Support Vector Machines

Virgil Pavlu October 8, 2008

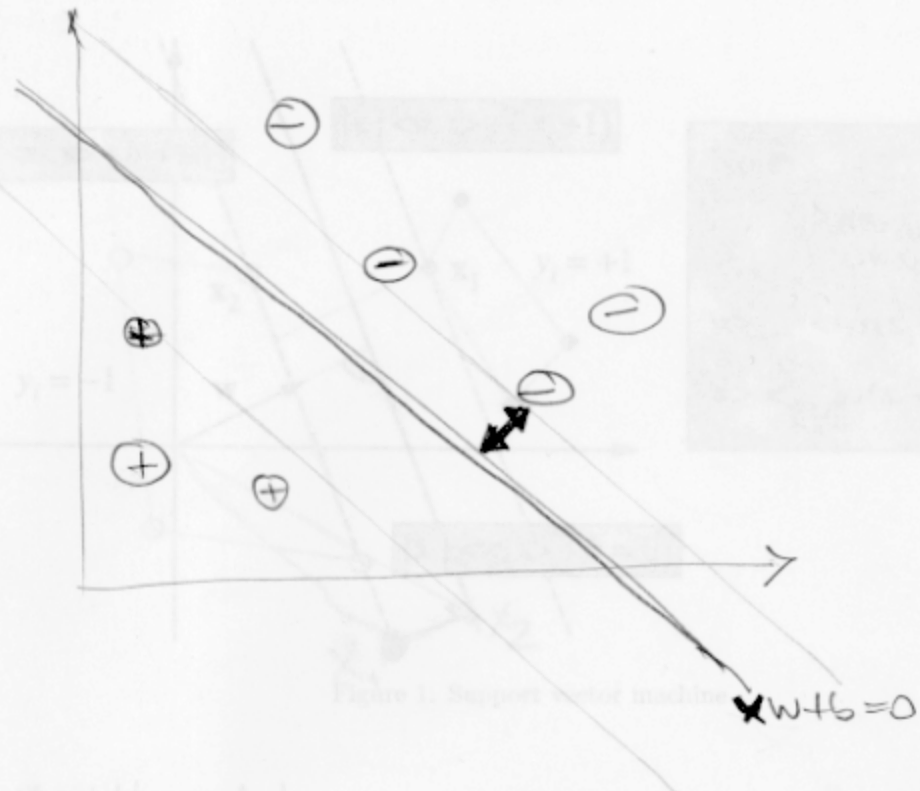
1 Linear discrimination

-we have talked about linear/logistic regression and the perceptron. There are other methods, essentially variations of gradient descent with specific objective functions.

-lets take a closer look at a hyperplane separating the classes in a binary problem.

= linear/logistic regression
- perceptron
- others (gradient descent)

2 Geometry of the hyperplanes (separable case)



line $wx+b=0$ can be written as $(wc)x+(bc)=0$ $\forall c$
so we settle for c such that
$\min |wx+b| = 1$ for x in X

