# Lecture 6
# Kernel Machines – Part 1

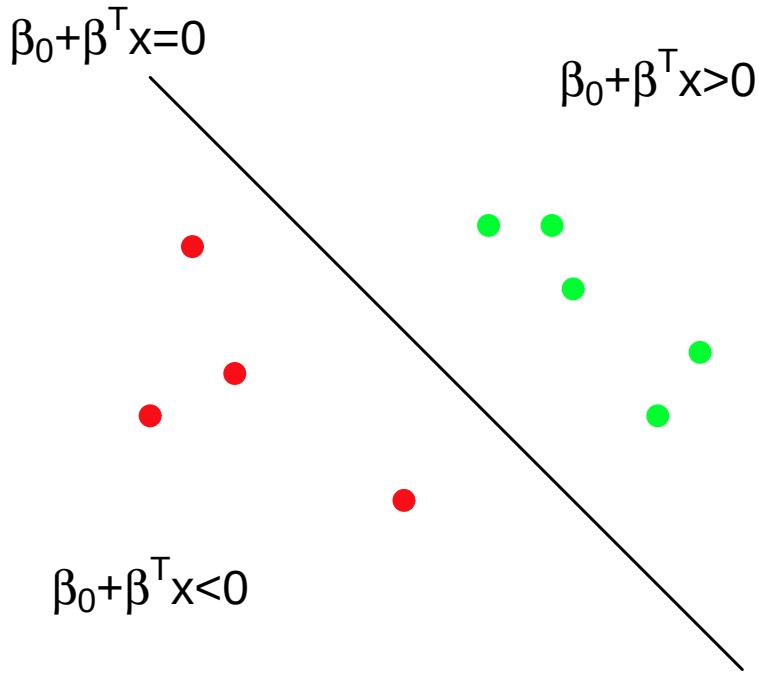
(February 13, 2015)

Mu Zhu
University of Waterloo

# Outline

- Part 1
  - support vector machine (SVM)
- Part 2 (after short break)
  - kernel machines

# Separating Hyperplane




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

a separating hyperplane if $\exists\ c > 0$ s.t. $y_i(\beta_0 + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i) \geq c\ \forall\ i$

# Scaling

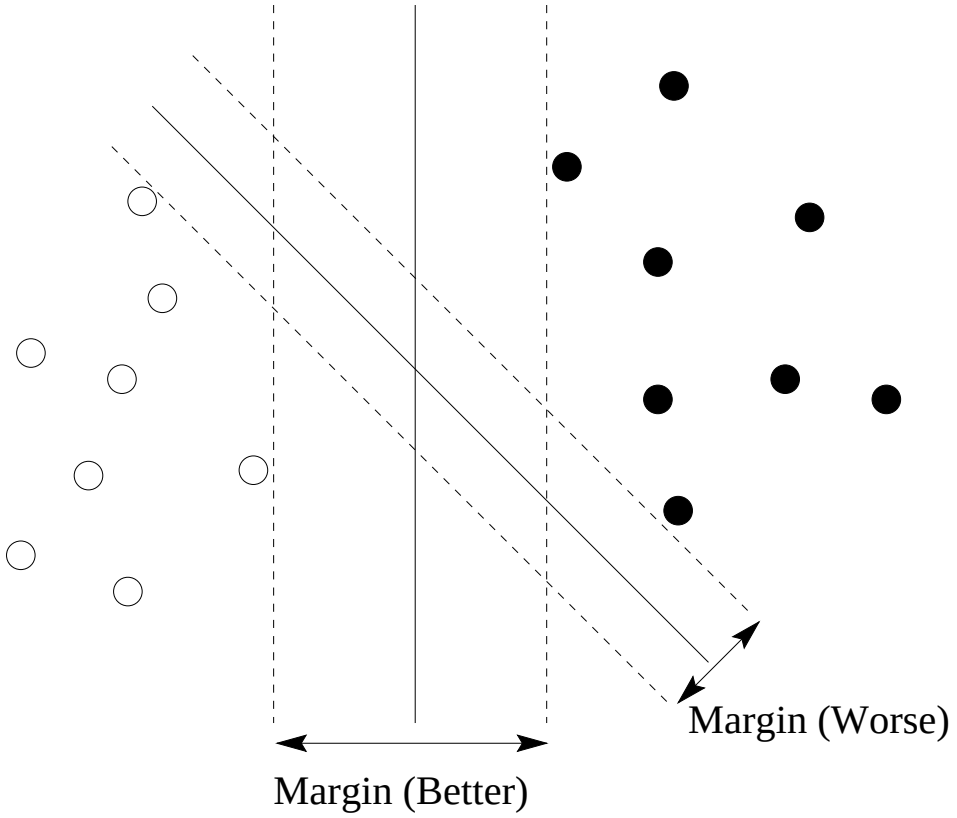
clearly, for any $s > 0$, $\quad \beta_0 + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x} = 0 \quad \Leftrightarrow \quad s(\beta_0 + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}) = 0$

