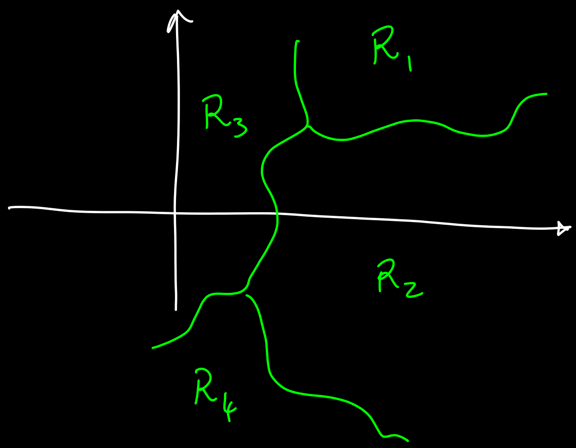
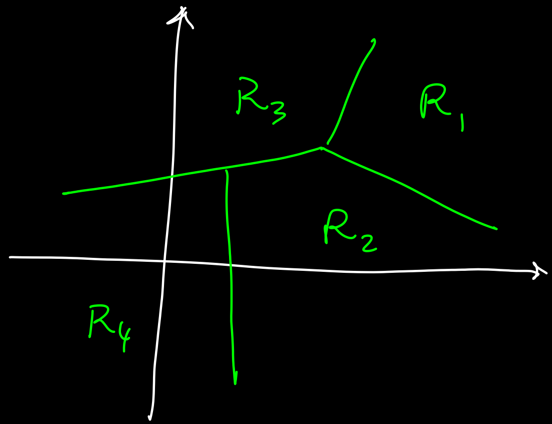
2022-09-16 Linear and Probabilistic Models for Classification

Decision regions

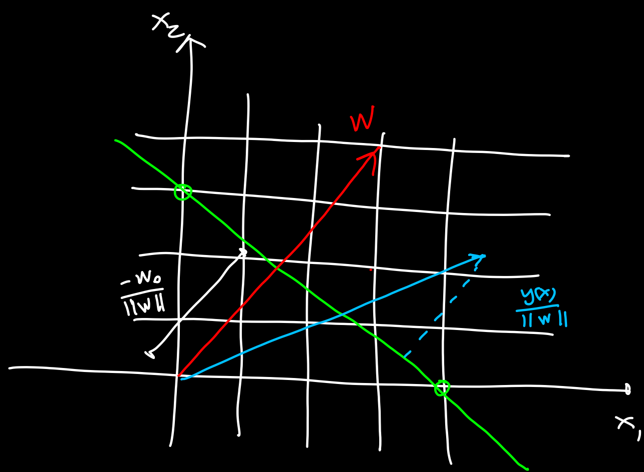


Decision regions (linear models)



Exampe: 2 classes

$$\begin{cases} y(x) = w^T x + w_0 \\ x \in R_1 \text{ iff } y(x) \geq 0 \\ x \in R_2 \text{ iff } y(x) < 0 \end{cases} \quad (\text{sign}[y(x)])$$



Example

$\overbrace{w_0 = -12, \; w_1 = 3, \; w_2 = 4}^{w}$

decision boundry : (x_1, x_2) s.t. $y(x) = 0$

$$3x_1 + 4x_2 - 12 = 0$$

$$\|w\| = \sqrt{3^2 + 4^2} = \sqrt{25} = 5$$

$$\frac{-w_0}{\|w\|} = \frac{12}{5} \qquad \frac{y(x)}{\|w\|}$$

Least squares

check derivations on the book

Fisher's linear discriminant

linear classification viewed as dimensionality reduction

$y = w^T x$ from D to 1 dimension

we can select the projection that maximizes class separation

Example 2-class

$C_1, N_1, \quad C_2, N_2$

$$\underline{m}_1 = \frac{1}{N_1} \sum_{n \in C_1} x_n \qquad \underline{m}_2 = \frac{1}{N_2} \sum_{n \in C_2} x_n$$

simple measure of separation:

