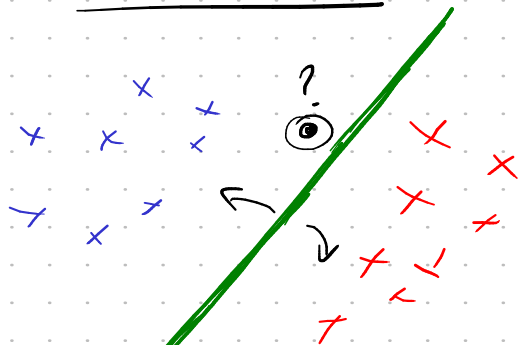
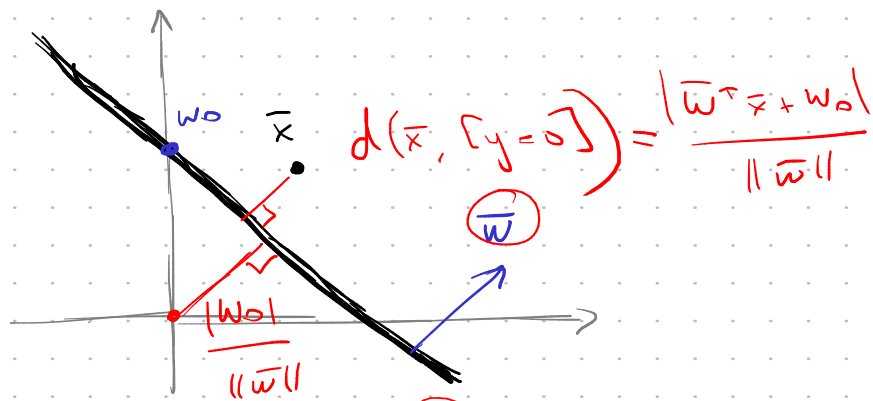


