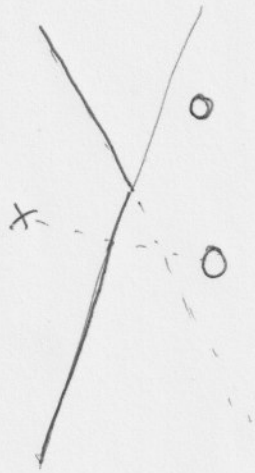
18.1



2 points
Tar bisector



18.2 $h = w^T x$ $\quad_{32\times32\times3}$ $\quad y$ = target $+1, -1$

fit by L2 $\quad w^* = \arg\min_w e = |h - y|^2 = |w^T x - y|^2$

$$\frac{\partial e}{\partial w} = (w^T x - y)\, x = 0$$

$$x\, x^T w - y x = 0$$

for multiple x_i's $\quad \sum_i x_i x_i^T w - \sum_i y_i x_i = 0$

let $X = \begin{pmatrix} \cdots x_1^T \cdots \\ \cdots x_2^T \cdots \end{pmatrix}$ $\quad X^T X w - X^T Y = 0$

$$w = (X^T X)^{-1} X^T Y$$

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \end{pmatrix}$$

18.3 $h = W^T x$

↑ Matrix

scores per class → $\begin{pmatrix} s_1 \\ s_2 \\ s_3 \\ \vdots \end{pmatrix} = \begin{pmatrix} w_1^T \\ w_2^T \\ w_3^T \\ \vdots \end{pmatrix} \underline{x}$

one vs all : want $s_i = 1$ if class of x is i

0 otherwise

18.4 $y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 \ldots$

solve for $a_0, a_1 \cdots a_3$

18.5 prevent overfitting by regularization

L2 penalty on weights / polynomial coeffs

$e = |y - M\underline{a}|^2 + \lambda |\underline{a}|^2$

18.6 Non linear opt.

$h - t \simeq J \Delta W + r$ ← num data eqns

num parameters ← 10^6 +

18.7

$$e = \frac{1}{2}|w^T x - t|^2$$

$$\frac{\partial e}{\partial w} = (w^T x - t)\, x^T$$

gradient descent $\quad w_{t+1} = w_t - \alpha \underbrace{\frac{\partial e}{\partial w}}_{=\nabla w}$

$$w_{t+1} = w_t - \alpha\,(w^T x - t)\, x^T$$

α ↑ learning rate.

18.8

$$V_{t+1} = \rho V_t + \nabla w_t \qquad \rho = 0.9 - 0.99$$

$$w_{t+1} = w_t - \alpha V_{t+1}$$

Improved versions Ada grad, rms prop, Adam