BIOSTAT 527

5/1/23

Lecture 11

SVM

Lecture V: Support Vector Machines and Kernel Machines

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

April, 2023

Linear SVM's

The margin and the expected classification error
Maximum Margin Linear classifiers
Linear classifiers for non-linearly separable data

optimization matters !!

Analysis of solution

Non linear SVM

The "kernel trick"
Kernels
Prediction with SVM

Extensions

L_1 SVM
Multi-class and One class SVM
SV Regression

Reading AoNPS Ch.: Ch. 12.1–3, HTF Ch.: Ch 14 (14.1,14.2–14.2.4 kernels, 14.4 and equations (14.28,14.29) kernel trick, 14.5.1.–3 Support Vector Machines)7.1–7.4, 7.7
Additional Reading: C. Burges - "A tutorial on SVM for pattern recognition"
These notes: Appendices (convex optimization) are optional.

Data Linearly separable $\Rightarrow$ ∞ linear classifiers

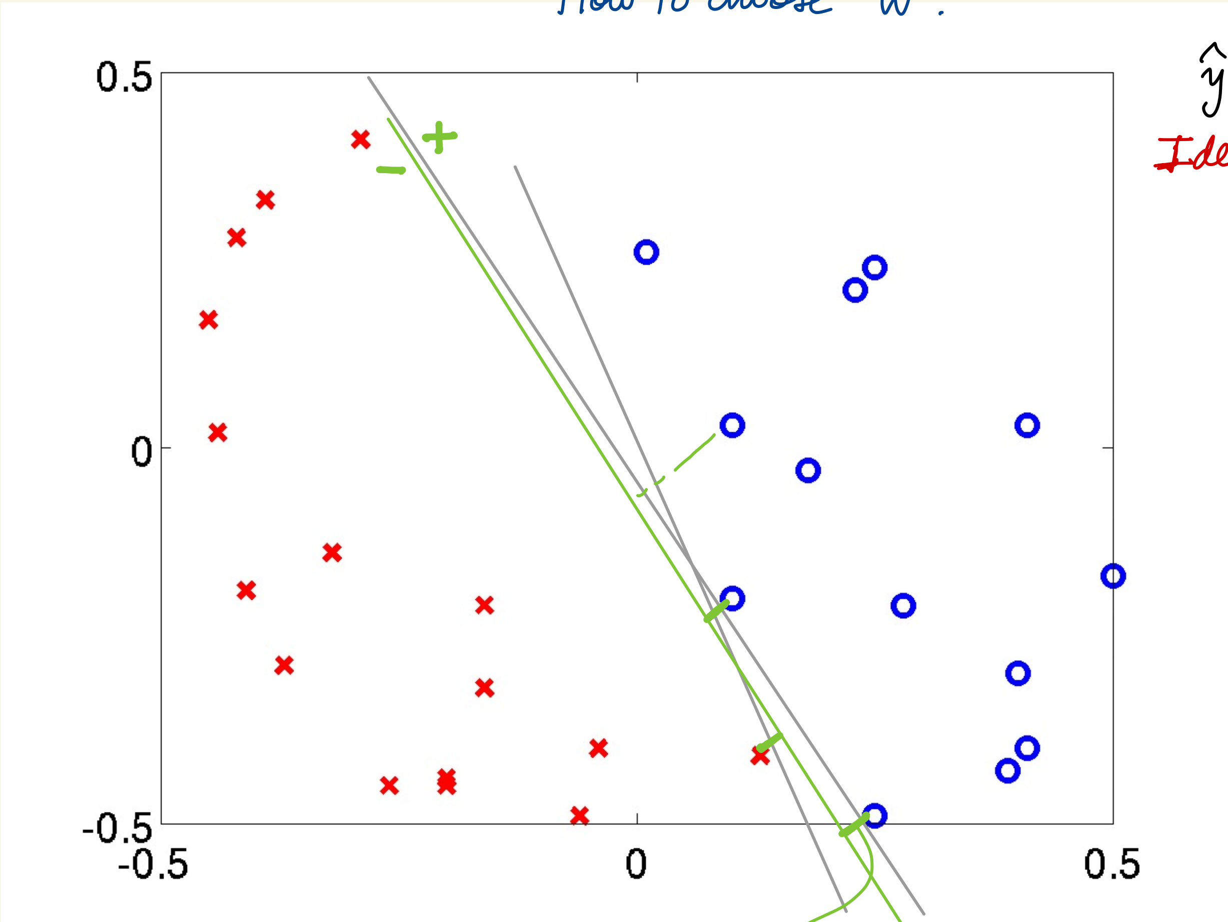
How to choose $\hat{w}$?