$$p(\bar{\theta} | D) = \frac{p(\bar{\theta}) p(D | \bar{\theta})}{p(D)}$$

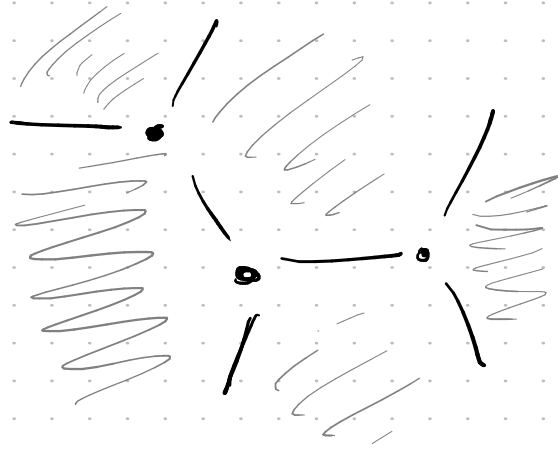
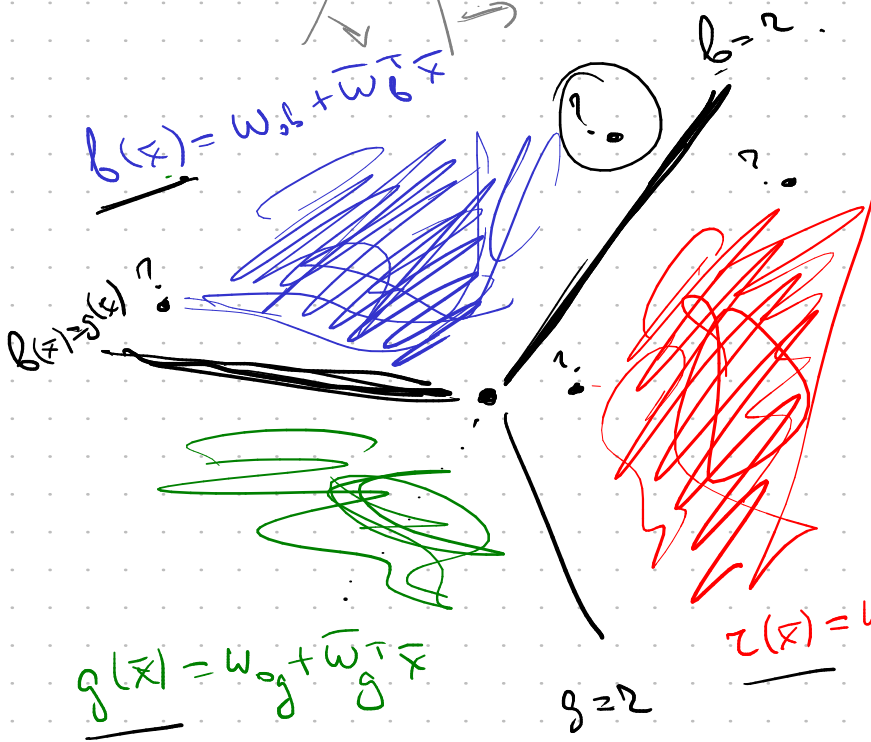
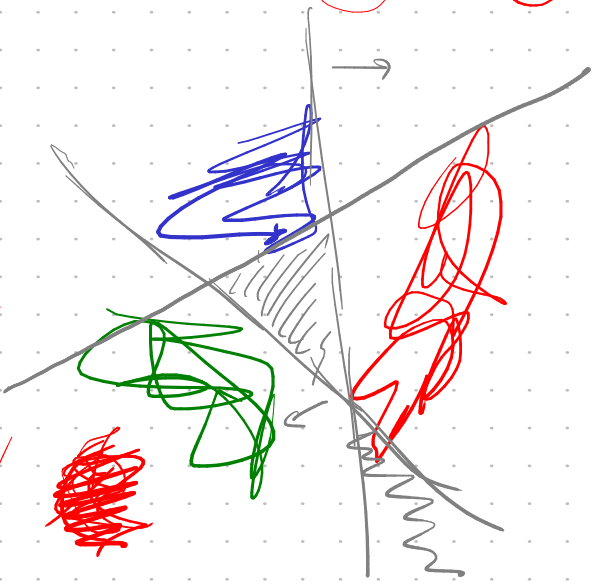
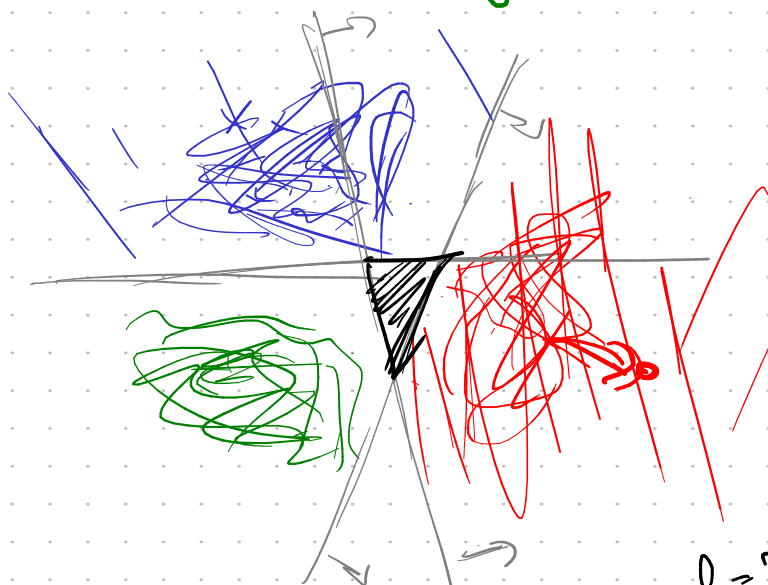
① Classification



