STAT 535

11/16/23

Lecture 15

SVM

optimization

C - SVM

Poll · future lectures

HW5 due

HW6 TB out

Q3 ≥ TxG

Lecture V: Support Vector Machines

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

November, 2023

Linear SVM's

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

optimization

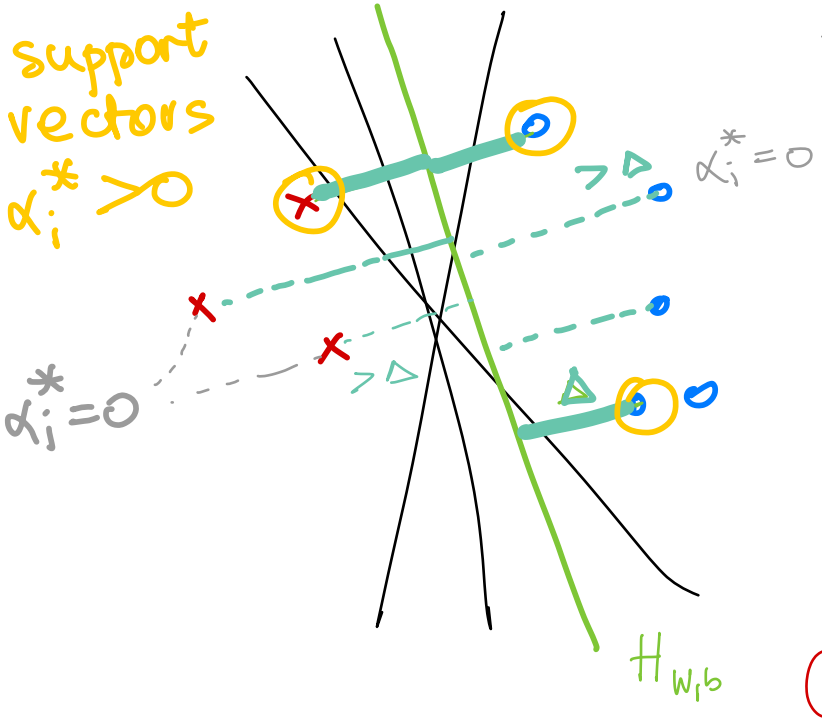
Non linear SVM

- The "kernel trick"
- Kernels
- Prediction with SVM

Extensions

- L_1 SVM
- Multi-class and One class SVM
- SV Regression

Reading HTF Ch.: Ch. 12.1–3, Murphy 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), Bach Ch.: 7.1–7.4, 7.7
Additional Reading: C. Burges - "A tutorial on SVM for pattern recognition"
These notes: Appendices (convex optimization) are optional.



Assume data linearly separable

margin

$\Delta_{w,b} = \min_i dist(x^i, H_{w,b})$

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

$H_{w,b} = \{x, f(x) = 0\}$

SVM pb max margin

want $\underset{w,b}{\text{argmax}}\ \Delta_{w,b}$ s.t.

(P1) $y^i(w^T x^i + b) > 0$

classify correctly

Remark 1.

(P2) .... s.t. $y^i(w^T x^i + b) \geq 1$

Remark 2

$d(x, H_{w,b}) = \dfrac{|w^T x + b|}{\|w\|}$

(P3) $\max_{w,b} \left(\min_i \dfrac{|w^T x^i + b|}{\|w\|} \right)$, where $\min_i |w^T x^i + b| = 1$

s.t $y^i(w^T x^i + b) \geq 1$

Remark 3. $\min_i y^i(w^T x^i + b) = 1$ $\Rightarrow$

(P4) $\max_{w,b} \frac{1}{\|w\|}$ s.t $y^i(w^T x^i + b) \geq 1$

(P5) $\min_{w,b} \|w\|^2$ s.t —"— CONVEX

(P6) $\min \|w\|^2$ s.t —"— QUADRATIC

easier to analyse !

$\Updownarrow$

(P) $\min_{w,b} \frac{1}{2}\|w\|^2$ s.t. $y^i(w^T x^i + b) \geq 1$ for all $i = 1:n$

Primal [opt. pb.]

