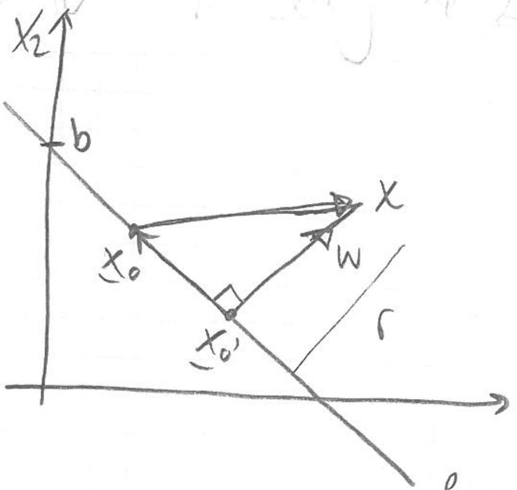
SVM



$$\mathcal{L} = \{\underline{x} : \underline{x}^T \underline{w} + b = 0\}$$

i) For $\underline{x}_0 \in \mathcal{L}$ har vi at $\underline{x}_0^T \underline{w} = -b$

ii) For $\underline{x}_0' \in \mathcal{L}$ og $\underline{x}_0 \in \mathcal{L}$ gælder:

$$(\underline{x}_0 - \underline{x}_0')^T w = -b - (-b) = 0$$

Dette medfører at $\underline{w}$ er normalvektor til $\mathcal{L}$

ii') Specielt er $w^* = \frac{w}{\|w\|}$ også normal (og har enhedslængde)

iii) Afstanden (med fortegn) fra punkt $\underline{x}$ til $\mathcal{L}$ er givet ved skalar projektionen:

$$(\underline{x} - \underline{x}_0)^T w^* = (\underline{x} - \underline{x}_0)^T \frac{w}{\|w\|} = \frac{1}{\|w\|}(x^T w - x_0^T w)$$

idet $x_0 \in \mathcal{L}$ $\longrightarrow$ $= \frac{1}{\|w\|}(x^T w + b)$

Definér $g(\underline{x}) = \underline{x}^T \underline{w} + b$. Da er $g'(x) = \frac{\partial g}{\partial x}(x) = \underline{w}$.

Dvs $(\underline{x} - \underline{x}_0)^T w^* = \frac{g(x)}{\|g'(x)\|} = r$

Yderemere $\rho = 2|r|$. Som vi senere kalder for marginen.

Anteg vi har data således (Y, X) udgør data for $i = 1, \dots, n$.

Vi kan vælge (for to klasser) at omkode Y til

$y_i = 1$ for klasse 1

$y_i = -1$ for klasse 2.

<u>Ønske</u>

Maximere afstanden mellem hyperplanen og punkterne. således alle klasse 1 observationer er på én side af planen, mens alle klasse 2 obs. er på modsat side:

$$\max_{\underline{w}, b, \|w\|=1} \rho \quad \text{således} \quad \begin{array}{ll} x_i^T w + b \geq r & y_i = 1 \\ x_i^T w + b \leq r & y_i = -1 \end{array}$$

Husk r med fortegn.

Bruger nu at $y_i = 1$ eller $y_i = -1$:

$$\left.\begin{array}{ll} y_i = 1 & x_i^T w + b \geq r \\ y_i = -1 & x_i^T w + b \leq r \end{array}\right\} \; y_i (x_i^T w + b) \geq r$$

Dvs

$$\max_{\underline{w}, b, \|w\|=1} \rho \quad \text{således} \quad y_i(x_i^T w + b) \geq r$$

~~Vælges $\|w\| = 1$~~ Som er ækvivalent med

$$\max_{\underline{w}, \underline{b}} \rho \quad \text{således} \quad \frac{1}{\|w\|} y_i (x_i^T w + b) \geq r$$

Vælges nu $\|w\| = \frac{1}{r}$ har vi

$$\max_{\underline{w},\, \underline{b}} \; \rho = 2|r| = \frac{2}{\|w\|} \quad \text{således} \quad y_i(x_i^T w + b) \geq 1$$

Dette problem har samme fix punkt (løsning) som

$$\min_{\underline{w},\, b} \; \cancel{\max} \; \frac{1}{2}\|w\|^2 \quad \text{således} \quad y_i(x_i^T w + b) \geq 1,$$

$$\text{dvs} \quad 1 - y_i(x_i^T w + b) \leq 0$$

Som er et Quadratic Programmerings Problem. (QP)

For store dataset / mængder er QP dog ikke beregnings-mæssigt hensigtsmæssig.