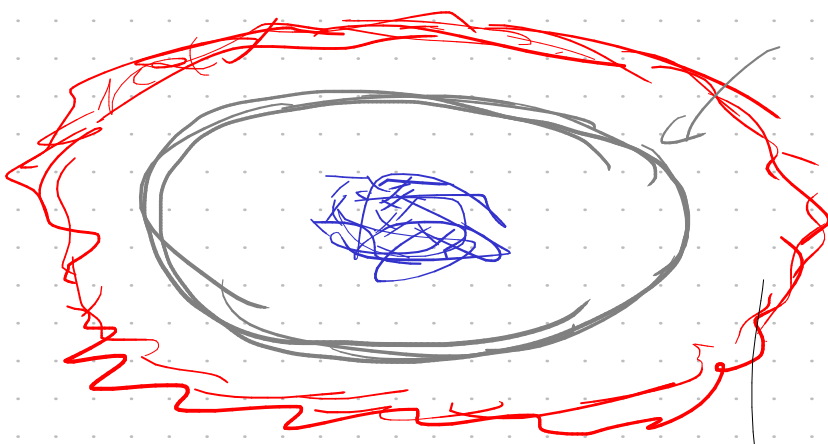
decision surface / boundary



$$\bar{w}^T \bar{x} + w_0 = 0$$



$$z(x) = w_0 + w_z^T x$$



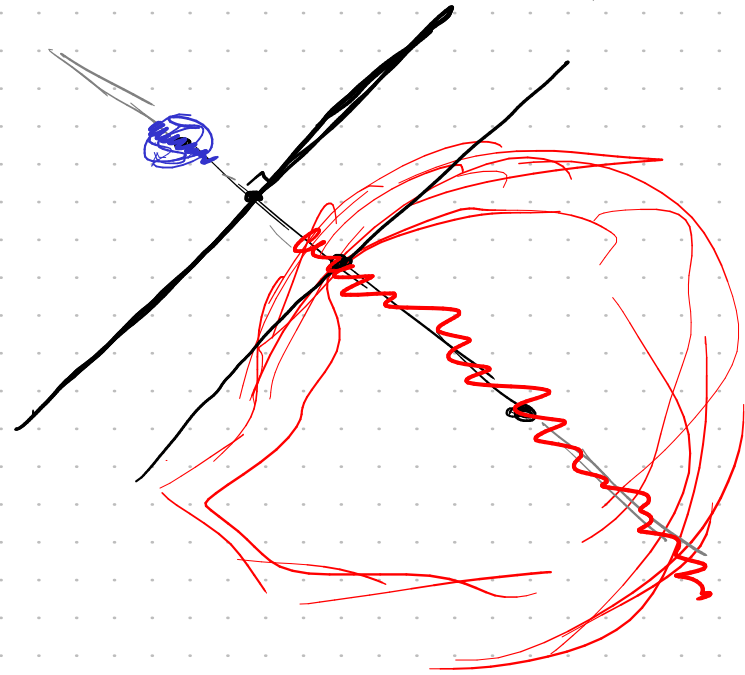
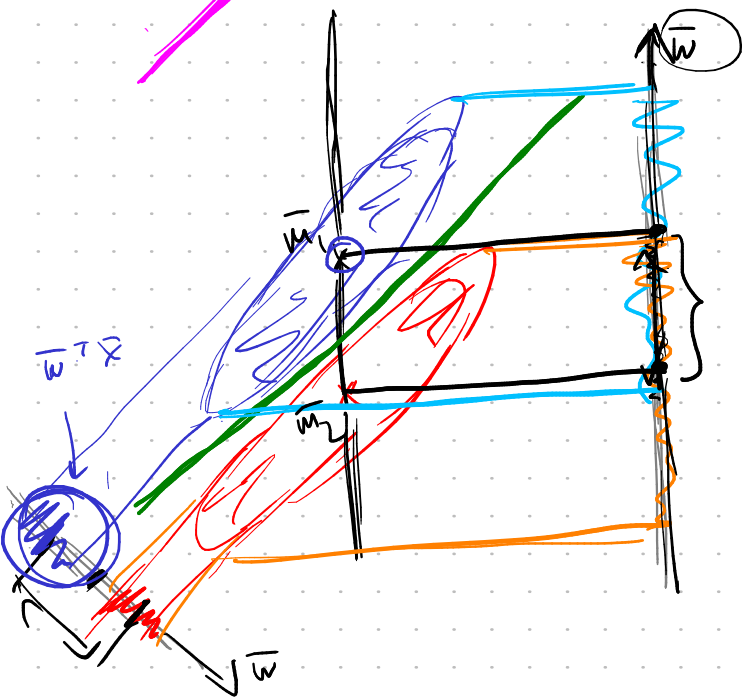
$$\frac{w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_1 x_2 + w_5 x_2^2}{1}$$

$$\begin{pmatrix} 1 \\ x_1 \\ x_2 \\ x_1^2 \\ x_1 x_2 \\ x_2^2 \end{pmatrix}$$

② Fisher's linear discriminant

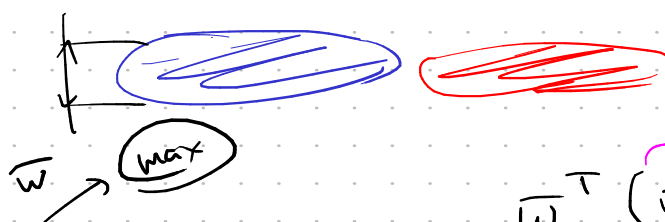
$$\bar{\mathbf{m}}_1 = \frac{1}{N_1} \sum_{\mathbf{x} \in C_1} \mathbf{x}$$