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

$\hat{y} = \text{sgn}\, f(x)$

Convex opt. analysis

Optimization with Lagrange multipliers

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

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

[KKT conditions]

At the optimum of (13)

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

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

(P) $\min_{w,b} \frac{1}{2}\|w\|^2$ s.t. $y^i(w^T x^i + b) \geq 1$ for all $i = 1:n$ ← α_i

objective — constraints

Primal [opt. pb.]

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

$\hat{y} = \text{sgn}\, f(x)$

$1 - y^i(w^T x^i + b) \leq 0$

Convex opt. analysis

↓

Lagrangean fction

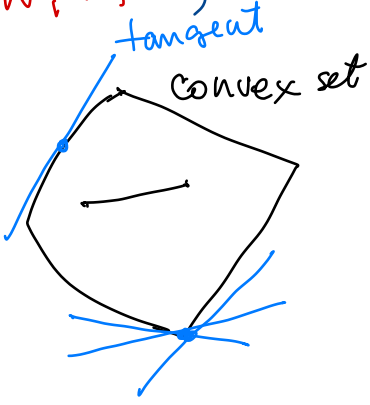
dual variable

L: $L = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i [1 - y^i(w^T x^i + b)] = L(w, b, \alpha)$

Optimum: (w^*, b^*, α^*) — min, max

KKT conditions

at opt $\begin{cases} \frac{\partial L}{\partial w} = 0 \\ \frac{\partial L}{\partial b} = 0 \end{cases}$



at opt $\frac{\partial L}{\partial w} = 0$

$w + \sum_i (-y^i \alpha_i x^i)$

$\Rightarrow \boxed{w^* = \sum_{i=1}^n \alpha_i^* y^i x^i}$ $\Rightarrow$ w^* = linear combination of x^i !!

$\alpha_i^* \geq 0$ (no proof)

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

$\|w\|^2 = w^T w$

$-w^T \sum \alpha_i y^i x^i = -w^T w$

$\frac{\partial L}{\partial b} = -\boxed{\sum \alpha_i^* y^i = 0}$ constraint on α_i's

$L(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y^i y^j \boxed{(x^i)^T x^j} + 0$

$= 1^T \alpha - \frac{1}{2} \alpha^T \bar{G} \alpha = g(\alpha)$ concave, quadratic

$\alpha = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix}$

$G = \left[(x^i)^T x^j\right]_{i,j=1:n}$ Gram matrix

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

dual pb

(D) $\max_{\alpha \in \mathbb{R}^n} g(\alpha)$ s.t. $\alpha_i \geq 0$

$\sum_{i=1}^n \alpha_i y^i = 0$!! linear

$\bar{G} = \left[y^i y^j (x^i)^T (x^j)\right]_{i,j}$

$= \begin{bmatrix} y^1 & \cdots \end{bmatrix} G \begin{bmatrix} y_1 \\ \vdots \end{bmatrix} \geq 0$

solution α^*

(P) Primal [opt. pb.]

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

objective

$= 1$

(D)

$$\max_{\alpha \in \mathbb{R}^n} g(\alpha) \quad \text{s.t.} \quad \alpha_i \geq 0, \quad \sum_{i=1}^{n} \alpha_i y^i = 0$$

linear

solution α^*

$$w^* = \sum_{\substack{i=1 \\ \alpha_i^* > 0}}^{n} \alpha_i^* y^i x^i$$

STATISTICS, finally

~ depends on subset of (x^i, y^i) directly → complexity

depends on data

params = # s.v.

⇒ low variance

b^*

for s.v.: $y^i(w^T x^i + b) = 1$

↳ solve for b^*

$w^T x^i + b = y^i$

Theory

$\alpha_i^* > 0$ tight

$\alpha_i^* = 0$ slack constraints

support vectors (y^i, x^i)

= x^i's on the margin

SVM Training

1. $\{(x^i, y^i)\}_i \rightarrow$ compute $G, \bar{G}$
2. Solve (D) dual pb $\Rightarrow \alpha_i^*$, $i = 1:n$
3. $w = \sum_{\substack{i=1 \\ \alpha_i^* > 0}}^{n} \alpha_i^* y^i x^i$
4. for i support vector : solve for b
   $\alpha_i^* > 0$

Prediction

new x : $f(x) = w^T x + b = \sum_{i, \alpha_i^* > 0} \alpha_i^* y^i (x^i)^T x + b$

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

- Sofar : linear f, data linearly separable ✓
- Next : —"— , data NOT —"— —"—
- —"— : NonLinear f

Computation efficient for $n \ll d$

(D) $\alpha \in \mathbb{R}^n$

n variables

$(n+1)$ constraints

(P) $d+1$ variables

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

$b \in \mathbb{R}$

n constraints

efficient for d small n large

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 $\bar{G} = \text{diag}\{y^{1:n}\}^T G \text{diag}\{y^{1:n}\}$.

