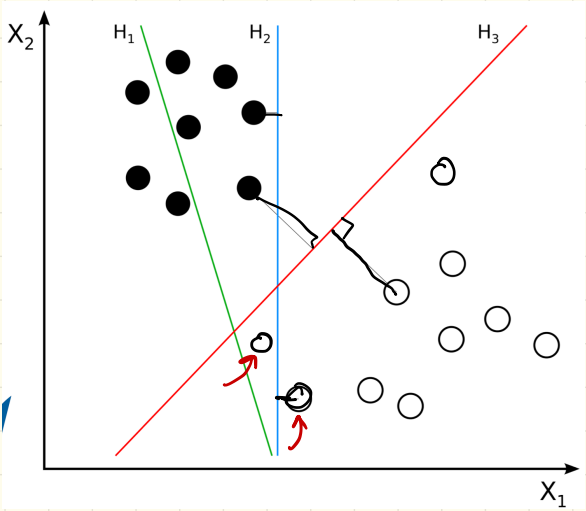
Topic 8

Support vector machines (SVM)

Plan:

1. Motivation for SVM
2. How to solve SVM optimisation problem?
3. Duality (a proof of the method of solving SVM)

Section 1: Motivation and derivation.



Binary classific.

Previously linear regression →

→ logistic func

→ classification

H_1 – wrong

H_2 – ok

H_3 – ok

Idea: maximise the total/average distance between points and a decision boundary.

1.2 Simple linear algebra/geometry

Assumptions: binary classification (based on linear regression), class labels $\{-1, 1\}$.

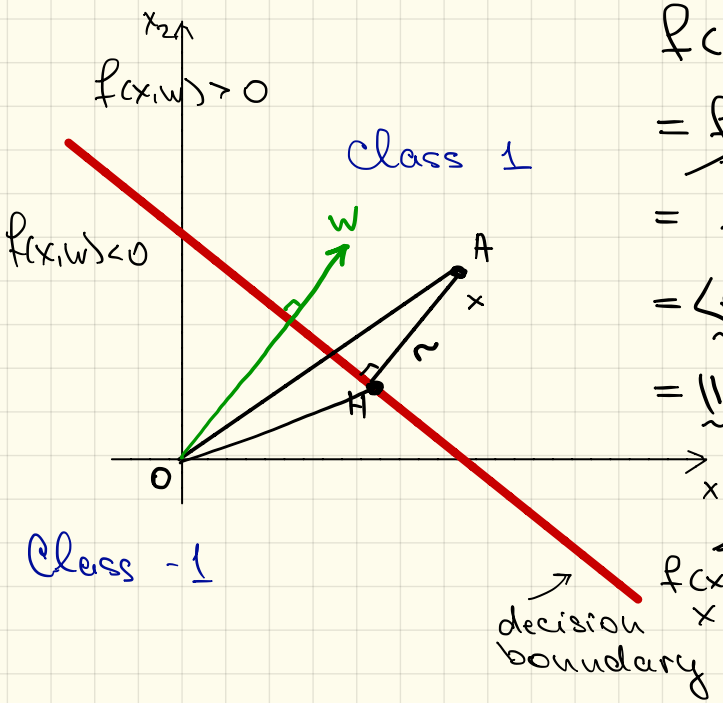
Start with $\{x^{(i)}, y_i\}_{i=1}^{s}$, $x^{(i)} \in \mathbb{R}^d$, $y_i \in \{-1, 1\}$.

$$f(x, w) = \langle \tilde{x}, \tilde{w} \rangle = w_0 + \langle x, w \rangle$$

$\tilde{x}$: extended, $\tilde{w} \in \mathbb{R}^{d+1}$

if $f(x, w) \geq 0 \Rightarrow$ label x with 1

if $f(x, w) < 0 \Rightarrow$ label x with -1



$$f(x,w) = f(x^{\perp} + r, w)$$

$$= \underbrace{f(x^{\perp}, w)}_{0} + f(r,w)$$

$$= f(r,w)$$

$$= \langle r, w \rangle + w_0$$

$$= \|r\| \cdot \|w\| + w_0$$

$$\langle x, w \rangle + w_0 = 0$$

$f(x,w) = 0$ for x on a boundary

Goal express r in terms of $f(x,w)$.