3 Margin

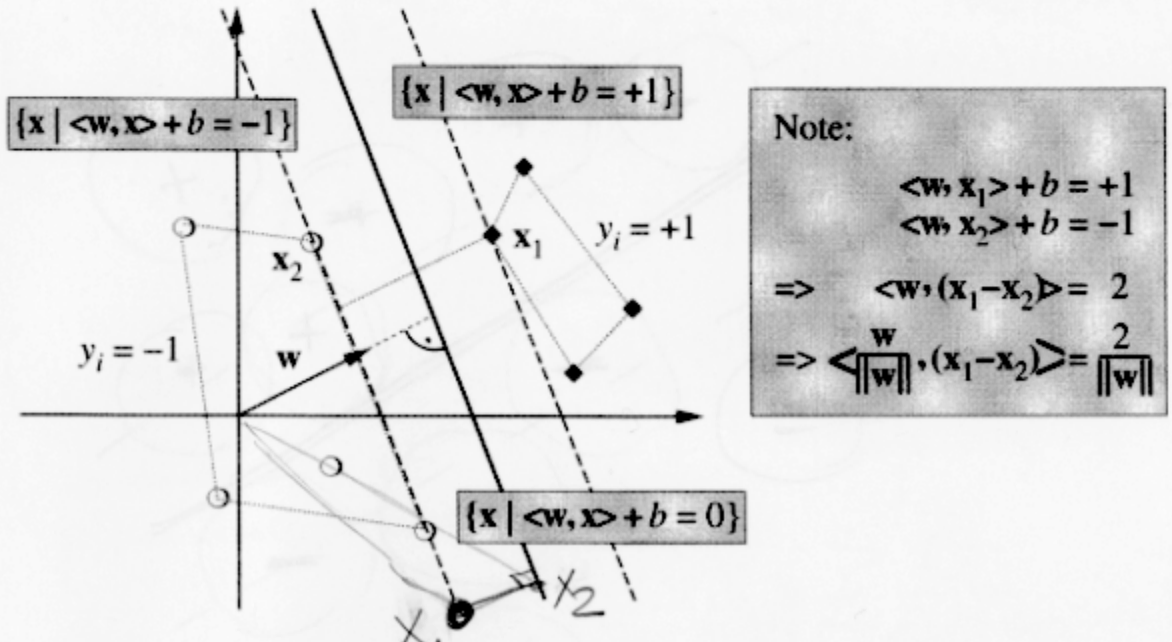


Figure 1: Support vector machine

$$\left.\begin{array}{l} X_1 W + b = 1 \\ X_2 W + b = 0 \end{array}\right| \Rightarrow (X_2 - X_1) W = -1 \Rightarrow \|X_2 - X_1\| = \frac{1}{\|w\|}$$

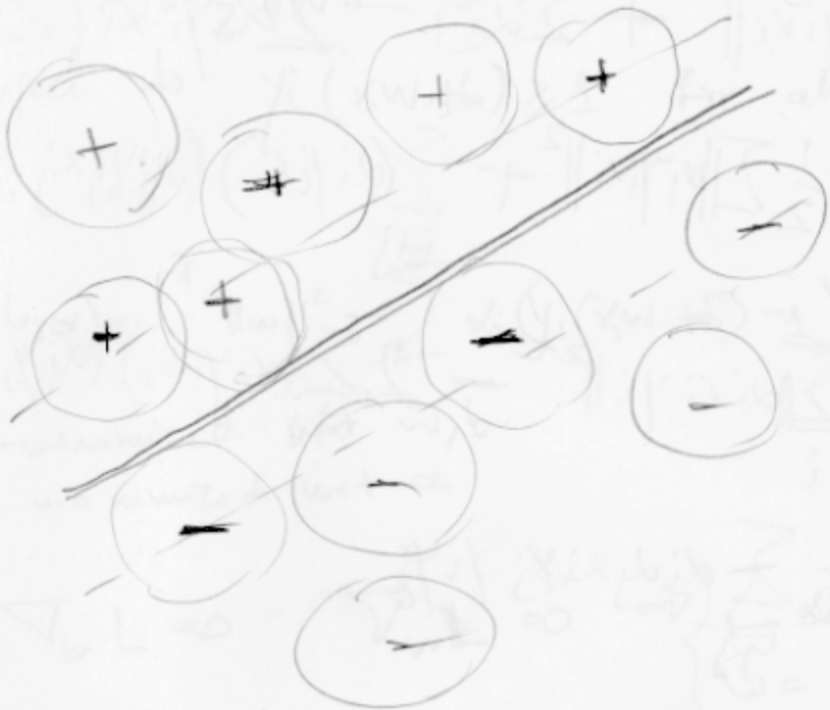
margin:

$$\rho(X_1, y) = \frac{y(X_1 w + b)}{\|w\|} = \frac{y\left(w x_1 + b - (w x_2 + b)\right)}{\|w\|} =$$

$$= \frac{y\left((X_1 - X_2) w\right)}{\|w\|} = y(x_1 - x_2) \cdot$$

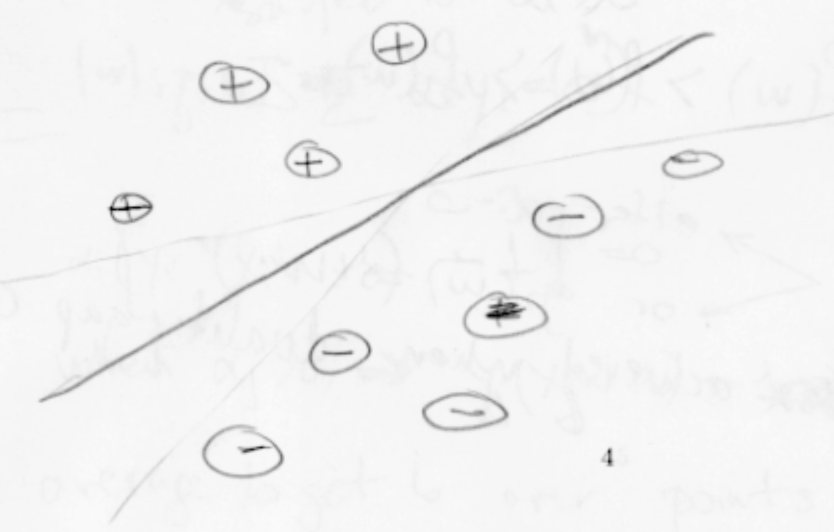
$$\text{"size"} = \frac{1}{\|w\|} = \|X_2 - X_1\| \cdot |y|$$