- 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

The **C-SVM**

C ↑ less slack ⇒ "∧" ⇒ bias ↘ var ↗

C>0 regularization — as little slack as possible

$$\begin{aligned} \text{minimize}_{w,b,\xi} \quad & \frac{1}{2}||w||^2 + C\sum_i \xi_i \\ \text{s.t.} \quad & y^i(w^T x^i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0 \end{aligned} \tag{20}$$

← slack variables

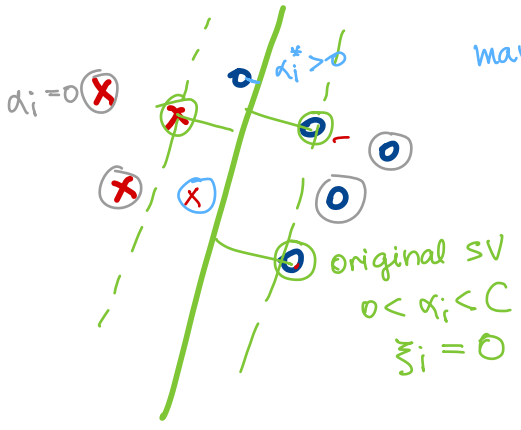
In the above, ξ_i are the slack variables. Dual[3]:

$$\begin{aligned} \text{maximize}_{\alpha} \quad & \sum_i \alpha_i - \frac{1}{2}\sum_i \alpha_i \alpha_j y^i y_j x^{i\,T} x_j \\ \text{s.t.} \quad & C \geq \alpha_i \geq 0 \text{ for all } i \\ & \sum_i \alpha_i y^i = 0 \end{aligned} \tag{21}$$

⇒ two types of SV

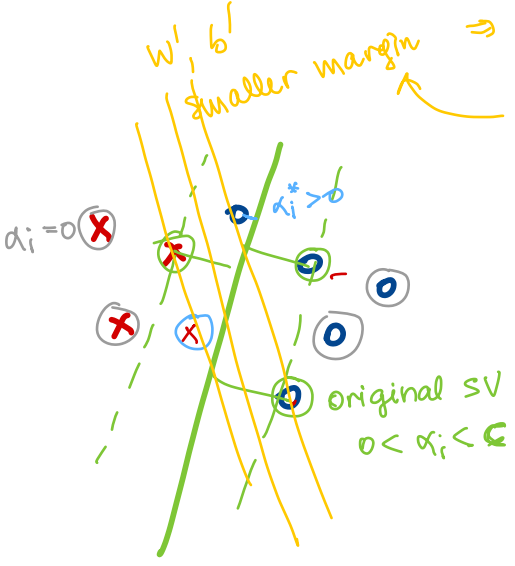
- $\alpha_i < C$ data point x^i is "on the margin" $\Leftrightarrow$ $y^i(w^T x^i + b) = 1$ (original SV)
- $\alpha_i = C$ data point x^i cannot be classified with margin 1 (**margin error**) $\Leftrightarrow$ $y^i(w^T x^i + b) < 1$

[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 $\alpha_i \geq 0,\ \xi_i \geq 0,\ \mu_i \geq 0$



margin errors $\xi_i > 0 \Leftrightarrow \alpha_i^* = C$

In general:
$\alpha_i^* \uparrow \Leftrightarrow$ point i harder to classify



Lectures 1.5-2 Boosting SVM

—"— [Kernel machines] → R Fourier F.* → Double* Descent

< 1 ↳ Neuro-tangent* + ...

1.5 — Clust K-means + EM

1.5 Spectral clustering

< 1 Model selection AIC, BIC, Struct Risk Min ← VC dim

Kernel trick

RKHS $\mathcal{H}$

4.5 lectures

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]:

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

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

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

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

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 $\alpha_i \geq 0,\ \delta \geq 0,\ \mu_i \geq 0$