$f(x) = w^T x + b$

$x, w \in \mathbb{R}^d$

$\hat{y} = \operatorname{sign} f(x)$

Idea 1. Max Margin



ρ = margin of $f_{w,b}$

Parametric

LDA

LR

Perceptron

Second idea

The **Second idea** is to formulate (10) as a **quadratic** optimization problem.

primal variables

constraints

$$\min_{w,b} \underbrace{\frac{1}{2}||w||^2}_{\text{objective}} \text{ s.t } \boxed{y^i(w^T x^i + b) \geq 1} \text{ for all } i = 1:n \tag{13}$$

convex opt. pb.

This is the **Linear SVM (primal) optimization problem**

- This problem has a strongly convex **objective** $||w||^2$, and **constraints** $y^i(w^T x^i + b)$ linear in (w, b).
- Hence this is a convex problem, and can be studied with the tools of convex optimization.

Optimization in the analysis of SVM — pb.

solution

convex opt. pb

Objective = convex fction

Constraints = linear functions

$\Uparrow$

Feasible set is convex

{ w, b that satisfy constraints }

Convex function

- $\nabla^2 f(x) \succcurlyeq 0$ for all x

  Hessian positive definite

  must exist

- 

- Convex set $A \iff$

  $x, x' \in A \Rightarrow \underbrace{tx + (1-t)x'}_{\text{line segment } [x, x']} \in A$

  $t \in [0,1]$

SVM

(P) $\min_{w,b} \frac{1}{2}\|w\|^2$ s.t. $\overbrace{-\left(y^i(w^Tx^i+b)-1\right)}^{g_i(w,b)} \le 0 \leftarrow \alpha_i$

standard form $\rightarrow \alpha_i \ge 0$

$x, w \in \mathbb{R}^d$

$d+1$ variables

$w \in \mathbb{R}^d$, $b \in \mathbb{R}$

n constraints

Lagrange function

$$L(\underbrace{w, b}_{\text{primal variables}}, \underbrace{\alpha_{1:n}}_{\text{dual variables}}) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i \Big[\underbrace{-y^i(w^Tx^i+b)+1}_{\text{constraint}}\Big]$$

KKT Conditions for optimum

$*$ $\frac{\partial L}{\partial w} = 0 = w - \sum_{i=1}^{n} \alpha_i y^i x^i \Rightarrow \boxed{w = \sum_{i=1}^{n} \alpha_i y^i x^i}$

$\frac{\partial L}{\partial b} = 0$

$\alpha_i[\cdots] = 0 \Rightarrow \begin{cases} \alpha_i = 0 \text{ and } [\cdots] \neq 0 \leftarrow g_i < 0 \quad \text{slack} \\ \alpha_i > 0 \text{ and } [\cdot\cdot] = 0 \leftarrow \text{constraint is tight} \end{cases}$

can't increase g_i

□ $\frac{\partial L}{\partial \alpha_i} = 0$ if $\alpha_i > 0$ $\quad \alpha_i = 0 = [\cdot\cdot]$ w.p. 0 ignore it!

$$W = \sum_{i=1}^{n} \alpha_i y^i x^i \implies$$

- w = linear combination of $x^{1:n}$

Ex: Logistic regression

when is $\hat{\beta} = \sum \alpha_i x^i$?

$\downarrow$

d parameters ? n params?