$AH \parallel \vec{w}$

H is a projection of A on a line

$\Rightarrow \vec{OH} = x^{\perp}$

Any x: $x = x^{\perp} + r$

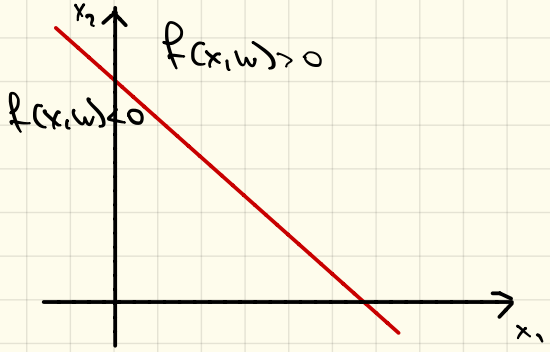
$x^{\perp} \in$ line; $r \perp$ to the line.

$$\|r\| = \frac{|f(x,w) - w_0|}{\|w\|}$$

↳ max

$$\Rightarrow \boxed{\max \frac{f(x,w)}{\|w\|}}$$

What about classification?



$\{x^{(i)}, y_i\}_{i=1}^{s}$

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

We want to have $f(x^{(i)}, w) > 0$ if $y_i = 1$ and $f(x^{(i)}, w) < 0$ if $y_i = -1$. $\Leftrightarrow$ $y_i \cdot f(x^{(i)}, w) \geq 0$

"$\Leftrightarrow$" $y_i \cdot f(x^{(i)}, w) \geq 1$.

Conclusion: $\max \sum_{i=1}^{s} \frac{|f(x^{(i)}, w) \cdot y_i|}{\|w\|}$

with $y_i f(x^{(i)}, w) \geq 1$ for $\forall i$.

Reformulate: $\min \|w\|$ with the same constraints

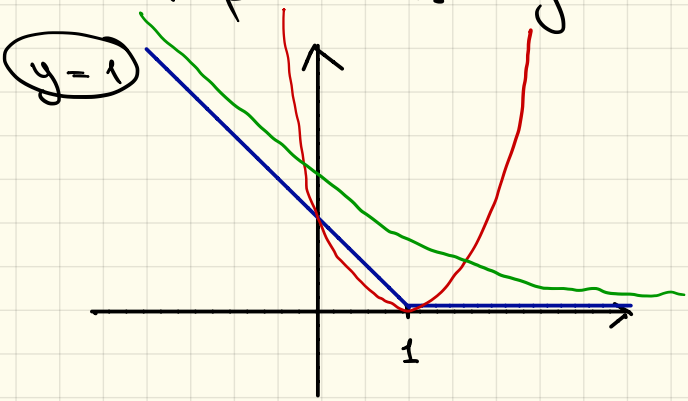
$$\min_w \left\{ \sum_{i=1}^{s} \max\{0, 1 - y_i f(x^{(i)}, w)\} + \frac{\alpha}{2} \|w\|^2 \right\}$$

Support vector machine optimisation problem.

$$f(x^{(i)}, w) = \langle \tilde{x}^{(i)}, \tilde{w} \rangle = w_0 + \langle x^{(i)}, w \rangle$$

$H(z) = \max(0, 1 - yz)$ - Hinge-loss

In our problems $y = \pm 1$



- $MSE(z) = (1 - yz)^2$
- $\text{Logistic}(z) = \log(1 + e^{-yz})$

2. Solving SVM optimisation

$$\min_w \left\{ \sum_{i=1}^{s} \max(0, 1 - y_i f(x^{(i)}, w)) + \frac{\alpha}{2}\|w\|^2 \right\} = L(w)$$

Problems: ① $\max(0, 1-yz)$ is non-differentiable.