always choose $s$ so that $c = 1$

call it a "canonical" separating hyperplane if

$$y_i(\beta_0 + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i) \geq 1 \quad \forall \quad i$$

# Margin




if data are separable, $\exists$ infinitely many separating hyperplanes

argument: the one with the largest margin is the best

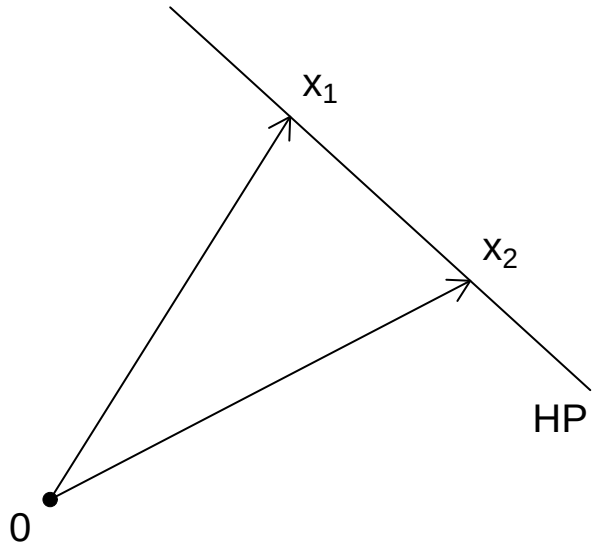
# Computing the Margin

(1) for any $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \text{HP}$,

$$\beta_0 + \boldsymbol{\beta}^{\text{T}}\boldsymbol{x}_1 = \beta_0 + \boldsymbol{\beta}^{\text{T}}\boldsymbol{x}_2 = 0,$$

so

$$\boldsymbol{\beta}^{\text{T}}(\boldsymbol{x}_1 - \boldsymbol{x}_2) = 0 \quad \Rightarrow \quad \boldsymbol{\beta} \perp \text{HP}$$




(2) for any $\boldsymbol{x}_0 \in \text{HP}$,

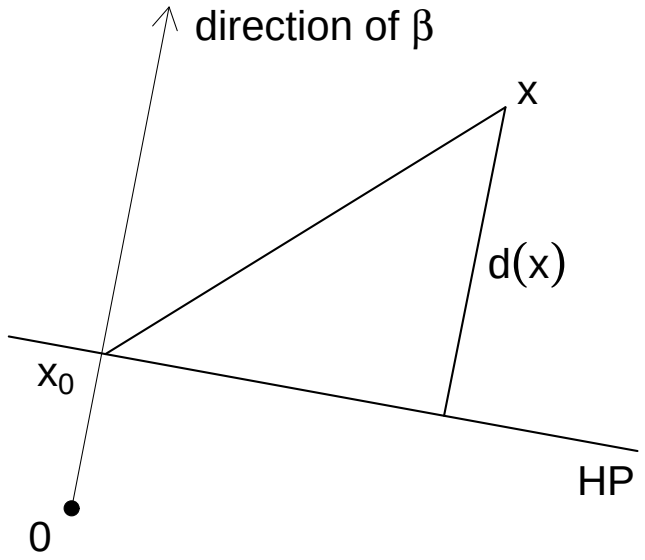
$$\beta_0 + \boldsymbol{\beta}^{\text{T}}\boldsymbol{x}_0 = 0 \quad \Rightarrow \quad \beta_0 = -\boldsymbol{\beta}^{\text{T}}\boldsymbol{x}_0$$