NO $\#\text{params} \cong \# \{\alpha_i > 0\}$

- if $\alpha_i = 0$ (SLACK) $\Rightarrow$ w does not depend on (x^i, y^i)
- $\alpha_i > 0 \implies x^i$ is support vector
[- if few $\alpha_i > 0$ then solution is sparse]

$$\frac{\partial L}{\partial b} = 0 = \left[-\sum_{i=1}^{n} \alpha_i y_i \right] \leftarrow \text{constraint on } \alpha_{1:n} \text{ !!}$$

for SV $\alpha_i > 0$

$$y^i (w^T x^i + b) = 1$$

$$\Rightarrow \left[b = y^i - w^T x^i \in \mathbb{R} \right]$$

same for all i S.V.

Solving SVM problem

- plug in w, constraint $\sum y^i \alpha_i = 0$ in L

Dual SVM problem

$$(D) \quad \max_{\alpha_{1:n}} \; \underline{1^T\alpha - \frac{1}{2}\alpha^T \tilde{G} \alpha} \quad \text{s.t} \quad \alpha_i \geq 0, \quad \sum_{i=1}^{n} \alpha_i y^i = 0$$

n variables ↑

quadratic objective + linear constraints

Dual objective

$$-\left(\sum_i \alpha^i y^i\right) b = 0$$

$$\frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i \left[- y^i (w^T x^i + b) + 1 \right] = 1^T\alpha - \frac{1}{2}\alpha^T \tilde{G} \alpha$$

$$\sum_{i=1}^{n} \alpha_i = 1^T \alpha$$

$$-\left(\sum \alpha_i y^i x^i\right)^T w = -\|w\|^2$$

$$1 = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \in \mathbb{R}^n$$

$$w^T w$$

$$\|w\|^2 = \left\| \sum \alpha_i y^i x^i \right\|^2 = \alpha^T \tilde{G} \alpha$$

$$\alpha = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} \qquad G = \left[(x^i)^T x^j \right]_{i,j=1:n} \leftarrow \text{Gram matrix} \qquad \tilde{G} = \left[y^i y^j (x^i)^T x^j \right]_{i,j=1:n}$$

Optimization with Lagrange multipliers

[2] The **Lagrangean** of (13) is

$$L(w, b, \alpha) \;=\; \frac{1}{2}||w||^2 - \sum_i \alpha_i [y^i (w^T x^i + b) - 1]. \tag{16}$$

[KKT conditions]

At the optimum of (13)

$$w \;=\; \sum_i \alpha_i y^i x^i \quad \text{with } \alpha_i \geq 0 \tag{17}$$

and $b = y^i - w^T x^i$ for any i with $\alpha_i > 0$.

- **Support vector** is a data point x^i such that $\alpha_i > 0$.
- According to (17), the final decision boundary is determined by the support vectors (i.e. does not depend explicitly on any data point that is not a support vector).

[2]The derivations of these results are in the Appendix

Dual SVM optimization problem

- Any convex optimization problem has a **dual** problem. In SVM, it is both illuminating and practical to solve the dual problem.
- The dual to problem (13) is

$$\max_{\alpha_{1:n}} \sum_i \alpha_i - \frac{1}{2} \sum_i \alpha_i \alpha_j y^i y^j x^{i\,T} x_j \text{ s.t } \alpha_i \geq 0 \text{ for all } i \text{ and } \sum_i \alpha_i y^i = 0. \tag{18}$$

- This is a quadratic problem with n variables on a convex domain.
- Dual problem in matrix form
  - Denote $\alpha = [\alpha_i]_{i=1:n}$, $y = [y^i]_{i=1:n}$, $G_{ij} = x^{i\,T} x_j$, $\bar{G}_{ij} = y^i y^j x^{i\,T} x_j$, $G = [G_{ij}] \in \mathbb{R}^{n\times n}$, $\bar{G} = [\bar{G}_{ij}] \in \mathbb{R}^{n\times n}$.

$$\max_{\alpha\in\mathbb{R}^n} 1^T\alpha - \frac{1}{2}\alpha^T \bar{G}\alpha \quad \text{s.t } \alpha \succeq 0 \text{ and } y^T\alpha = 0. \tag{19}$$

- $g(\alpha) = 1^T\alpha - \frac{1}{2}\alpha^T\bar{G}\alpha$ is the **dual objective function**
- G is called the **Gram matrix** of the data. Note that $\bar{G} = \mathrm{diag}\,\{y^{1:n}\}^T G \mathrm{diag}\,\{y^{1:n}\}$.