$$\bar{\mathbf{m}}_2 = \frac{1}{N_2} \sum_{\mathbf{x} \in C_2} \mathbf{x}$$



$$\left(\bar{\mathbf{w}}^T (\bar{\mathbf{m}}_1 - \bar{\mathbf{m}}_2) \right)^2 \xrightarrow{\bar{\mathbf{w}}} \max$$

$$\sum_{\mathbf{x} \in C_1} \left(\bar{\mathbf{w}}^T \mathbf{x} - \bar{\mathbf{w}}^T \bar{\mathbf{m}}_1 \right)^2 + \sum_{\mathbf{x} \in C_2} \left(\bar{\mathbf{w}}^T \mathbf{x} - \bar{\mathbf{w}}^T \bar{\mathbf{m}}_2 \right)^2 \xrightarrow{\bar{\mathbf{w}}} \min$$



S_B between-class covariance

$$J(\bar{\mathbf{w}}) = \frac{\bar{\mathbf{w}}^T (\bar{\mathbf{m}}_1 - \bar{\mathbf{m}}_2) (\bar{\mathbf{m}}_1 - \bar{\mathbf{m}}_2)^T \bar{\mathbf{w}}}{\bar{\mathbf{w}}^T \left(\sum_{\mathbf{x} \in C_1} (\mathbf{x} - \bar{\mathbf{m}}_1) (\mathbf{x} - \bar{\mathbf{m}}_1)^T + \sum_{\mathbf{x} \in C_2} (\mathbf{x} - \bar{\mathbf{m}}_2) (\mathbf{x} - \bar{\mathbf{m}}_2)^T \right) \bar{\mathbf{w}}}$$

S_w - within-class cov.

$$J(\bar{w}) = \frac{\bar{w}^T S_B \bar{w}}{\bar{w}^T S_w \bar{w}} \xrightarrow{\bar{w}} \max$$

$$S_B + S_w = \sum_{c_1, c_2} (\bar{x} - \bar{m})(\bar{x} - \bar{m})^T$$

$$\nabla_{\bar{w}} J = \frac{2 S_B \bar{w} (\bar{w}^T S_w \bar{w}) - 2 S_w \bar{w} (\bar{w}^T S_B \bar{w})}{\bar{w}^T S_w \bar{w}} = 0$$

$$(\bar{w}^T S_w \bar{w}) \cdot S_B \bar{w} = (\bar{w}^T S_B \bar{w}) S_w \bar{w}$$

$$S_B \bar{w} \propto S_w \bar{w}$$

$$(\bar{m}_1 - \bar{m}_2) \left((\bar{m}_1 - \bar{m}_2)^T \bar{w} \right)$$

$$S_w \bar{w} \propto \bar{m}_1 - \bar{m}_2$$

$$\bar{w} \propto S_w^{-1} (\bar{m}_1 - \bar{m}_2)$$

③ Generative models



generative models:
discriminative models:

$$p(c_1|\bar{x}) = \frac{p(\bar{x}|c_1) \cdot p(c_1)}{p(\bar{x}|c_1)p(c_1) + p(\bar{x}|c_2)p(c_2)}$$

Optimal Bayes classifier



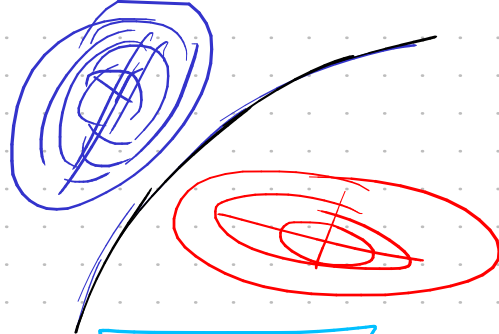
$$p(c_1|\bar{x}) = p(c_2|\bar{x}); \quad p(\bar{x}|c_1)p(c_1) = p(\bar{x}|c_2)p(c_2)$$