4 Large margin classification



- perceptron with data+noise ⇒ better hyperplane

freedom do modify (slightly) the hyperplane and still get a good generalization



5 SVM - optimal hyperplane wrt margin

primal opt | minimize $\frac{1}{2}\|w\|^2$ [$\Leftrightarrow$ maximize margin]
subject to $y_i(xw+b) \geq 1$ for all i

if $y_i(x_i w+b) < 1 \Rightarrow L \to \infty$ so w, b forced to do $y_i(x_i w+b) \geq 1$

$$L(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \sum \alpha_i \left(y_i(x_i w+b) - 1\right)$$

- minimized wrt w, b
- maximized wrt α

$$\nabla_b L = 0 \quad \nabla_w L = 0 \Rightarrow \begin{cases} \sum \alpha_i y_i = 0 \\ w = \sum \alpha_i y_i x_i \end{cases} \rightarrow \text{lin. comb}$$

remember $\alpha_i > 0 \Leftrightarrow y_i(xw+b) = 1 \Leftrightarrow (x_i, y_i)$ SUPPORT VECTOR

dual: maximize
$$W(\alpha) = \sum \alpha_i - \frac{1}{2}\sum \alpha_i\alpha_j y_i y_j x_i x_j^T$$
subject to $\alpha_i \geq 0$
$$\sum \alpha_i y_i = 0$$

EASIER PB (no w)
- quadratic solver
- heuristic

$$\alpha_i\left[y_i(x_i w+b) - 1\right] = 0$$

when $\alpha_j > 0 \Rightarrow y_j(x_j w+b) = 1 \Rightarrow \sum_i \alpha_i y_i x_i x_j^T + b = y_j$

average to get b over points with $\alpha_j > 0$

$$\frac{1}{2}\|w\|^2 = \frac{1}{2}\Big\|\sum \alpha_i y_i x_i\Big\|^2$$

$$L = \frac{1}{2}\Big\|\sum \alpha_i y_i x_i\Big\|^2 + \sum \alpha_i - \sum \alpha_i y_i x_i \Big(\sum \alpha_j y_j x_j\Big)$$

$$= \sum \alpha_i + \frac{1}{2}\sum \|\alpha_i y_i x_i\|^2 + \sum_{i \neq j} (\alpha_i y_i x_i)(\alpha_j y_j x_j$$

$$= \sum_i \|\alpha_i x_i y_i\|^2 - 2\sum_{i \neq j} (\alpha_i y_i x_i)(\alpha_j y_j x_j)$$

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

$\sum \alpha_i g_i(w)$ $\quad g_i =$ constraints $\quad$ Karush Kuhn Tucker

$\|$ duality gap (or KKT gap)

$\tilde{w}$ solution $\Rightarrow$ any w, α with $\left\{\begin{array}{l} \alpha > 0 \\ \partial_w L = 0 \\ \partial_\alpha L = 0 \end{array}\right\}$ has