- At the dual optimum
  - $\alpha_i > 0$ for constraints that are satisfied with equality, i.e. **tight**
  - $\alpha_i = 0$ for the **slack** constraints

Non-linearly separable problems and their duals

regularization constant

The **C-SVM**

$$\text{minimize}_{w,b,\xi} \quad \frac{1}{2}||w||^2 + C\boxed{\sum_i \xi_i} \tag{20}$$

$$\text{s.t.} \quad y^i(w^T x^i + b) \geq 1 - \boxed{\xi_i}$$

$$\boxed{\xi_i} \geq 0$$

slack variables

In the above, ξ_i are the slack variables. Dual[3]:

$1^T\alpha - \frac{1}{2}\alpha^T\tilde{G}\alpha$

$$\text{maximize}_\alpha \quad \sum_i \alpha_i - \frac{1}{2}\sum_i \alpha_i\alpha_j y^i y_j x^{iT} x_j \tag{21}$$

$$\text{s.t.} \quad \boxed{C \geq \alpha_i} \geq 0 \text{ for all } i$$

$$\sum_i \alpha_i y^i = 0$$

new constraint

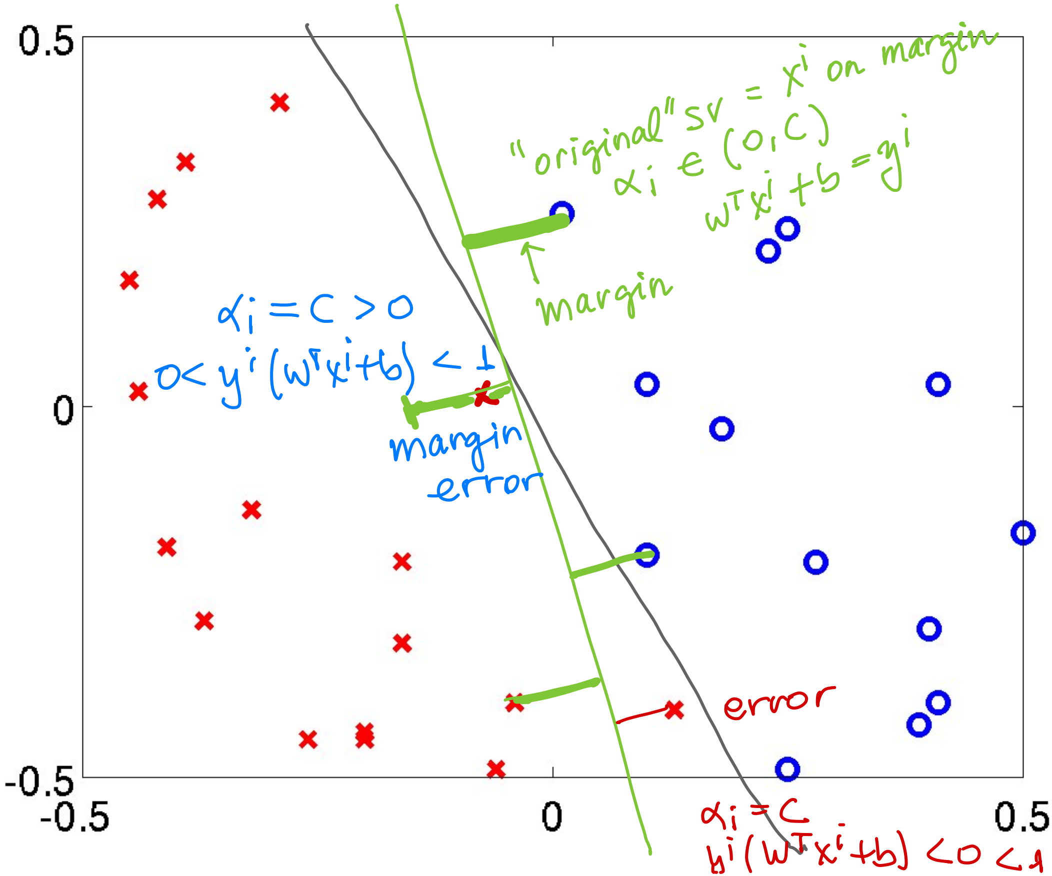
$\Rightarrow$ two types of SV

- $\alpha_i < C$ data point x^i is "on the margin" $\Leftrightarrow$ $y^i(w^T x^i + b) = 1$ (original SV) $\xi_i = 0$
- $\boxed{\alpha_i = C}$ data point x^i cannot be classified with margin 1 (**margin error**) $\Leftrightarrow$ $y^i(w^T x^i + b) < 1$ $\xi_i > 0$
- $\alpha_i = 0$ x^i not s.v.

$y^i(w^T x^i + b) < 1$ → < 0 error

→ ≥ 0 proper margin error

[3] Lagrangean $L(w, b, \xi, \alpha, \mu) = \frac{1}{2}||w||^2 + C\sum_i \xi_i - \sum_i \alpha_i[y^i(w^T x^i + b) - 1 + \xi_i] - \sum_i \mu_i \xi_i$ with $\alpha_i \geq 0,\ \xi_i \geq 0,\ \mu_i \geq 0$



0.5
0
-0.5
-0.5
0
0.5
"original" sv = x^i on margin