$$\log \frac{p(\bar{x}|c_1)}{p(\bar{x}|c_2)} + \log \frac{p(c_1)}{p(c_2)} = 0$$

④ LDA, QDA

$$p(\bar{x}|C_1) = \mathcal{N}(\bar{x} | \bar{\mu}_1, \Sigma_1)$$

$$p(\bar{x}|C_2) = \mathcal{N}(\bar{x} | \bar{\mu}_2, \Sigma_2)$$



$$\log p(C_1) + \log p(\bar{x}|C_1) = \log p(C_2) + \log p(\bar{x}|C_2)$$

$$\begin{aligned} \log \frac{p(C_1)}{p(C_2)} - \frac{d}{2} \log 2\pi - \frac{1}{2} \log \det \Sigma_1 - \frac{1}{2} (\bar{x} - \bar{\mu}_1)^T \Sigma_1^{-1} (\bar{x} - \bar{\mu}_1) &= \\ &= -\frac{d}{2} \log 2\pi - \frac{1}{2} \log \det \Sigma_2 - \frac{1}{2} (\bar{x} - \bar{\mu}_2)^T \Sigma_2^{-1} (\bar{x} - \bar{\mu}_2) \end{aligned}$$

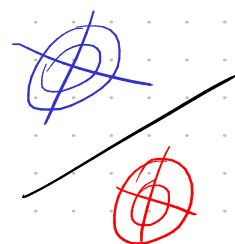
$$\begin{aligned} \log \frac{p(C_1)}{p(C_2)} + \frac{1}{2} \log \frac{\det \Sigma_2}{\det \Sigma_1} - \frac{1}{2} \left(\bar{x}^T \Sigma_1^{-1} \bar{x} - 2 \bar{x}^T (\Sigma_1^{-1} \bar{\mu}_1) + \bar{\mu}_1^T \Sigma_1^{-1} \bar{\mu}_1 \right) &= \\ = -\frac{1}{2} \left(\bar{x}^T \Sigma_2^{-1} \bar{x} - 2 \bar{x}^T (\Sigma_2^{-1} \bar{\mu}_2) + \bar{\mu}_2^T \Sigma_2^{-1} \bar{\mu}_2 \right) \end{aligned}$$

$$-\frac{1}{2} \bar{x}^T (\Sigma_1^{-1} - \Sigma_2^{-1}) \bar{x} + \bar{x}^T (\Sigma_1^{-1} \bar{\mu}_1 - \Sigma_2^{-1} \bar{\mu}_2) + \dots = 0$$

Optimal Bayes classifier

$$\begin{aligned} \hat{\bar{\mu}}_c &= \frac{1}{N} \sum_c \bar{x} \\ \hat{\Sigma}_c &= \frac{1}{N} \sum_c (\bar{x} - \hat{\bar{\mu}})(\bar{x} - \hat{\bar{\mu}})^T \end{aligned}$$

QDA
quadratic
discriminant
analysis



LDA - Linear Disc. analysis

$$\Sigma_1 = \Sigma_2 = \Sigma$$

$$\hat{\Sigma} = \frac{N_1}{N_1 + N_2} \hat{\Sigma}_1 + \dots \hat{\Sigma}_n$$

$$(\bar{\mu}_1, \bar{\mu}_2, \Sigma)$$

$$p(\mathbf{0} | \bar{\mu}_1, \bar{\mu}_2, \Sigma) = \left[\prod_{\bar{x} \in C_1} \mathcal{N}(\bar{x} | \bar{\mu}_1, \Sigma) \right] \cdot \left[\prod_{\bar{x} \in C_2} \mathcal{N}(\bar{x} | \bar{\mu}_2, \Sigma) \right] \rightarrow \max_{\bar{\mu}_1, \bar{\mu}_2, \Sigma}$$

$$\bar{w}^T \bar{x} \rightarrow \boxed{???} \rightarrow p(C_1 | \bar{x}) \downarrow \mu_1, \Sigma_1$$

⑤ Logistic regression / discriminative model / linear model $\bar{w}^T \bar{x}$

$$y = p(C_1 | \bar{x}) = \frac{p(\bar{x} | C_1) p(C_1)}{p(\bar{x} | C_1) p(C_1) + p(\bar{x} | C_2) p(C_2)}$$

$$= \frac{1}{1 + \frac{p(\bar{x} | C_2) p(C_2)}{p(\bar{x} | C_1) p(C_1)}}$$

odds $\frac{p(C_2 | \bar{x})}{p(C_1 | \bar{x})} = \frac{p(\bar{x} | C_2) p(C_2)}{p(\bar{x} | C_1) p(C_1)} \in (0, \infty) \approx \bar{w}^T \bar{x}?$

