IE 506 - Machine Learning: Principles and Techniques

Support Vector Machines for Classification

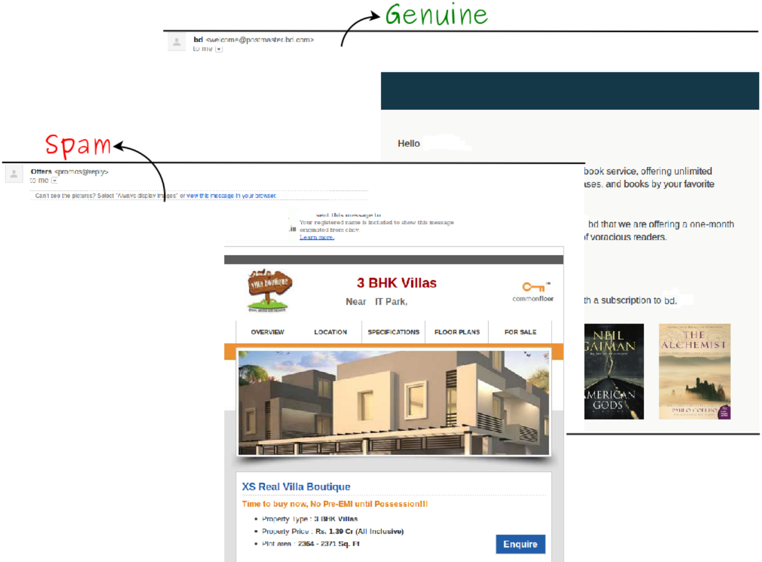
February 16, 2024.

1. Recap
   - Supervised Machine Learning
2. Support Vector Machines

Recap: Supervised Machine Learning

Binary Classification

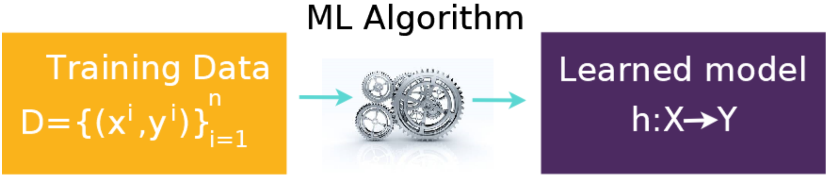
Recall: e-mail Spam Classification



Binary Classification

- **Input:** e-mail messages $\implies$ some feature space $\subseteq \mathbb{R}^d$.
  - $x \in \mathcal{X} \subseteq \mathbb{R}^d$
- **Output:** Spam/Not spam $\implies$ $\{+1, -1\}$
  - $y \in \mathcal{Y} = \{+1, -1\}$
- Generally n input/output pairs $\{(x^i, y^i)\}_{i=1}^n \subset (\mathcal{X} \times \mathcal{Y})$ are given for learning the machine learning model.
- $D = \{(x^i, y^i)\}_{i=1}^n$ called the training data.

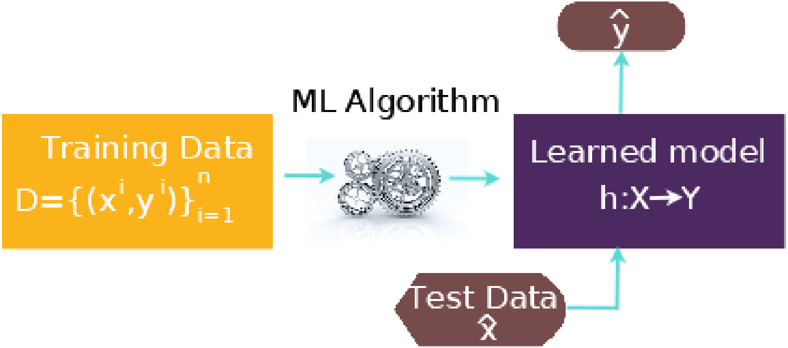
General Nature of a Supervised Machine Learning Task