# Computing the Margin

(3) for any $\boldsymbol{x}$, to compute

$$d(\boldsymbol{x}) = \text{signed distance to HP},$$

simply grab any $\boldsymbol{x}_0 \in$ HP, and project $\boldsymbol{x} - \boldsymbol{x}_0$ onto direction of $\boldsymbol{\beta}$ (see figure), so




$$\begin{aligned} d(\boldsymbol{x}) &= \frac{\boldsymbol{\beta}^{\mathrm{T}}}{\|\boldsymbol{\beta}\|}(\boldsymbol{x} - \boldsymbol{x}_0) \\ &= \frac{1}{\|\boldsymbol{\beta}\|}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x} + \beta_0) \end{aligned}$$

by (2) above

# Computing the Margin

$$\begin{aligned}
\text{margin} &= 2 \times \min_i \left[ y_i d(\boldsymbol{x}_i) \right] \\
&= 2 \times \min_i \left[ \frac{y_i(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_i + \beta_0)}{\|\boldsymbol{\beta}\|} \right] \\
&= \frac{2}{\|\boldsymbol{\beta}\|}
\end{aligned}$$

$$\text{since} \quad \underbrace{y_i(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_i + \beta_0) \geq 1 \quad \forall \quad i}_{\text{separating HP}}$$

# Optimization Problem

$$\max_{\boldsymbol{\beta},\beta_0} \quad \frac{2}{\|\boldsymbol{\beta}\|} \quad \Leftrightarrow \quad \min_{\boldsymbol{\beta},\beta_0} \quad \frac{1}{2}\|\boldsymbol{\beta}\|^2$$

$$\text{s.t.} \quad y_i(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i + \beta_0) \geq 1 \quad \forall \quad i$$

assumes classes are separable ... for now

extend to non-separable case later

**Remark** Going after maximum margin is not only intuitive, but also backed by rich VC-theory due to Vapnik & Chervonenkis.

# Flavor of VC-Theory

- suppose [using hyperplanes (HP)]
  - all data $\boldsymbol{x} \in \mathbb{R}^d$ are confined within a ball of radius $r$
  - training data perfectly separated (ideal case) w/ margin $\Delta$
- then, w/ probability $1 - \delta$, error ($\varepsilon$) on test data bounded by

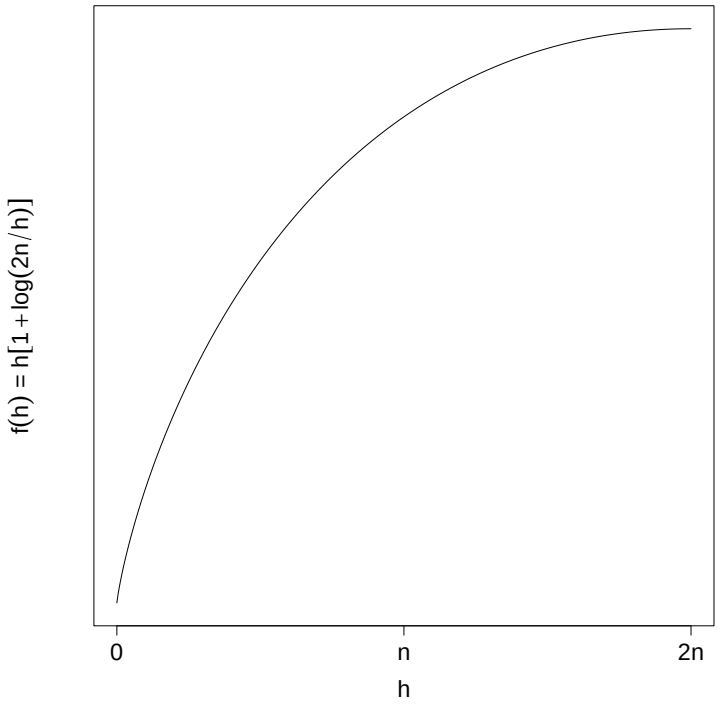
$$\varepsilon \leq 4 \times \frac{h[1 + \log(2n/h)] - \log(\delta/4)}{n}$$