log-odds $\frac{1}{1 + e^{-\log \frac{p(\bar{x} | C_1) p(C_1)}{p(\bar{x} | C_2) p(C_2)}}} = \frac{1}{1 + e^{-\bar{w}^T \bar{x}}} \in (-\infty, \infty)$

$$p(C_1 | \bar{x}) = \frac{1}{1 + e^{-\bar{w}^T \bar{x}}}$$

$$\sigma(a) = \frac{1}{1 + e^{-a}}$$

logistic sigmoid

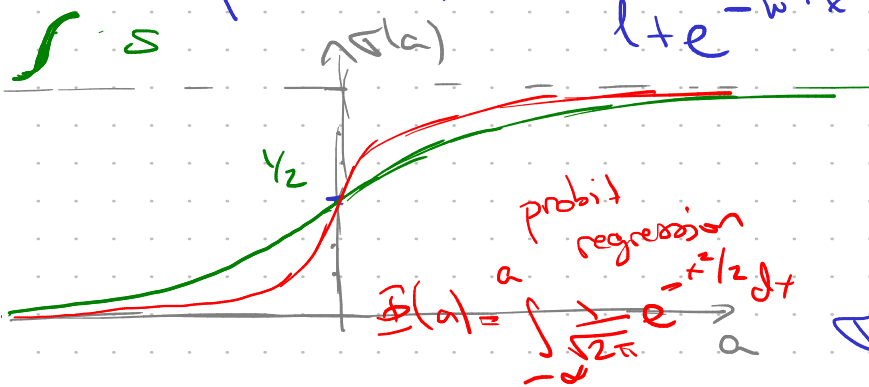
$$\sigma(-a) = 1 - \sigma(a)$$

$$\sigma'(a) = \frac{1}{(1 + e^{-a})^2} \cdot (+e^{-a}) = \sigma(a) \cdot (1 - \sigma(a))$$

$$\sigma'(a) = \frac{1}{1 + e^{-a}} \cdot \frac{e^{-a}}{1 + e^{-a}}$$

$$\sigma'(a) = \sigma(a) \cdot (1 - \sigma(a))$$

$$p(\mathbf{D} | \bar{w}) = \prod_{n=1}^N p(y_n | \bar{w}, \bar{x}_n) = \prod_{n=1}^N \begin{cases} y_n = 1 \Rightarrow \sigma(\bar{w}^T \bar{x}_n) \\ y_n = 0 \Rightarrow 1 - \sigma(\bar{w}^T \bar{x}_n) \end{cases}$$



$$= \prod_n \left(y_n \cdot \sigma(\bar{w}^T \bar{x}_n) + (1-y_n) \cdot (1-\sigma(\bar{w}^T \bar{x}_n)) \right) =$$

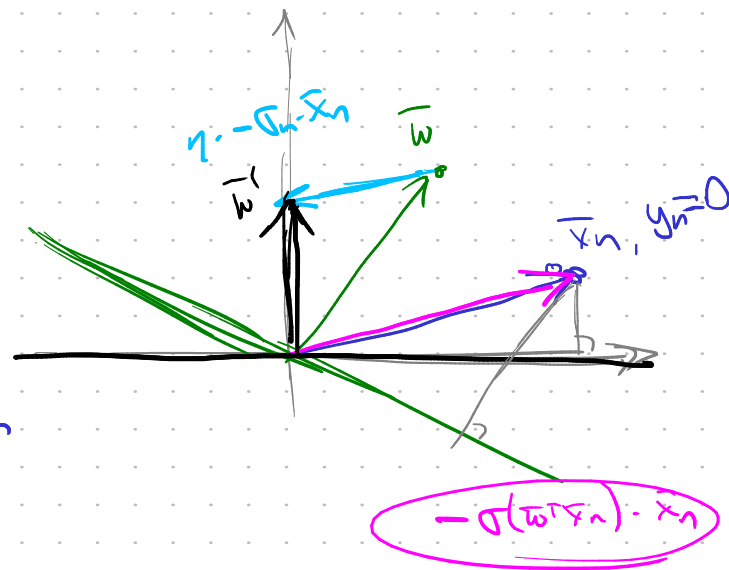
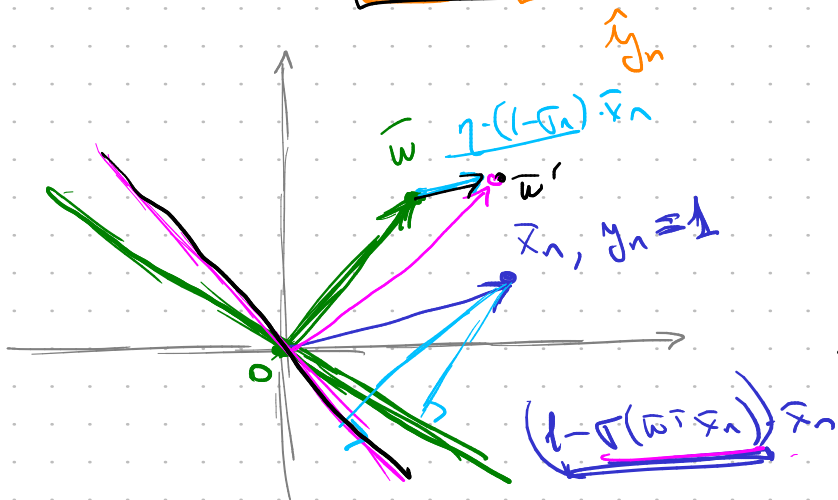
$$= \prod_{n=1}^N \underbrace{\sigma(\bar{w}^T \bar{x}_n)}^{y_n} \underbrace{(1-\sigma(\bar{w}^T \bar{x}_n))^{1-y_n}}_{\xrightarrow{\bar{w}} \max}$$