Training

- **Input:** Training data $D = \{(x^i, y^i)\}_{i=1}^{n}$
- **Aim:** Learn a model $h : \mathcal{X} \to \mathcal{Y}$

General Nature of a Supervised Machine Learning Task



Training

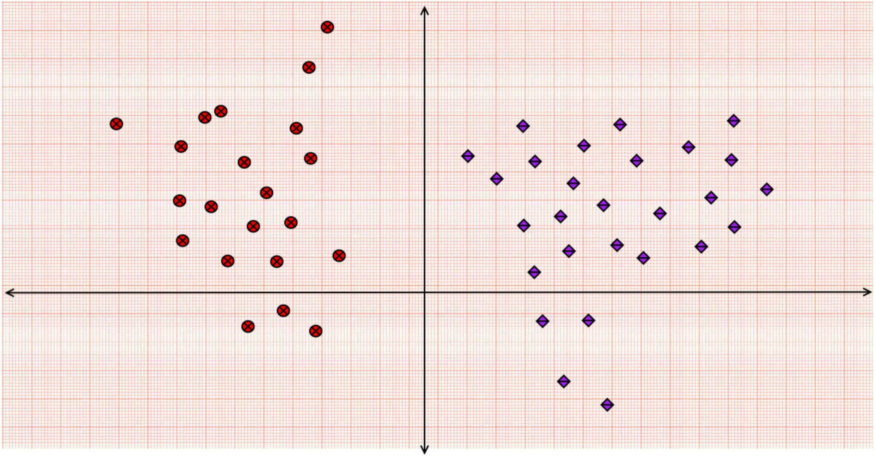
- **Input:** Training data $D = \{(x^i, y^i)\}_{i=1}^n$
- **Aim:** Learn a model $h : \mathcal{X} \to \mathcal{Y}$

Testing

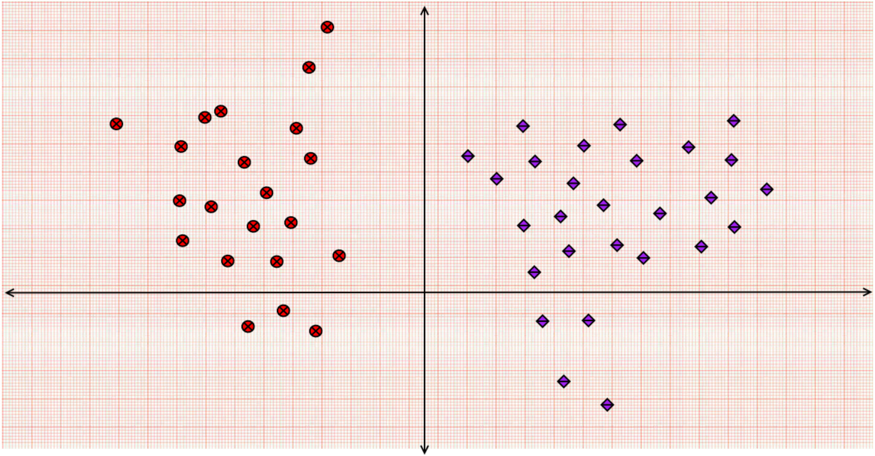
- Given $\hat{x}$, predict $\hat{y} = h(\hat{x})$