  where

$$h \leq \min\left(\frac{r^2}{\Delta^2}, d\right) + 1$$

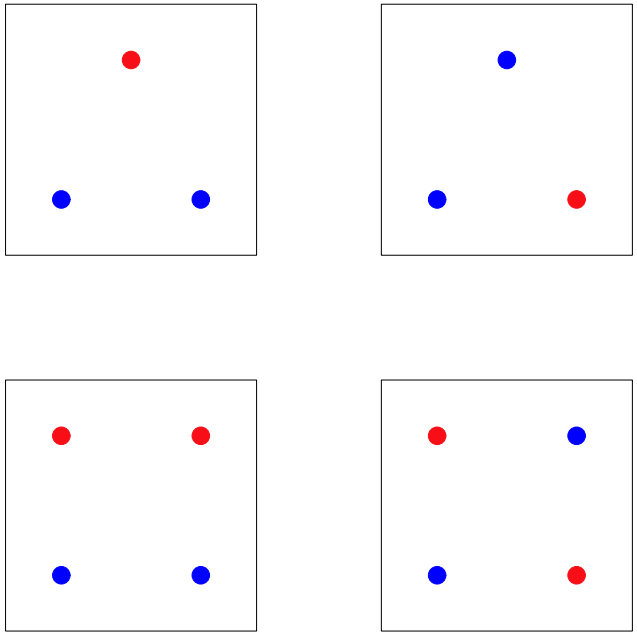
  is the VC dimension of the separating HP
- basically, $\Delta \uparrow \Rightarrow h \downarrow \Rightarrow \varepsilon \downarrow$ ... $\exists$ many variations

# Some Details




*Above*: $f(h) = h[1+\log(2n/h)]$ is an increasing function of $h$ on $(0, 2n)$.

*Below*: VC dimension = maximum number of points a function can "shatter", a measure of complexity. In 2D, HPs have VC dim = 3.



# Optimization

- Fermat ($\sim$ 1650): unconstrained optimization

$$\min_{\boldsymbol{x}} \quad f(\boldsymbol{x})$$

- Lagrange ($\sim$ 1800): equality constraints

$$\min_{\boldsymbol{x}} \quad f(\boldsymbol{x}) \quad \text{s.t.} \quad g_i(\boldsymbol{x}) = 0$$

- Kuhn-Tucker ($\sim$ 1950): inequality constraints

$$\min_{\boldsymbol{x}} \quad f(\boldsymbol{x}) \quad \text{s.t.} \quad \underbrace{g_i(\boldsymbol{x}) \geq 0}_{\text{main difficulty}} \quad \text{and} \quad \underbrace{h_j(\boldsymbol{x}) = 0}_{\substack{\text{treat ``as usual''} \\ \text{we'll ignore}}}$$

# The Primal Problem

- minimizing the Lagrangian function,

$$L(\boldsymbol{x}, \boldsymbol{\alpha}) \equiv f(\boldsymbol{x}) - \sum_i \alpha_i g_i(\boldsymbol{x}),$$

over $\boldsymbol{x}$, using $\alpha_i \geq 0$, gives lower bound on $f(\boldsymbol{x})$ for feasible $\boldsymbol{x}$:

$$\begin{aligned} L^*(\boldsymbol{\alpha}) &\equiv \inf_{\boldsymbol{x}} \left\{ f(\boldsymbol{x}) - \sum_i \alpha_i g_i(\boldsymbol{x}) \right\} \\ &\leq f(\boldsymbol{x}) - \sum_i \alpha_i g_i(\boldsymbol{x}) \quad [\text{having minimized over } \boldsymbol{x}] \\ &\leq f(\boldsymbol{x}) \quad [\alpha_i \geq 0 \text{ and } g_i(\boldsymbol{x}) \geq 0 \text{ for feasible } \boldsymbol{x}] \end{aligned}$$