Løsning er Lagrange Multipliers eller rettere "Karush-Kuhn-Tucker" ~~med~~ betingelser. // Wolfe dual

Det primære Lagrange funktion som skal minimeres mht. $\underline{w}$ og $\underline{b}$ er således:

burde være "+", men vi har vendt uligheden: $y_i(x_i^T w + b) - 1 \leq 0$

$$L_p = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i [y_i(x_i^T w + b) - 1]$$

Differentiere mht $\underline{w}$ og $\underline{b}$ giver det hhv (husk $\|w\|^2 = w^T w$)

$$\left.\begin{aligned} \frac{\partial}{\partial \underline{w}} L_p &= w - \sum_{i=1}^{n} \alpha_i y_i x_i = \underline{0} \;\Rightarrow\; \hat{\underline{w}} = \sum_{i=1}^{n} \alpha_i y_i x_i \\ \frac{\partial}{\partial b} L_p &= -\sum_{i=1}^{n} \alpha_i y_i = 0 \;\Rightarrow\; \sum_{i=1}^{n} \alpha_i y_i = 0 \end{aligned}\right\} (*)$$

$$\alpha_i \geq 0 \quad \text{for alle } i = 1, \ldots, n.$$

Bibetingelses ligninger

Indsættes (*) i L_P får vi det såkaldte Wolfe dual

$$L_D = \frac{1}{2}\hat{w}^T\hat{w} - \sum_{i=1}^{n} \alpha_i \left[y_i (x_i^T \hat{w} + \text{"}\hat{b}\text{"}) - 1 \right]$$

$\alpha[y_i x_i^T w + y_i b - 1] = \alpha y_i x_i^T w + \alpha y_i b - \alpha$

$$= \frac{1}{2}\left(\sum_{i=1}^{n} \alpha_i y_i x_i\right)^T \left(\sum_{j=1}^{n} \alpha_j y_j x_j\right) - \sum_{i=1}^{n} \alpha_i \left[y_i \left(x_i^T \left\{ \sum_{j=1}^{n} \alpha_j y_j x_j \right\} + \underbrace{\left\{ \sum_{i=1}^{n} \alpha_i y_i \right\}}_{0} \right) \right]$$

$$= \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^T x_j - \sum_{i,j} \alpha_i y_i \alpha_j y_j x_i^T x_j$$

$$- \underbrace{\left\{ \sum_{i=1}^{n} \alpha_i y_i \right\}}_{=0} \underbrace{\left\{ \sum_{j=1}^{n} \alpha_j y_j \right\}}_{=0} + \sum_{i=1}^{n} \alpha_i$$

$$= \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^T x_j \quad \text{således} \quad \begin{cases} \sum_{i=1}^{n} \alpha_i y_i = 0 \\ \alpha_i \geq 0 \\ !! \ \alpha_i [y_i (x_i^T w + b) - 1] = 0 \quad \forall i = 1, \dots \end{cases}$$

<u>Heraf ser vi at (!!):</u>

- Hvis $\alpha_i > 0$ så må $\underline{y_i(x_i^T w + b)} = 1$. Dvs at x_i ligger på randen/kanten af vores område

Husk: $y_i(x_i^T \hat{w} + b) \geq 1$

- Hvis $\underline{y_i(x_i^T w + b) > 1}$ så $\alpha_i = 0$.

<u>DERFOR:</u> Punkter med $\alpha_i > 0$ er således eneste bidrag til $$\hat{w} = \sum_{i=1}^{n} \alpha_i y_i x_i$$

og kaldes derfor <u>SUPPORT VECTORS</u>

Bemærk

$\hat{g}(x) = x^T\hat{w} + \hat{b}$ giver anledning til en ~~klassifier~~ klassifier: $G(\underline{x}) = \text{sign}\{\hat{g}(x)\}$

Sidebemærkning:

Logistisk regression giver tilsvarende hyperplan i det separable tilfælde:

Fit: glm($\underbrace{\text{class}}_{y}$ ~ ., family = "binomial")

Planen er givet for $P(\text{klasse } i \mid x)) = 0.5$

$$0 = \text{logit } 0.5 = \log\frac{0.5}{1-0.5} = \text{logit } P(\text{klasse} \mid \underline{x}) = \beta^T\underline{x} + \beta_0.$$

Så for $\underline{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ har vi med $\beta = (\beta_1, \beta_2)$:

$$0 = \beta^T x + \beta_0 = \beta_1 x_1 + \beta_2 x_2 + \beta_0 \Rightarrow x_2 = -\frac{\beta_0 + \beta_1 x}{\beta_2}.$$

Det ikke-separable tilfælde

En mulig løsning er at indføre "slack" variable, som tillader at ikke alle observationer er på den korrekte side af hyperplanen.

"A few rotten apples in the basket".

Sidebetingelser bliver nu:

$$y_i(x_i^T \underline{w} + b) \geq 1 - \xi_i$$

hvor $\xi_i \geq 0$ samt $\sum_{i=1}^{n} \xi_i \leq C$, hvor C er en "tunings-parameter". I R-implementering kaldes C "cost".

ξ_i kan fortolkes som afstanden punkt i ligger på den forkerte side af $x^T w + b = y_i$.