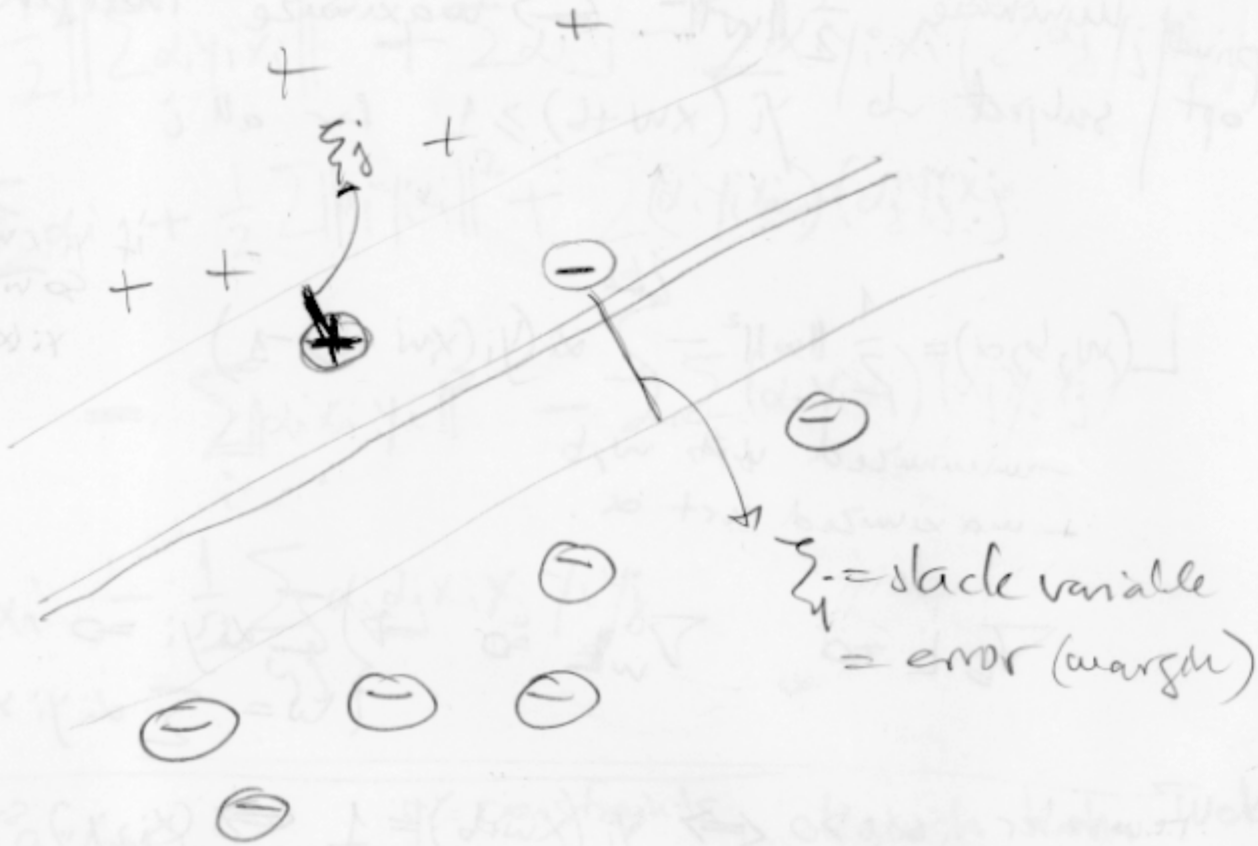
$$f(w) > f(\tilde{w}) > f(w) + \sum \alpha_i g_i(w)$$

---

best $\tilde{w}, \tilde{\alpha}$:

$\sum \tilde{\alpha}_i g_i(\tilde{w}) = 0$ $\rightarrow$ either $\alpha_i = 0$ / or $g_i(\tilde{w}) = 0$

so $f(\tilde{w})$ is ~~max~~ achieved when duality gap closes.



ξ_j
ξ_i = slack variable
= error (margin)

6 Non-separable data

-slack variables

<u>-soft margin hyperplanes</u>

constraints $y_i(\cancel{x}w+b) \geq 1-\xi_i$

ξ_i = slack variable

$$\begin{cases} \text{minimize} & \frac{1}{2}\|w\|^2 + \frac{C}{m}\sum \xi_i \\ \text{subject to} & \xi_i \geq 0 \\ & y_i(xw+b) \geq 1-\xi_i \end{cases}$$

dual

$$\begin{cases} \text{maximize} & \bar{W}(\alpha) = \sum \alpha_i - \frac{1}{2}\sum \alpha_i \alpha_j y_i y_j x_i x_j^T \\ \text{subj to} & 0 \leq \alpha_i \leq \frac{C}{m} \\ & \sum \alpha_i y_i = 0 \end{cases}$$

to compute b,

$$L(\quad) = \frac{1}{2}\|w\|^2 - \sum_{i \in SV} \alpha_i \left(y_i(xw+b) - 1 + \xi\right) + \frac{C}{m}\sum \xi$$

$$\boxed{\text{why } \alpha < \frac{C}{m} ?}$$

$$-\alpha(yxw + yb - 1 + \xi) + \frac{C}{m}\xi$$

⇒ coefficient of ξ is $-\alpha + \frac{C}{m}$

$\xi > 0$; for optimization (minimization) to make sense: $-\alpha + \frac{C}{m} > 0$, otherwise we can send $\xi \to +\infty$ and minimize to $-\infty$.