~~②~~ rewrite everything in terms of X, y, w

$$L(w) = \sum_{i=1}^{s} \max\left(0, \mathbb{1}_i - \underbrace{(YXw)_i}_{\text{vector}}\right) + \frac{\alpha}{2}\|w\|^2$$

$$\mathbb{1} = \underbrace{(1, 1, \ldots, 1)}_{s}^{T} \qquad Y = \begin{pmatrix} y_1 & & 0 \\ & \ddots & \\ 0 & & y_s \end{pmatrix}$$

$$\max(0, z) = \max_{\lambda \in [0,1]} \lambda z \cdot w$$

$$L(w) = \sum_{i=1}^{s} \max_{\lambda_i \in [0,1]} \left\{ \lambda_i \cdot (\mathbb{1}_i - (YXw)_i) \right\} + \frac{\alpha}{2}\|w\|^2$$

$$L(w) = \sum_{i=1}^{s} \max_{\lambda_i \in [0,1]} \left\{ \lambda_i \cdot \left(\mathbb{1}_i - \left(Y X w \right)_i \right) \right\} + \frac{\alpha}{2} \|w\|^2$$

$$= \boxed{\max_{\Lambda} \sum_{i=1}^{s} \left(\Lambda \left(\mathbb{1} - Y X w \right) \right)_i + \frac{\alpha}{2} \|w\|^2}$$

$$\Lambda = \begin{pmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ 0 & & \ddots \\ & & & \lambda_s \end{pmatrix}$$

SVM problem:

$$\min_{w} \max_{\Lambda} \sum_{i=1}^{s} \left(\Lambda \left(\mathbb{1} - Y X w \right) \right)_i + \frac{\alpha}{2} \|w\|^2$$

$$\hat{w} = \arg\min_{w} \max_{\Lambda} \sum_{i=1}^{s} \left(\Lambda \left(\mathbb{1} - Y X w \right) \right)_i + \frac{\alpha}{2} \|w\|^2$$

Crucial idea: change min $\leftrightarrow$ max order !

Example: $f(x,y) = \sin(x+y)$

$$\min_x \max_y f(x,y) = \min_x 1 = 1$$

$$\max_y \min_x f(x,y) = \max_y -1 = -1$$

we can't change min-max order here.

Assume we can change min-max in SVM optimisation problem.

$$\max_{\Lambda} \min_{w} \underbrace{\sum_{i=1}^{n} \Lambda_i (1 - Y_i X_i w) + \frac{\alpha}{2}\|w\|^2}_{L(w,\Lambda)}$$

Observation:

- $L(w,\Lambda)$ is differentiable in w and Λ.
- $L(w,\Lambda)$ is convex in w, is concave in Λ.

$L(w, \Lambda)$ is minimised in w's when $\nabla_w L(w, \Lambda) = 0$.

$$L(w, \Lambda) = \sum_{i=1}^{s} \lambda_i \left(1 - y_i \langle x^{(i)}, w \rangle\right) + \frac{\alpha}{2} \|w\|^2$$

$$\nabla_w L(w, \Lambda) = \sum_{i=1}^{s} -\lambda_i y_i x^{(i)} + \alpha w = 0$$

$$\boxed{w = \frac{1}{\alpha} \sum_{i=1}^{s} \lambda_i y_i x^{(i)} = \frac{1}{\alpha} X^T Y \lambda}$$

$$\max_{\lambda \in [0,1]^s} L\left(\frac{1}{\alpha} X^T Y \lambda, \Lambda\right)$$

$$= \max_{\Lambda} \left\{ \sum_{i=1}^{s} \left(\Lambda \left(\mathbb{1} - \frac{1}{\alpha} Y X X^T Y \lambda\right)\right)_i + \frac{\alpha}{2} \frac{1}{\alpha^2} \|X^T Y \lambda\|^2 \right\}$$

$$= \max_{\Lambda} \left\{ \langle \lambda, \mathbb{1} \rangle - \frac{1}{\alpha} \sum_{i=1}^{s} \left(\Lambda Y^T X X^T Y \lambda\right)_i + \frac{1}{2\alpha} \|X^T Y \lambda\|^2 \right\}$$

$\lambda = (\lambda_1, \ldots, \lambda_s)^T$

$$= \max_{\lambda \in [0,\infty)^s} \left\{ \langle \lambda, \mathbb{1} \rangle - \frac{1}{2} \sum_{i=1}^{s} (\Lambda Y^T X X^T Y \lambda)_i \right.$$
$$\left. + \frac{1}{2\alpha} \| X^T Y \lambda \|^2 \right\}$$