Igen får vi et kvadratisk programmerings problem (QP)

$$\min_{\underline{w}, b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i$$

Note $C = \infty$ ~ separable

således

$$\xi_i \geq 0 \quad \text{og} \quad y_i(x_i^T w + b) \geq 1 - \xi_i.$$

Det primære Lagrange funktional

$$L_P = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i [y_i(x_i^T w + b) - (1 - \xi_i)] - \sum_{i=1}^{n} \mu_i \xi_i$$

minimer mht $\underline{w}$, b og ξ_i

Differentiation mht $\underline{w}$, b og ξ_i giver:

$$\frac{\partial}{\partial w} L_P = w - \sum_{i=1}^{n} \alpha_i y_i x_i = 0 \Rightarrow \hat{w} = \sum_{i=1}^{n} \alpha_i y_i x_i$$

$$\frac{\partial}{\partial b} L_P = -\sum_{i=1}^{n} \alpha_i y_i = 0 \Rightarrow \sum_{i=1}^{n} \alpha_i y_i = 0$$

$$\frac{\partial}{\partial \xi_i} L_P = C - \alpha_i - \mu_i = 0 \Rightarrow \underbrace{\alpha_i = C - \mu_i}_{(**)} \quad i = 1, \dots, n.$$

(**)

Indsættes (**) samt $\alpha_i \geq 0$ $\mu_i \geq 0$ og $\xi_i \geq 0$ $i = 1, \dots, n$

i L_P får vi igen Wolfe dual:

$$L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^T x_j$$ således at

$0 \leq \alpha_i \leq C$ og $\sum_{i=1}^{n} \alpha_i y_i = 0$, hvor maximering af

L_D også skal opfylde Karush-Kuhn-Tucker betingelser:

$$\left.\begin{aligned} \alpha_i [y_i(x_i^T w + b) - (1 - \xi_i)] &= 0 \\ \mu_i \xi_i &= 0 \\ y_i(x_i^T w + b) &\geq (1 - \xi_i) \end{aligned}\right\} \quad i = 1, \dots, n \quad (*)$$

Igen har vi ~~til~~ ~~tilfældet~~

- $\alpha_i > 0$ må medføre $y_i(x_i^T w + b) = 1 - \xi_i$

Dvs på "randen" jf ξ_i forskydning

~~$\alpha_i \xi_i = 0$~~ { Nogle vil have $\xi_i = 0$ og være på randen. ~~Dvs $\xi_i = 0$~~ Dvs $0 < \alpha_i < C$ for disse

{ For $\alpha_i > 0$ og $\xi_i > 0$ må $\mu_i = 0$ så (*) og (**) giver at $\alpha_i = C$.

Generelt ser vi at

$$L_D = \sum \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

$$= \sum \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle$$

← indre produkt.

Kernel trick

Består i at afbillede x ind i højere dimensionelt rum H vha ~~...~~ $\phi : X \to H$

Dvs

$$L_D = \sum \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \langle \phi(x_i), \phi(x_j) \rangle$$

Dvs $g(x)$ som før var $g(x) = x^T w + b$

nu bliver til

$$\tilde{g}(x) = \phi(x)^T w + b$$

Hvor vi tidligere fik $\hat{w} = \sum_{i=1}^{n} \alpha_i y_i x_i$ så har vi nu

$$\tilde{w} = \sum_{i=1}^{n} \alpha_i y_i \phi(x_i)$$

$$\tilde{g}(x) = \phi(x)^T \sum \alpha_i y_i \phi(x_i) + b$$

$$= \sum_{i=1}^{n} \alpha_i y_i \langle \phi(x), \phi(x_i) \rangle + b$$

Lad $K(x, x') = \langle \phi(x), \phi(x') \rangle$ da har vi

altså

$$L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j).$$

~~altså~~

~~For~~ d-grads polynomium:

$$K(x, x') = (1 + \langle x, x' \rangle)^d.$$

Fx $d = 2$: og $x = (x_1, x_2)$ og $z = (z_1, z_2)$

$$(1 + \langle x, z \rangle)^d = (1 + x_1 z_1 + x_2 z_2)^2$$

$$= (1 + (x_1 z_1)^2 + (x_2 z_2)^2 + 2 x_1 z_1 + 2 x_2 z_2 + 2 x_1 x_2 z_1 z_2$$

Dette svarer til at vi afbilder $\phi : X \to H$, hvor

$\phi : \mathbb{R}^2 \to \mathbb{R}^6$, hvor $\phi(x) = (\phi_1(x), \ldots, \phi_6(x))$ hvor

$$\left.\begin{aligned}
\phi_1(x) &= 1 \\
\phi_2(x) &= \sqrt{2}\, x_1 \\
\phi_3(x) &= \sqrt{2}\, x_2 \\
\phi_4(x) &= x_1^2 \\
\phi_5(x) &= x_2^2 \\
\phi_6(x) &= \sqrt{2}\, x_1 x_2
\end{aligned}\right\}$$

$$K(\phi(x), \phi(z))$$

$$= 1 + 2 x_1 z_1 + 2 x_2 z_2 + x_1^2 z_1^2 + x_2^2 z_2^2 + 2 x_1 z_1 x_2 z_2.$$