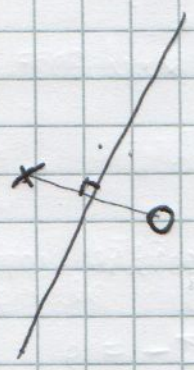
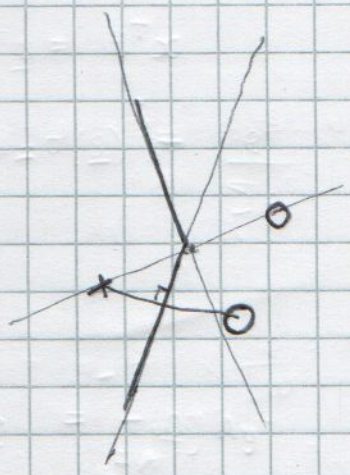
6.1



2 points
1 ar bisector



add more bisectors..

6.1 a

want $|x_i - w_j|^2$ is $\begin{cases} \text{small}, & c_i = j \\ \text{large}, & c_i \neq j \end{cases}$

$$|x_i - w_j|^2 = |x_i|^2 + |w_j|^2 - 2 w_j^T x_i$$

suppose $|x_i|^2 = |w_j|^2 = 1$

Then $w_j^T x_i = \begin{cases} \text{large}, & c_i = j \\ \text{small}, & c_i \neq j \end{cases}$

6.2 $h = \underline{w}^T \underline{x}$

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

$\frac{\partial e}{\partial w} = 2(w^T x - y)\, x$

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

many x's $\sum_i x_i x_i^T w - \sum_i y_i x_i = 0$

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

$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \end{pmatrix}$ $\quad \underline{w = (X^T X)^{-1} X^T Y}$

w vector same size as data x

$h = \underline{w}^T \underline{x} + b_0$ $\quad \underline{h} = w^T X + \underline{b}$ $\quad \begin{bmatrix} \underline{x} \\ 1 \end{bmatrix}$

3072×1 vectors; b_0: bias

6.3 $h = w^T x$

$\begin{pmatrix} s_1 \\ s_2 \\ s_3 \\ \vdots \end{pmatrix} = \begin{pmatrix} w_1^T \\ w_2^T \\ w_3^T \\ \vdots \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \end{pmatrix}$ $\quad s_i = w_i^T x$

w_i are "templates" for class i

6.4 $y = a_0 + a_1 x + a_2 x^2 + a_3 x^3$

y_i, x_i

4 unknowns, solve for a_0, a_1, a_2, a_3 using 4 points

N^{th} order poly $N+1$ points

6.5 $y = Ma$

L2 loss $e = |y - Ma|^2$

regularization $e = |y - Ma|^2 + \lambda |a|^2$

6.7 $e = \frac{1}{2} |w^T x - t|^2$

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

update rule $w_{t+1} = w_t - \alpha \frac{\partial e}{\partial w}$ $= \nabla w$

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

learning rate

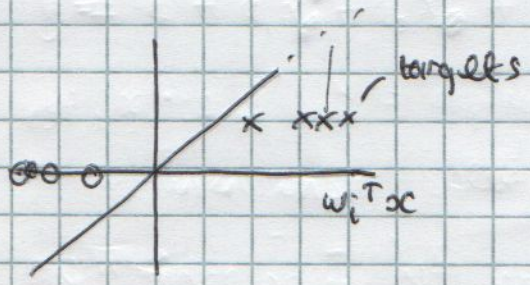
6.8 SGD + momentum

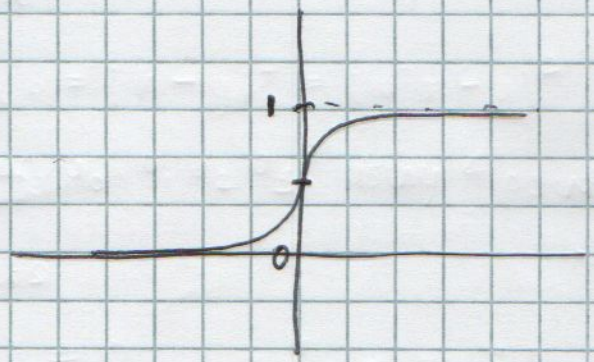
$v_{t+1} = \rho v_t + \nabla w_t$ $\rho = 0.9 - 0.99$

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

Improved versions Ada grad, RMS prop, Adam.

6.9





$$h(s_i) = \frac{1}{1+e^{-s}} \quad \text{logistic}$$

$$\underline{s} = w^T \underline{x}$$

$$\underline{h}(\underline{s}) = \frac{e^{\underline{s}}}{\sum_i e^{s_i}}$$

~~6.10~~

6.10

define scores $\underline{s} = W^T x$

softmax predictor $h = \sigma(\underline{s}) = \frac{e^{\underline{s}}}{\sum_i e^{s_i}}$ $\quad \sigma_i = \frac{e^{s_i}}{\sum_j e^{s_j}}$

cross entropy loss

$$e = -\sum_i t_i \log h_i = -\log h_y$$

$$\frac{\partial e}{\partial \underline{s}} = -\frac{1}{h_y} \frac{\partial}{\partial \underline{s}} h_y = -\frac{1}{h_y} \frac{\partial}{\partial \underline{s}} \frac{e^{s_y}}{\sum_i e^{s_i}}$$

$$\frac{\partial}{\partial s_k} \frac{e^{s_y}}{\sum_i e^{s_i}} = \begin{cases} \dfrac{-e^{s_y}}{\left(\sum_i e^{s_i}\right)^2} e^{s_k} & k \neq y \\ \dfrac{e^{s_y}}{\sum_i e^{s_i}} - \dfrac{e^{s_y} e^{s_k}}{\left(\sum_i e^{s_i}\right)^2}, & k = y \end{cases}$$

$$= \begin{cases} -h_y h_k, & k \neq y \\ h_y - h_y^2, & k = y \end{cases}$$

$$\frac{\partial e}{\partial \underline{s}} = -\frac{1}{h_y} \frac{\partial}{\partial s} h_y = \begin{cases} h_k, & k \neq y \\ -1 + h_k, & k = y \end{cases}$$

$$\frac{\partial e}{\partial \underline{s}} = \underline{h} - \underline{t}$$

where $\underline{h} = \sigma(W^T x)$ $\quad s = W^T x$

e is cross entropy loss.

$$\frac{\partial e}{\partial W} = \frac{\partial e}{\partial s} \frac{\partial s}{\partial W} = (\underline{h} - \underline{t}) x^T$$