Support Vector Machines: Geometric intuition

SVMs: Geometric idea

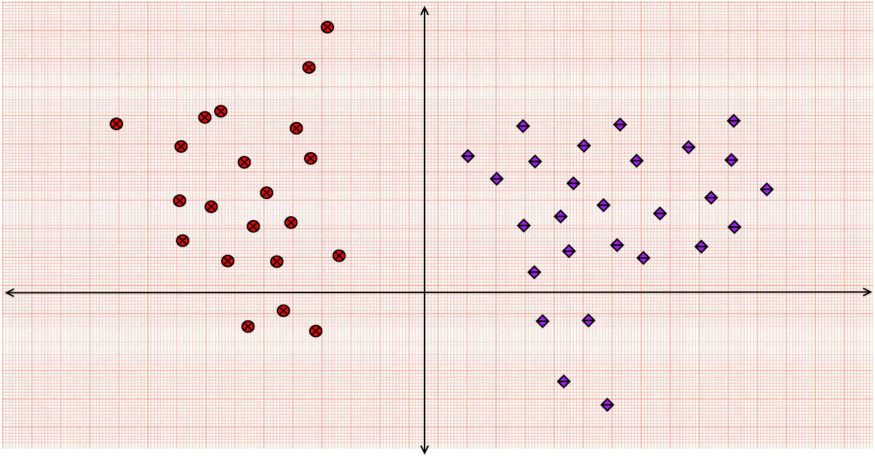


SVMs: Geometric idea



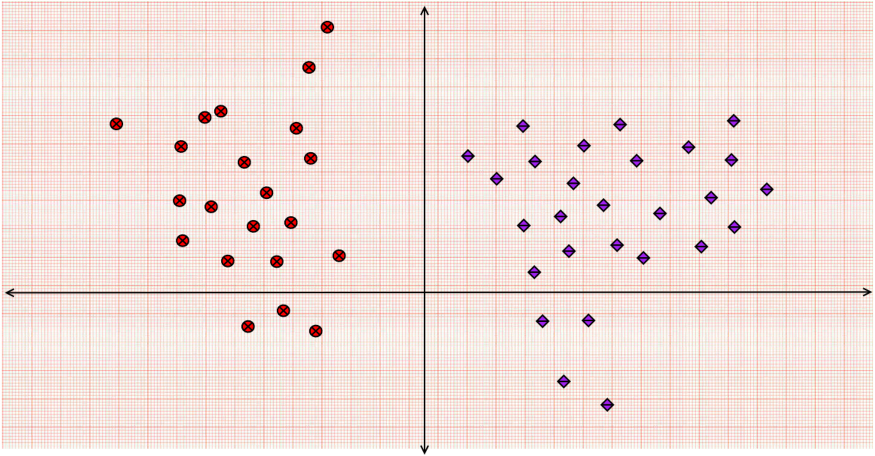
- Training data $D = \{(x^i, y^i)\}_{i=1}^n$

SVMs: Geometric idea



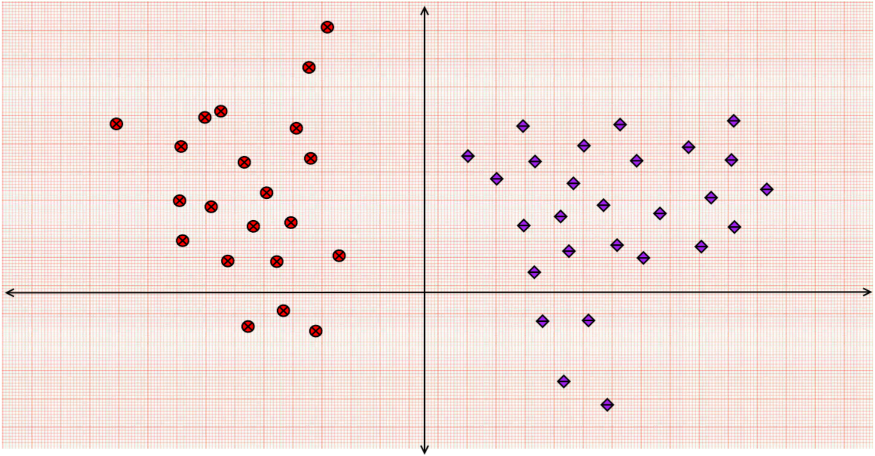
- Training data $D = \{(x^i, y^i)\}_{i=1}^n$
- Data features $x^i = (x_1^i, x_2^i, \ldots, x_d^i) \in \mathbb{R}^d, \ \forall i \in \{1, 2, \ldots, n\}$ are represented in a high dimensional space. (Here $d = 2$ is assumed.)