$$\log p(D|\bar{w}) = \sum_n \left(y_n \log \sigma_n + (1-y_n) \log (1-\sigma_n) \right)$$

$$\nabla_{\bar{w}} \log p(D|\bar{w}) = \sum_n \left(y_n \frac{\sigma_n(1-\sigma_n)}{\sigma_n} \cdot \bar{x}_n - (1-y_n) \frac{\sigma_n(1-\sigma_n)}{1-\sigma_n} \bar{x}_n \right)$$

$$= \sum_n \left(\overset{y_n - y_n \sigma_n - \sigma_n + y_n \sigma_n}{y_n(1-\sigma_n) - (1-y_n)\sigma_n} \right) \bar{x}_n =$$

$$= \sum_n \left(y_n - \sigma(\bar{w}^T \bar{x}_n) \right) \bar{x}_n$$



$$L(\bar{w}) \approx L(\bar{w}_0) + \nabla_{\bar{w}} L(\bar{w}_0)^T (\bar{w} - \bar{w}_0) + \frac{1}{2} (\bar{w} - \bar{w}_0)^T H|_{\bar{w}_0} (\bar{w} - \bar{w}_0)$$

$$\nabla_{\bar{w}} L(\bar{w}_0) + H(\bar{w}_* - \bar{w}_0) = 0$$

$$\bar{w}_* = \bar{w}_0 - H^{-1} \cdot \nabla_{\bar{w}} L(\bar{w}_0)$$

$$\frac{\partial^2 \log p(D|\bar{w})}{\partial w_i \partial w_j} = \frac{\partial}{\partial w_i} \left(\sum_n (y_n - \sigma(\bar{w}^T \bar{x}_n)) \cdot x_{nj} \right) =$$

$$= \sum_n \left(-\sigma_n (1 - \sigma_n) \cdot x_{ni} \right) x_{nj} =$$

$$= - \sum_n \sigma_n (1 - \sigma_n) x_{ni} x_{nj} = (H)_{ij}$$

$$H = - X^T \begin{pmatrix} \sigma_1(1-\sigma_1) & & 0 \\ & \ddots & \\ 0 & & \sigma_N(1-\sigma_N) \end{pmatrix} X$$

$d \times d$ $d \times N$ $N \times N$ $N \times d$

$\sigma_n = \sigma(\bar{w}^T \bar{x}_n)$

$$\bar{w}^{(k+1)} = \bar{w}^{(k)} + (X^T R X)^{-1} X^T (\bar{y} - \bar{\sigma})$$

$$\bar{w}_* = (X^T X)^{-1} X^T \bar{y}$$

IRLS - iterative reweighted least squares

$$p(\bar{w}|D) \propto p(\bar{w}) p(D|\bar{w})$$

$$\log p(\bar{w}|D) = \text{const} + \log p(\bar{w}) + \log p(D|\bar{w})$$

$$p(\bar{w}) = \mathcal{N}(\bar{w}|\bar{0}, \sigma_0^2 \mathbf{I}) \quad \sum_n (y \log \sigma - \dots)$$