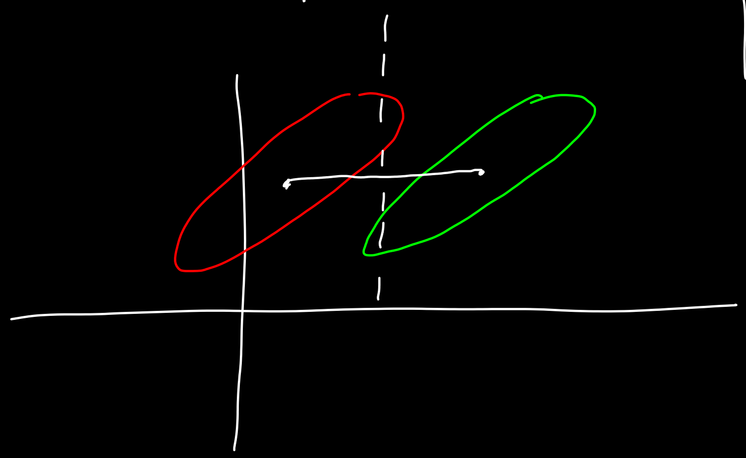
$$m_2 - m_1 = w^T (\underline{m}_2 - \underline{m}_1)$$

but this can be arbitrarily large (increasing $\|w\|$)

we can constrain $\sum_i w_i = 1$

Fig 4.6

Problem:



Idea: 1) maximize distance of projected means

2) minimize within-class projected variance

$$s_k^2 = \sum_{n \in C_k} (y_n - m_k)^2$$

$$y(x_n) = w^T x_n$$

total within-class variance: $s_1^2 + s_2^2$

$$J(w) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2} \qquad \text{Fisher's criterion}$$

$$= \frac{(w^T \underline{m}_2 - w^T \underline{m}_1)^2}{\sum_{n \in C_1} (w^T x_n - w^T \underline{m}_1)^2 - \sum_{n \in C_2} (w^T x_n - w^T \underline{m}_2)^2} =$$

$$w^T (x_n - \underline{m}_1) \left[w^T (x_n - \underline{m}_1) \right]^T$$

$$w^T (x_n - \underline{m}_1)(x_n - \underline{m}_1)^T w$$

$$= \frac{w^T (\underline{m}_2 - \underline{m}_1)(\underline{m}_2 - \underline{m}_1)^T w}{w^T \left[\sum_{n \in C_1} (x_n - \underline{m}_1)(x_n - \underline{m}_1)^T + \sum_{n \in C_2} (x_n - \underline{m}_2)(x_n - \underline{m}_2)^T \right] w}$$

$$= \frac{w^T S_B w}{w^T S_W w}$$

$S_B \leftarrow$ between-class covariance

$S_W \leftarrow$ within-class covariance

differentiating $\Rightarrow$ $w \propto S_W^{-1} (\underline{m}_2 - \underline{m}_1)$

$y = w^T x$ is roughly gaussian because of the central limit theorem.

least-squares $\rightarrow$ minimize error w.r.t. t

fisher $\longrightarrow$ maximize class separation in output space

If the target for class C_1 is $\frac{N}{N_1}$ (reciprocal of prior)

C_2 is $-\frac{N}{N_2}$

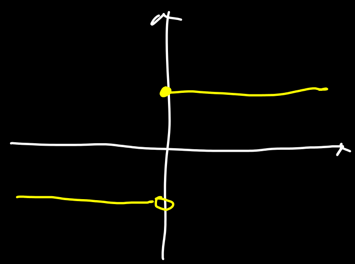
$\Rightarrow$ least squares $\equiv$ fisher