- natural to ask for the best lower bound

# The Dual Problem

- maximizing $L^*(\boldsymbol{\alpha})$ over $\alpha_i \geq 0$ gives the greatest lower bound
- under "good" conditions, can get

$$\max_{\alpha_i \geq 0 \ \forall \ i} \quad L^*(\boldsymbol{\alpha}) \quad \Leftrightarrow \quad \min_{g_i(\boldsymbol{x}) \geq 0 \ \forall \ i} \quad f(\boldsymbol{x}),$$

  known as strong duality

- can see that, when this happens, must have

$$\alpha_i g(\boldsymbol{x}_i) = 0 \quad \forall \quad i,$$

  known as complementary slackness

**Exercise** What's the intuition behind complementary slackness?

# Karush-Kuhn-Tucker Conditions

Let $f(\boldsymbol{x})$ be a continuously differentiable function for $\boldsymbol{x} \in \mathbb{R}^d$. Let $S = \{\boldsymbol{x} \in \mathbb{R}^d : g_i(\boldsymbol{x}) \geq 0\}$ be a set of affine constraints. Define

$$L \equiv f(\boldsymbol{x}) - \sum_i \alpha_i g_i(\boldsymbol{x})$$

to be the Lagrangian. If $f$ has local minimum on $S$ at $\boldsymbol{x}^*$, then

(a) $$\frac{\partial L}{\partial \boldsymbol{x}}\Big|_{\boldsymbol{x}=\boldsymbol{x}^*} = 0 \quad \Rightarrow \quad f'(\boldsymbol{x}^*) = \sum_i \alpha_i g_i'(\boldsymbol{x}^*);$$

(b) $$\alpha_i \geq 0 \quad \text{for all} \quad i;$$

(c) $$\alpha_i g_i(\boldsymbol{x}^*) = 0 \quad \text{for all} \quad i.$$

If $f$ is also (strictly) convex and (a)–(c) are satisfied, then $\boldsymbol{x}^*$ is a (unique) global minimum.

# The KKT Conditions

$$\min_{\boldsymbol{\beta},\beta_0} \quad \frac{1}{2}\|\boldsymbol{\beta}\|^2 \quad \text{s.t.} \quad y_i(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i + \beta_0) \geq 1 \quad \forall \quad i$$

$$L \equiv \frac{1}{2}\|\boldsymbol{\beta}\|^2 - \sum_i \alpha_i \left[y_i(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i + \beta_0) - 1\right]$$

(a) $$\frac{\partial L}{\partial \boldsymbol{\beta}} = 0 \Rightarrow \boldsymbol{\beta} = \sum_i \alpha_i y_i \boldsymbol{x}_i; \quad \frac{\partial L}{\partial \beta_0} = 0 \Rightarrow \sum_i \alpha_i y_i = 0$$

(b) $$\alpha_i \geq 0 \quad \text{for all} \quad i$$

(c) $$\alpha_i \left[y_i(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i + \beta_0) - 1\right] = 0 \quad \text{for all} \quad i$$

plug (a) into $L$, and obtain $L^*(\boldsymbol{\alpha})$

# The Dual Problem

$$\max_{\boldsymbol{\alpha}} \quad L^*(\boldsymbol{\alpha})$$

$$= \quad \frac{1}{2}\|\boldsymbol{\beta}\|^2 - \sum_i \alpha_i \left[ y_i(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i + \beta_0) - 1 \right]$$

$$= \quad \sum_i \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j$$

$$\text{s.t.} \quad \underbrace{\sum_i \alpha_i y_i = 0}_{\text{KKT (a)}} \quad \text{and} \quad \underbrace{\alpha_i \geq 0}_{\text{KKT (b)}}$$

# Quadratic Programming

- dual problem can be written as

$$\max_{\boldsymbol{\alpha}} \quad \mathbf{1}^{\mathrm{T}}\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{D}_{\boldsymbol{y}}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}_{\boldsymbol{y}}\boldsymbol{\alpha} \quad \text{s.t.} \quad \boldsymbol{y}^{\mathrm{T}}\boldsymbol{\alpha} = 0, \quad \boldsymbol{I}\boldsymbol{\alpha} \geq \mathbf{0}$$

$$\text{where} \quad \boldsymbol{D}_{\boldsymbol{y}} = \begin{bmatrix} y_1 & - & - \\ - & \ddots & - \\ - & - & y_n \end{bmatrix}_{n\times n} \qquad \text{and} \quad \boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_n^{\mathrm{T}} \end{bmatrix}_{n\times d}$$

- a standard quadratic programming problem of the form

$$\max_{\boldsymbol{\alpha}} \quad \boldsymbol{p}^{\mathrm{T}}\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{\alpha} \quad \text{s.t.} \quad \boldsymbol{A}\boldsymbol{\alpha} = \boldsymbol{b}, \quad \boldsymbol{C}\boldsymbol{\alpha} \geq \boldsymbol{d}$$

# "What's in a name?"

"That which we call a support vector machine by any other name would sound as esoteric."