$\lambda = (\lambda_1, \ldots, \lambda_s)^T$

$$= \max_{\lambda \in [0,\infty)^s} \left\{ \langle \lambda, \mathbb{1} \rangle - \frac{1}{2} \| X^T Y \lambda \|^2 \right.$$
$$\left. + \frac{1}{2\alpha} \| X^T Y \lambda \|^2 \right\}$$

$$= \max_{\lambda \in [0,\infty)^s} \left\{ \langle \lambda, \mathbb{1} \rangle - \frac{1}{2\alpha} \| X^T Y \lambda \|^2 \right\}$$

$L(\lambda)$

$L(\lambda)$ is differentiable, concave

$$\nabla_\lambda L(\lambda) = \mathbb{1} - \frac{1}{2\alpha} \nabla_\lambda \langle Y X X^T Y \lambda, \lambda \rangle$$

$$= \mathbb{1} - \frac{1}{\alpha} Y X X^T Y \lambda .$$

Bad idea: $\lambda = \alpha (Y X X^T Y)^{-1} \mathbb{1}$.

Restrictions !

Initial $\max_{\lambda \in [0,1]^3}$ problem can be written as

$$\max_{\lambda \in \mathbb{R}^3} \{L(\lambda) - \chi_{[0,1]^3}(\lambda)\}$$

Proximal gradient acsent (max)

We can find λ's by iterative procedure:

$$\lambda^{(k+1)} = \text{proj}_{[0,1]^3}\left(\lambda^{(k)} + \tau\left(\mathbb{1} - \frac{1}{2} Y X X^T Y \lambda^{(k)}\right)\right)$$

Algorithm:

- Take $\lambda = (0, \dots, 0)$ and repeat the above steps until "convergence".
- Calculate $\hat{w} = \frac{1}{2} X^T Y \lambda$.

Recap: SVM problem

$$\hat{w} = \arg\min_{w} \left\{ \sum_{i=1}^{s} \max(0, 1 - y_i \langle x^{(i)}, w \rangle) + \frac{\alpha}{2} \|w\|^2 \right\}$$

$$\max(0, z) = \max_{\lambda \in [0,1]} \lambda z$$

$$L(w, \lambda) = \sum_{i=1}^{s} \lambda_i (1 - y_i \langle x^{(i)}, w \rangle) + \frac{\alpha}{2} \|w\|^2$$

$$\hat{w} = \arg\min_{w} \max_{\lambda} L(w, \lambda) \; ?$$

$$\Leftrightarrow \max_{\lambda} \min_{w} L(w, \lambda) \; .$$

$$\hat{w} = \frac{1}{\alpha} X^T Y \lambda$$

$$\hat{\lambda} \longrightarrow \hat{w}$$

Duality

Question: Let $f(x,y)$ be a real valued function. What could be proven about $\min_x \max_y f(x,y)$ and $\max_y \min_x f(x,y)$?

Statement 1: (Max-Min inequality)

$$\max_y \min_x f(x,y) \leq \min_x \max_y f(x,y)$$

For every y we look for $x(y)$ such that

$f(x(y),y) \leq f(x,y)$ for any x.

↓ max over y

we find y^* such that

$f(x(y^*), y^*) \geq f(x(y), y)$

$f(x, y(x)) \geq f(x,y)$ $\forall y$.

Our goal to have

$$\max_y \min_x f(x,y) = \min_x \max_y f(x,y)$$

Theorem (von Neumann, 1928)
Let $X \subset \mathbb{R}^m$, $Y \subset \mathbb{R}^n$ be compact convex sets. If $f: X \times Y \to \mathbb{R}$ is continuous and convex-concave, that means

$f(\cdot, y): X \to \mathbb{R}$ - convex
$f(x, \cdot): Y \to \mathbb{R}$ - concave

then

$$\min_{x \in X} \max_{y \in Y} f(x,y) = \max_{y \in Y} \min_{x \in X} f(x,y)$$

Topic 9: Interpolation with graphs.

Another method of solving classification problems.