SVMs: Geometric idea



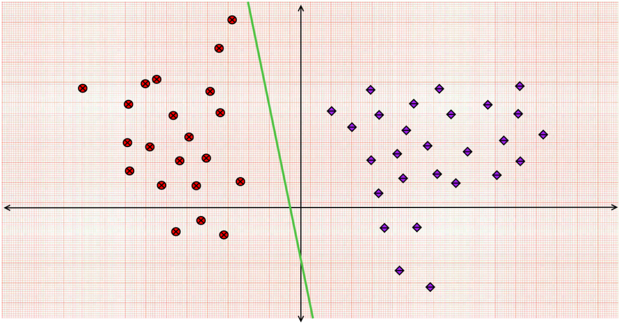
- Training data $D = \{(x^i, y^i)\}_{i=1}^n$
- Data features $x^i = (x_1^i, x_2^i, \ldots, x_d^i) \in \mathbb{R}^d$, $\forall i \in \{1, 2, \ldots, n\}$ are represented in a high dimensional space. (Here $d = 2$ is assumed.)
- Data labels are represented using different colors.

SVMs: Geometric idea



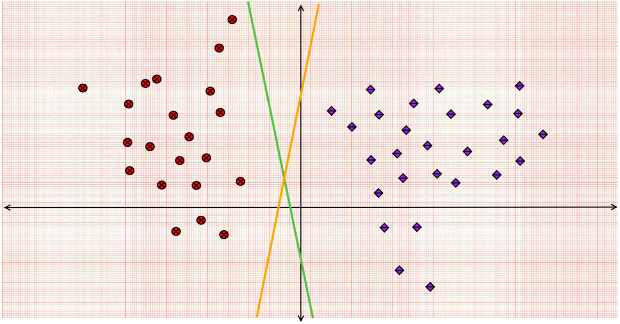
- **Question:** What is the geometric idea of separating samples belonging to one class from another?

SVMs: Geometric idea



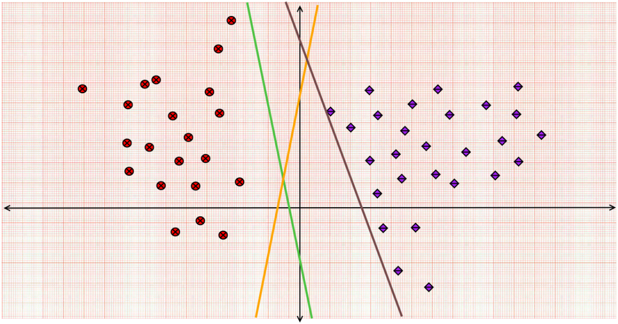
- **Question:** What is the geometric idea of separating samples (or points) belonging to one class from another?
- **One possible solution:** Find a line (or hyperplane) which can separate the points of one class from another.

SVMs: Geometric idea



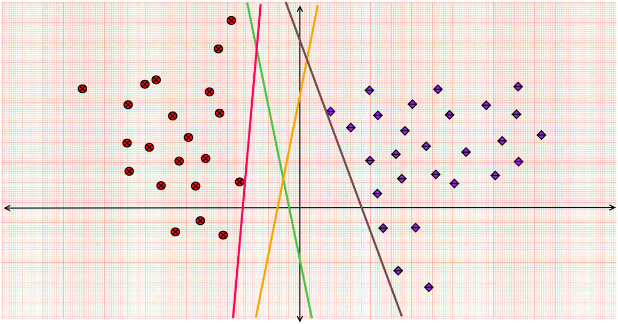
- **Note:** There may be multiple such separating lines (or hyperplanes).