- KKT(b)+(c) imply

$$\alpha_i > 0 \quad \Rightarrow \quad y_i(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i + \beta_0) - 1 = 0$$

and

$$y_i(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i + \beta_0) - 1 > 0 \quad \Rightarrow \quad \alpha_i = 0$$

- training observation $\boldsymbol{x}_i$ called a support vector if $\alpha_i > 0$;

$$\boldsymbol{\beta} = \sum_i \alpha_i y_i \boldsymbol{x}_i$$

only affected by terms with $\alpha_i > 0$ — SVs are on the margin

# SVM Steps

1. Solve dual problem (quadratic programming) for $\alpha_i$.

2. Let
$$\boldsymbol{\beta} = \sum_i \alpha_i y_i \boldsymbol{x}_i.$$

3. Solve for $\beta_0$. [How?]

4. Predict with
$$f(\boldsymbol{x}) = \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_i + \beta_0 = \beta_0 + \sum_i \alpha_i y_i \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}.$$

# SVM Steps

1. Solve dual problem (quadratic programming) for $\alpha_i$.

2. Let
$$\boldsymbol{\beta} = \sum_i \alpha_i y_i \boldsymbol{x}_i.$$

3. Solve for $\beta_0$ using any support vector [those with $\alpha_i > 0$, since all SVs satisfy $y_i(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i + \beta_0) - 1 = 0$].

4. Predict with
$$f(\boldsymbol{x}) = \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i + \beta_0 = \beta_0 + \sum_i \alpha_i y_i \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}.$$

# Non-separable Case

$$\min_{\boldsymbol{\beta},\beta_0} \quad \frac{1}{2}\|\boldsymbol{\beta}\|^2 + \gamma \sum_i \xi_i$$

$$\text{s.t.} \qquad y_i(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i + \beta_0) \geq 1 - \xi_i \quad \forall \quad i$$

$$\xi_i \geq 0 \quad \forall \quad i$$

[different formulations possible]

# The KKT Conditions

$$L \equiv \frac{1}{2}\|\boldsymbol{\beta}\|^2 + \gamma \sum_i \xi_i - \sum_i \alpha_i \left[ y_i(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i + \beta_0) - 1 + \xi_i \right] - \sum_i \lambda_i \xi_i$$

(a) $\dfrac{\partial L}{\partial \boldsymbol{\beta}} = 0 \Rightarrow \boldsymbol{\beta} = \sum_i \alpha_i y_i \boldsymbol{x}_i$; $\dfrac{\partial L}{\partial \beta_0} = 0 \Rightarrow \sum_i \alpha_i y_i = 0$;

$\dfrac{\partial L}{\partial \xi_i} = 0 \Rightarrow \gamma - \alpha_i - \lambda_i = 0$ or $\alpha_i = \gamma - \lambda_i$

(b) $\alpha_i \geq 0$ for all $i$; $\lambda_i \geq 0$ for all $i$

(c) $\alpha_i \left[ y_i(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i + \beta_0) - 1 \right] = 0$ for all $i$;

$\lambda_i \xi_i = 0$ for all $i$

# Two Typs of Support Vectors

- extra KKT (a)+(b)+(c):

$$\alpha_i = \gamma - \lambda_i; \quad \lambda_i \geq 0; \quad \lambda_i \xi_i = 0$$

- together, they imply

$$\alpha_i < \gamma \quad \Rightarrow \quad \lambda_i > 0 \quad \Rightarrow \quad \xi_i = 0$$

and

$$\xi_i > 0 \quad \Rightarrow \quad \lambda_i = 0 \quad \Rightarrow \quad \alpha_i = \gamma$$

- support vectors

| | | |
|---|---|---|
| either | on the margin | $(\xi_i = 0, \alpha_i < \gamma)$ |
| or | over the margin | $(\xi_i > 0, \alpha_i = \gamma)$ |

# Non-separable Case

## Exercises

(a) Show that, in this particular formulation of the non-separable case, the dual remains the same as in the separable case, except for an additional set of constraints that $\alpha_i \leq \gamma$ for all $i$.