$$\log p(\bar{w}|D) = \dots - \frac{1}{2\sigma_0^2} \bar{w}^T \bar{w}$$

$$\nabla_{\bar{w}} \log p = \dots - \frac{1}{\sigma_0^2} \bar{w}$$

$$H = \dots - \frac{1}{\sigma_0^2} \mathbf{I}$$

$$\approx \bar{w}_k^T \bar{x}$$

⑥ Multiclass logistic regression

$$p(C_k|\bar{x}) = \frac{p(\bar{x}|C_k) p(C_k)}{\sum_{s=1}^K p(\bar{x}|C_s) p(C_s)} = \frac{e^{\log p(\bar{x}|C_k) p(C_k)}}{\sum_s e^{\log p(\bar{x}|C_s) p(C_s)}}$$

$$\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_K)$$

$$\hat{y}_k = \frac{e^{\bar{\mathbf{w}}_k^T \bar{\mathbf{x}}}}{\sum_s e^{\bar{\mathbf{w}}_s^T \bar{\mathbf{x}}}}$$

affinity

$$\bar{\mathbf{a}} = (a_1, a_2, \dots, a_K)$$

$$\bar{\mathbf{a}} \mapsto \left(\dots, \frac{e^{a_k}}{\sum_s e^{a_s}}, \dots \right)$$

softmax

$$p(D|\bar{\mathbf{w}}) = \prod_{n=1}^N \left(\prod_{k=1}^K p(C_k | \bar{\mathbf{x}}_n, \bar{\mathbf{w}})^{y_{nk}} \right)$$

$$\log p(D|\bar{\mathbf{w}}) = \sum_n \sum_k y_{nk} \log \left(\frac{e^{\bar{\mathbf{w}}_k^T \bar{\mathbf{x}}_n}}{\sum_s e^{\bar{\mathbf{w}}_s^T \bar{\mathbf{x}}_n}} \right)$$

one-hot encoding

$$\bar{\mathbf{y}}_n = (0, 0, 1, 0)$$

$\bar{\mathbf{x}}_n \in C_k$

$$p(D|\bar{\mathbf{w}}) = \prod_n \left(p(C_1 | \bar{\mathbf{x}}_n, \bar{\mathbf{w}})^{y_{n1}} p(C_2 | \bar{\mathbf{x}}_n, \bar{\mathbf{w}})^{y_{n2}} \dots p(C_K | \bar{\mathbf{x}}_n, \bar{\mathbf{w}})^{y_{nK}} \right)$$

$$p(C_1 | \bar{\mathbf{x}}_n, \bar{\mathbf{w}})^{y_{n1}} p(C_2 | \bar{\mathbf{x}}_n, \bar{\mathbf{w}})^{y_{n2}} \dots p(C_K | \bar{\mathbf{x}}_n, \bar{\mathbf{w}})^{y_{nK}}$$

~~$$y_{n1} \cdot p_1 + y_{n2} \cdot p_2 + \dots + y_{nK} \cdot p_K$$~~

$$\nabla_{\bar{\mathbf{w}}_j} \log p(D|\bar{\mathbf{w}}) = \sum_n (y_{nj} - \hat{y}_{nj}) \cdot \bar{\mathbf{x}}_n$$

$$p(C_1 | \bar{\mathbf{x}}, D) = \int p(C_1 | \bar{\mathbf{x}}, \bar{\mathbf{w}}) \cdot \underline{p(\bar{\mathbf{w}} | D)} d\bar{\mathbf{w}} =$$

$$= \int \sigma(\bar{\mathbf{w}}^T \bar{\mathbf{x}}) \cdot \underline{p(\bar{\mathbf{w}} | D)} d\bar{\mathbf{w}}$$

$$\propto p(\bar{\mathbf{w}}) p(D|\bar{\mathbf{w}})$$

$$p(C_1 | \bar{\mathbf{x}}, D) = \sigma \left(\alpha \cdot \bar{\mathbf{w}}_{\text{MAP}}^T \bar{\mathbf{x}} \right) = \sigma \left(\frac{1}{\sqrt{1 + \dots}} \cdot \bar{\mathbf{w}}_{\text{MAP}}^T \bar{\mathbf{x}} \right)$$