$\alpha_i \in (0, C)$
$w^T x^i + b = y^i$
margin
$\alpha_i = C > 0$
$0 < y^i(w^T x^i + b) < 1$
margin error
error
$\alpha_i = C$
$y^i(w^T x^i + b) < 0 < 1$

The ν-SVM

$$\text{minimize}_{w,b,\xi,\rho} \quad \frac{1}{2}||w||^2 - \nu\rho + \frac{1}{n}\sum_i \xi_i \tag{22}$$

$$\text{s.t.} \quad y^i(w^T x^i + b) \ \geq \ \rho - \xi_i \tag{23}$$

$$\xi_i \ \geq \ 0 \tag{24}$$

$$\rho \ \geq \ 0 \tag{25}$$

where $\nu \in [0,1]$ is a parameter.
Dual[4]:

$$\text{maximize}_{\alpha} \quad -\frac{1}{2}\sum_i \alpha_i \alpha_j y^i y^j x^{iT} x^j \tag{26}$$

$$\text{s.t.} \quad \frac{1}{n} \ \geq \ \alpha_i \geq 0 \text{ for all } i \tag{27}$$

$$\sum_i \alpha_i y^i \ = \ 0 \tag{28}$$

$$\sum_i \alpha_i \ \geq \ \nu \tag{29}$$

Properties If $\rho > 0$ then:

- ν is an upper bound on #margin errors/n (if $\sum_i \alpha_i = \nu$)
- ν is a lower bound on #(original support vectors + margin errors)/n
- ν-SVM leads to the same w, b as C-SVM with $C = 1/\nu$

[4] Lagrangean $L(w, b, \xi, \rho, \alpha, \mu, \delta) \ = \ \frac{1}{2}||w||^2 - \nu\rho + \frac{1}{n}\sum_i \xi_i - \sum_i \alpha_i[y^i(w^T x^i + b) - \rho + \xi_i] - \sum_i \mu_i \xi_i - \delta\rho$ with $\alpha_i \geq 0, \ \delta \geq 0, \ \mu_i \geq 0$

A simple error bound

$$L_{01}(f_n) \;\leq\; E\left[\frac{\#\text{support vectors of } f_{n+1}}{n+1}\right] \tag{30}$$

where f_n denotes the SVM trained on a sample of size n.

Exercise Use the Homework 6 to prove this result.