Perception

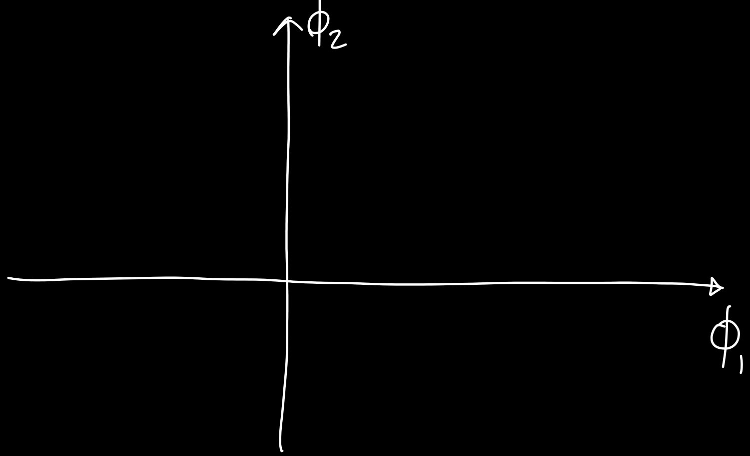
Rosenblatt 1962

$y(x) = f(w^T \phi(x))$

$$f(a) = \begin{cases} +1, & a \geq 0 \\ -1, & a < 0 \end{cases}$$



optimization of w through error minimization



$t \in \{-1, +1\}$

$w^T \phi(x_n) > 0 \rightarrow C_1$

$w^T \phi(x_n) < 0 \rightarrow C_2$

we would like $w^T \phi(x_n) t_n > 0$ for all patterns

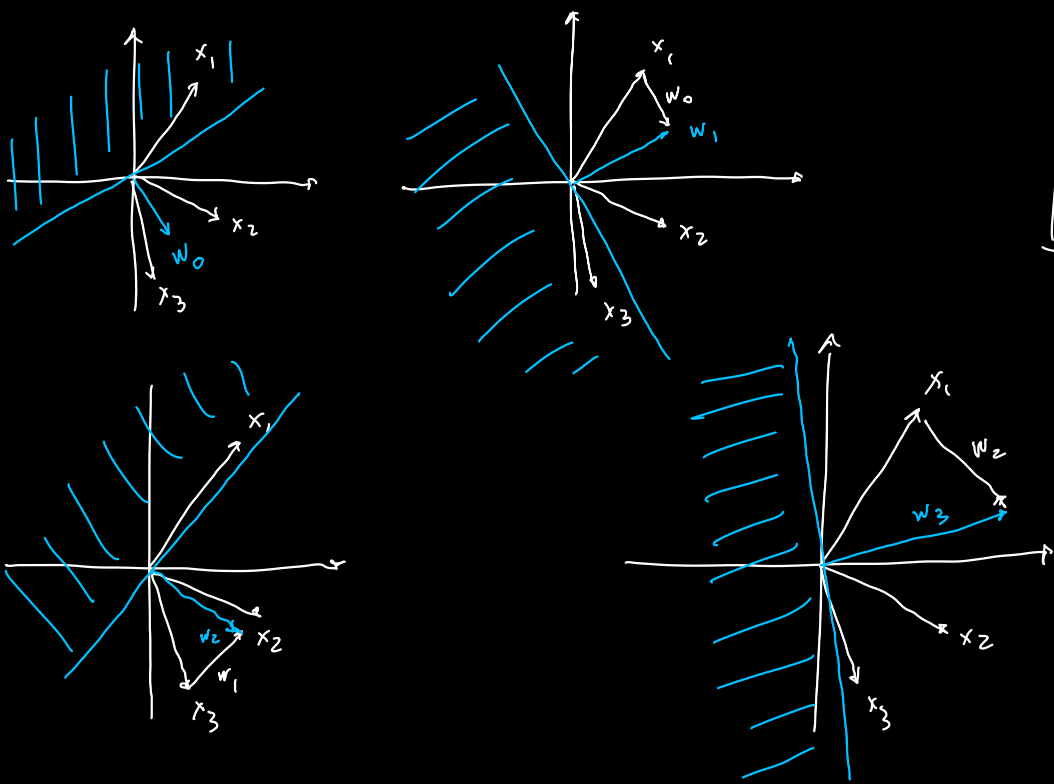
$E_P(w) = -\sum_{n \in \mathcal{M}} w^T \phi(x_n) t_n$ $\quad \mathcal{M}$ = misclassified

Stochastic gradient descent

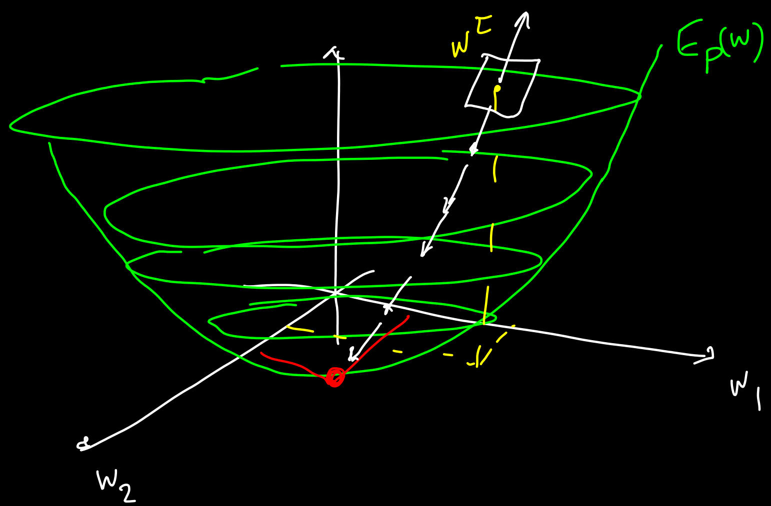
$$w^{(\tau+1)} = w^{(\tau)} - \underbrace{\eta}_{\text{learning rate}} \nabla E_p(w) = w^{(\tau)} + \eta \phi(x_n) t_n$$