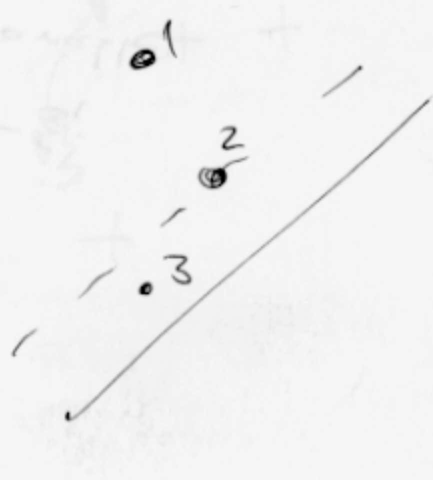
In fact
$$L(\quad) = \frac{1}{2}\|w\|^2 - \sum \alpha_i \left(y_i(xw+b) - 1 + \xi\right) + \frac{C}{m}\sum \xi_i - \beta_i \xi_i$$

→ Lag. multiplier for constraint $\xi_i \geq 0$

$$\frac{\partial L}{\partial \xi} = -\alpha - \beta + C/m = 0 \Rightarrow C/m = \alpha_i + \beta_i$$

KKT: $\beta_i \xi_i = 0$ co

1. $\xi_i = 0$, $\alpha_i = 0$, $\beta_i = 0$?
2. $\xi_i = 0$, $0 < \alpha_i < C/m$, $\beta_i > 0$?
3. $\xi_i > 0$, $\alpha_i = \frac{C}{m}$, $\beta_i = 0$



- SVM recap: margin, support vector. margin $\simeq \frac{1}{\|w\|}$

- Primal: minimize $\frac{1}{2}\|w\|^2$
  subject to $y_i(xw+b) \geq 1$

- Lagrangian $\begin{cases} L(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \sum \alpha_i (y_i(x_i w+b)-1) \\ \text{minimize with respect to } w, b \\ \text{max wrt } \alpha. \end{cases}$

- KKT theorem: nec + suf condition for solution

$\partial_{w,b} L = 0; \quad \alpha_i(y_i(x_i w+b)-1) = 0$ → Sp. vector / → irrelev. constrain

$\alpha_i \geq 0 \quad ; \quad y_i(xw+b) \geq 1$

$\tilde{w}, \tilde{\alpha}$ solution $\Rightarrow L(w,\tilde{\alpha}) \geq L(\tilde{w},\tilde{\alpha}) \geq L(\tilde{w},\alpha)$.

- saddle points, dual problem.
  max: $W(\alpha) = \sum \alpha_i - \frac{1}{2}\sum \alpha_i \alpha_j y_i y_j x_i x_j^T$
  subject to $\alpha_i \geq 0, \ \sum \alpha_i y_i = 0$.

- Non separable data: soft margin hyperplane
  ξ_i = slack variables

Primal $\begin{cases} \text{minimize} & \frac{1}{2}\|w\|^2 + \frac{C}{m}\sum \xi_i \\ \text{subject to} & \xi_i \geq 0, \ y_i(x_i w+b) \geq 1-\xi_i \end{cases}$

Dual $\begin{cases} \text{max} & W(\alpha) = \sum \alpha_i - \frac{1}{2}\sum \alpha_i \alpha_j y_i y_j x_i x_j^T \\ \text{subject to} & 0 \leq \alpha_i \leq \frac{C}{m} \end{cases}$

- Quadratic Solvers. SMO = sequential minimal optimizat

coordinate ascent: Loop
maxim one coordinate at the time.

$\alpha_i := \arg\max W(\alpha_1, \alpha_2, \ldots \alpha_{i-1}, \hat{\alpha}_i, \alpha_{i+1}, \ldots \alpha_m)$

order 1, 2, .. m MATTERS

SMO: Loop

choose a pair α_i, α_j (by heuristic → max update)

$\alpha_i, \alpha_j = \arg\max_{i,j} W(\alpha)$

efficient argmax:

$\alpha_1 y_1 + \alpha_2 y_2 = T$ constant (because of constraint $\sum \alpha_i y_i = 0$).

$\alpha_1 = (T - \alpha_2 y_2) y_1$

$W = W((T - \hat{\alpha}_2 y_2) y_1, \hat{\alpha}_2, \ldots \ldots)$ quadratic in α_2

→ interior point methods:
- Solve KKT equations, iteratively
- need a trick on $\alpha \xi = 0 \Rightarrow \alpha \xi = \mu$, make $\mu \to 0$
- follow path primal - dual
- cholesky decomposition, LAPACK (lin. alg library)

- Chunking working sets: try to identify a likely support vector set, only work with this data