(b) Specify the corresponding dual if the non-separable case is formulated as

$$\min_{\boldsymbol{\beta},\beta_0} \quad \|\boldsymbol{\beta}\|^2 + \gamma \sum_i \xi_i^2$$

$$\text{s.t.} \qquad y_i(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i + \beta_0) \geq 1 - \xi_i \quad \forall \quad i$$

instead. How to solve the corresponding dual?

(c) Solve for $\beta_0$ in the non-separable case. [both formulations]

# A Different Point of View

$$\min_{\boldsymbol{\beta},\beta_0} \quad \frac{1}{2}\|\boldsymbol{\beta}\|^2 + \gamma \sum_i \xi_i$$

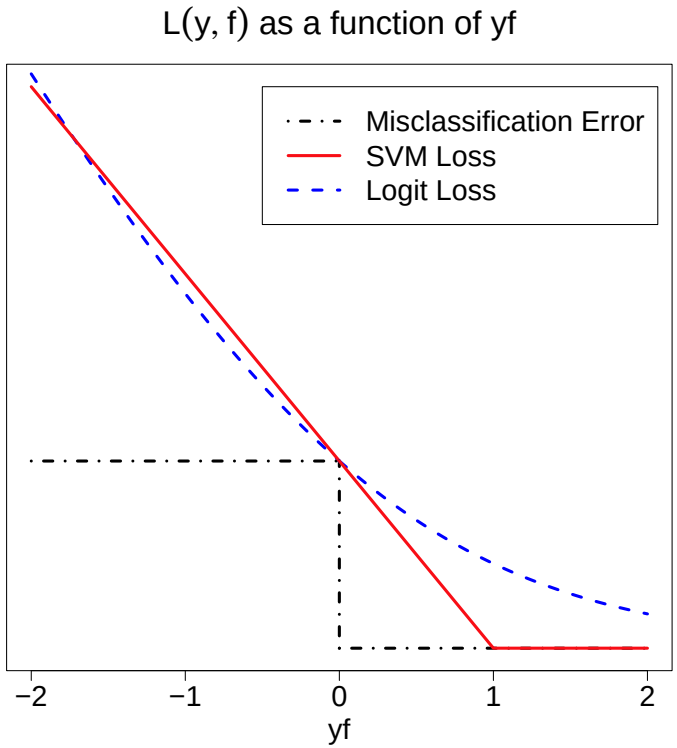
$$\text{s.t.} \quad y_i(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i + \beta_0) \geq 1 - \xi_i \quad \text{and} \quad \xi_i \geq 0 \quad \forall \quad i$$

$$\Updownarrow$$

$$\min_{\boldsymbol{\beta},\beta_0} \quad \underbrace{\sum_{i=1}^{n} [1 - y_i(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i + \beta_0)]_+}_{\text{loss}} + \underbrace{\lambda\|\boldsymbol{\beta}\|^2}_{\text{penalty}}$$

$$\text{where} \quad [z]_+ = \begin{cases} z, & z > 0; \\ 0, & z \leq 0 \end{cases}$$

# The "Hinge" Loss




$$L[y, f(\boldsymbol{x})]$$

Error : $I[yf(\boldsymbol{x}) < 1]$

Logit : $\log_2[1 + e^{-yf(\boldsymbol{x})}]$

SVM : $[1 - yf(\boldsymbol{x})]_+$

in a way, SVM not too different from ridge logistic regression

# Summary

- Part 1
  - separating hyperplanes; margin; VC-theory
  - optimization; primal; dual; KKT conditions
  - quadratic programming
  - support vector machine (SVM); "hinge" loss
- Part 2 (next, after short break)
  - kernel SVM; reproducing kernel Hilbert space
  - kernel ridge regression
  - kernel canonical correlation analysis