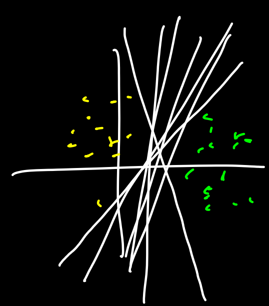
because $w *$ constant does not change the solution, we can set $\eta = 1$

not guaranteed to reduce error at each step



Fig





Probabilistic models

Maximum Likelihood

(D uda makes class independence an assumption)

In general we would like to maximize

$$\ln \mathcal{L} = \sum_{n=1}^{N} \ln p(x_n, t_n \mid \theta) \quad \dots \quad \left[p(x_n, t_n \mid \theta) = p(x_n \mid t_n, \theta)\, p(t_n \mid \theta) \right]$$

$$= \sum_{n=1}^{N} \left[\ln p(x_n \mid t_n, \theta) + \ln p(t_n \mid \theta) \right] \quad \dots$$

$$= \sum_{k=1}^{K} \sum_{n:\{t_n = k\}} \left[\ln p(x_n \mid C_k, \theta_k) + \ln p(C_k \mid \theta_k) \right]$$

When we differentiate w.r.t. the parameters for class k, only data points belonging to that class influence the derivative.

Example: Gaussian class-conditional likelihoods

$$p(C_k) = \pi_k \qquad \theta = \{ \pi_1 \dots \pi_K, \mu_1 \dots \mu_K, \Sigma_1 \dots \Sigma_K \}$$

$$p(x \mid C_k) = \mathcal{N}(x \mid \mu_k, \Sigma_k)$$

$$D \to p(x, C_k) \equiv$$

$$D \to \{ D_1, D_2, \dots, D_k, \dots, D_K \}, \quad D_k \to ML \to \begin{cases} p(C_k) \to \pi_{ML} = \dfrac{N_k}{N} \\ p(x \mid C_k) \to \begin{cases} \mu_k = \dfrac{1}{N_k} \displaystyle\sum_{n:t_n=k} x_n \\ \Sigma_k = \dfrac{1}{N_k} \displaystyle\sum_{n:t_n=k} (x_n - \mu_k)^T (x_n - \mu_k) \end{cases} \end{cases}$$

Special Case: equal Σ

<u>Example</u> <u>Gaussian distributions</u>

$$p(x|C_k) = \overbrace{\frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma_k|^{1/2}}}^{c} \exp\Big\{ \underbrace{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1} (x-\mu_k)}_{= c} \Big\}$$

quadratic

if 2-class, decision boundries:

$$p(C_1|x) = 0{,}5 \Leftrightarrow \ln p(C_1|x) = c \Leftrightarrow$$

if 2-classes and $\Sigma_1 = \Sigma_2 = \Sigma \Rightarrow$

$$p(C_1|x) = \sigma(a)$$

$$a = \ln \frac{\mathcal{N}(x|\mu_1,\Sigma)\, p(C_1)}{\mathcal{N}(x|\mu_2,\Sigma)\, p(C_2)} = \cancel{\ln c} - \frac{1}{2}(x-\mu_1)^T \Sigma^{-1} (x-\mu_1) + \ln p(C_1)$$
$$- \cancel{\ln c} + \frac{1}{2}(x-\mu_2)^T \Sigma^{-1} (x-\mu_2) - \ln p(C_2)$$

$$= \cancel{-\frac{1}{2} x^T \Sigma^{-1} x} \cancel{+\frac{1}{2} x^T \Sigma^{-1} x} - \frac{1}{2}\mu_1^T \Sigma^{-1} \mu_1 + \frac{1}{2}\mu_2^T \Sigma^{-1} \mu_2$$
$$+ \mu_1^T \Sigma^{-1} x - \mu_2^T \Sigma^{-1} x + \ln \frac{p(C_1)}{p(C_2)}$$

$$= \underbrace{(\mu_1 - \mu_2)^T \Sigma^{-1}}_{w^T} x \underbrace{- \frac{1}{2}\mu_1^T \Sigma^{-1} \mu_1 + \frac{1}{2}\mu_2^T \Sigma^{-1} \mu_2 + \ln \frac{p(C_1)}{p(C_2)}}_{w_0}$$

$$= w^T x + w_0$$ linear!