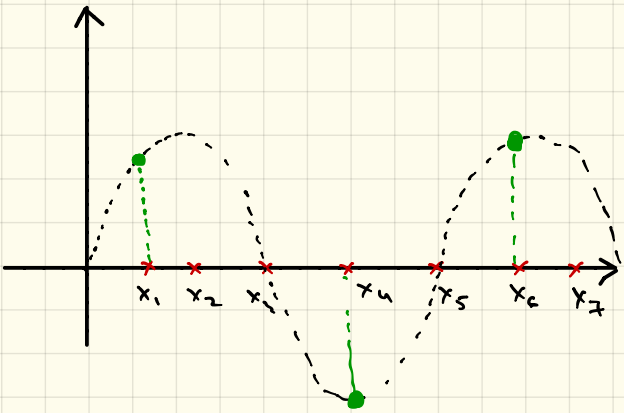
Goal: Having a list of data inputs with partial list of outputs classify the rest of inputs based on their similarities

Problem: $\{x_i\}_{i \in I_1} \subset X$

↑ not even numbers / vectors

for $i \in I_2 \subset I_1$ we are given class labels y_i. To find y_i for $i \in I_1$

Example:



$I_1 = \{1, 2, \dots, 7\}$

$I_2 = \{1, 4, 6\}$

+ additional information about x_i's.

$\Downarrow$

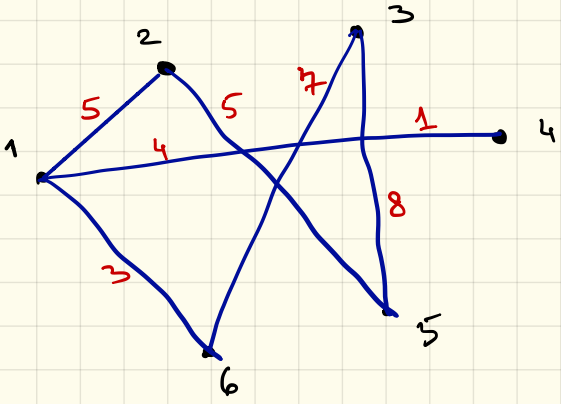
$f(x)$ for $x \in I_1$

What is this additional information?

In this case this will be a similarity between points.

9.1 Graphs as a tool for similarity representation.

Graphs = nodes and edges.
We will consider undirected weighted graphs.



- edges don't have directions
- every edge (u,v) is entitled with the weight w_{uv}.

Def The incidence matrix of a graph $G = (V, E)$ (V: nodes, E: edges) is a matrix of a size $|E| \times |V|$ of the form

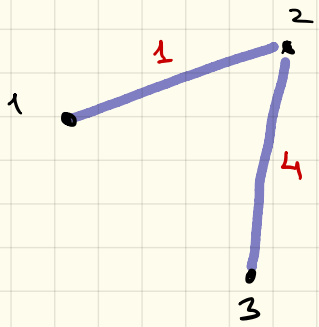
$$M_{ev} = \begin{cases} -\sqrt{w_{ij}} & , e = (i,j) , v = i \\ \sqrt{w_{ij}} & , e = (i,j) , v = j \\ 0 & , e = (i,j) , v \neq i,j \end{cases}$$

(e: edge, v: node)

The Laplacian matrix of G is given by

$$L = M^T M$$

Example:



$$V = \begin{pmatrix} v_1 & v_2 & v_3 \\ 1, & 2, & 3 \end{pmatrix}$$

$$E = \begin{pmatrix} (1,2) & ; & (2,3) \\ e_1 & & e_2 \end{pmatrix}$$

$$M = \begin{array}{c} \\ e_1 \\ e_2 \end{array} \begin{array}{c} \begin{matrix} v_1 & v_2 & v_3 \end{matrix} \\ \begin{pmatrix} -1 & 1 & 0 \\ 0 & -2 & 2 \end{pmatrix} \end{array}$$

$$L = M^T M = \begin{pmatrix} -1 & 0 \\ 1 & -2 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} -1 & 1 & 0 \\ 0 & -2 & 2 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & -1 & 0 \\ -1 & 5 & -4 \\ 0 & -4 & 4 \end{pmatrix}$$

$-w_{uv}$

↑ total weight

$$L \begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix} \approx \begin{pmatrix} \tilde{f}'(x_1) \\ \vdots \\ \tilde{f}'(x_n) \end{pmatrix}$$

↑ weighted derivative