SVMs: Geometric idea



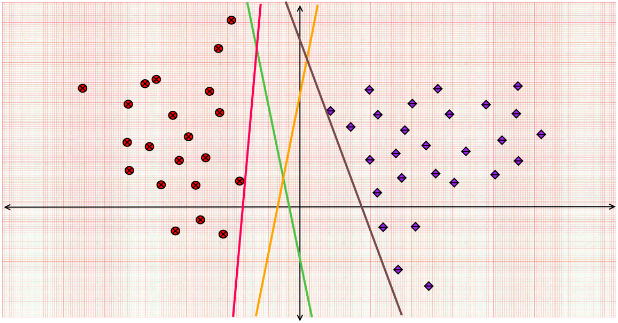
- **Note:** There may be multiple such separating lines (or hyperplanes).

SVMs: Geometric idea



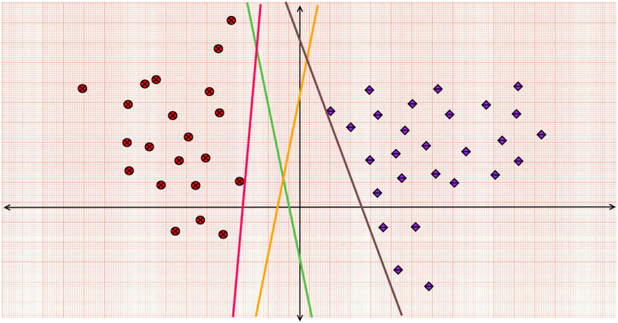
- **Note:** There may be multiple such separating lines (or hyperplanes).

SVMs: Geometric idea



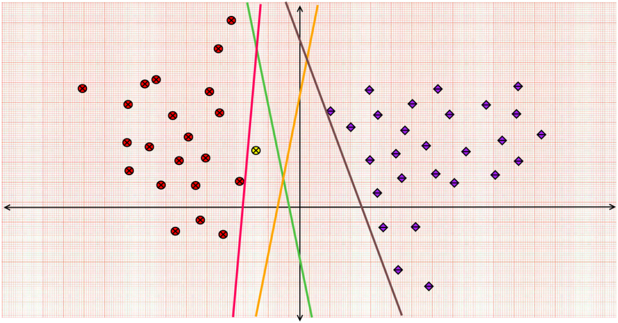
- **Note:** There may be multiple such separating lines (or hyperplanes).
- **Question:** How to choose one of these separating hyperplanes?

SVMs: Geometric idea



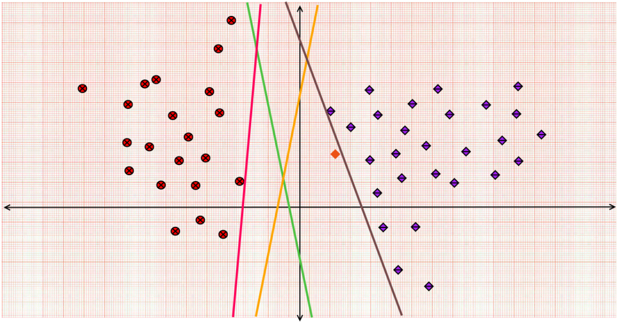
- **Note:** There may be multiple such separating lines (or hyperplanes).
- **Question:** How to choose one of these separating hyperplanes?
- **Related question:** Can we choose hyperplanes which are close to the samples?

SVMs: Geometric idea



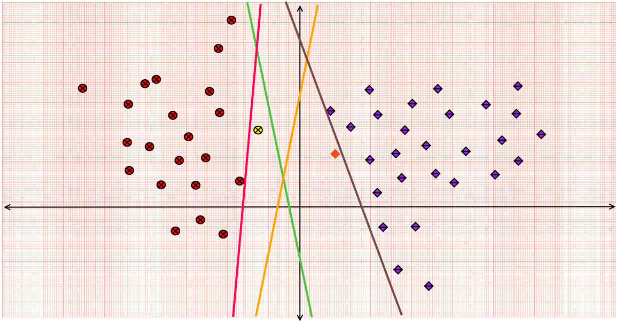
- **Note:** There may be multiple such separating lines (or hyperplanes).
- **Question:** How to choose one of these separating hyperplanes?
- **Related question:** Can we choose hyperplanes which are close to the samples?

SVMs: Geometric idea



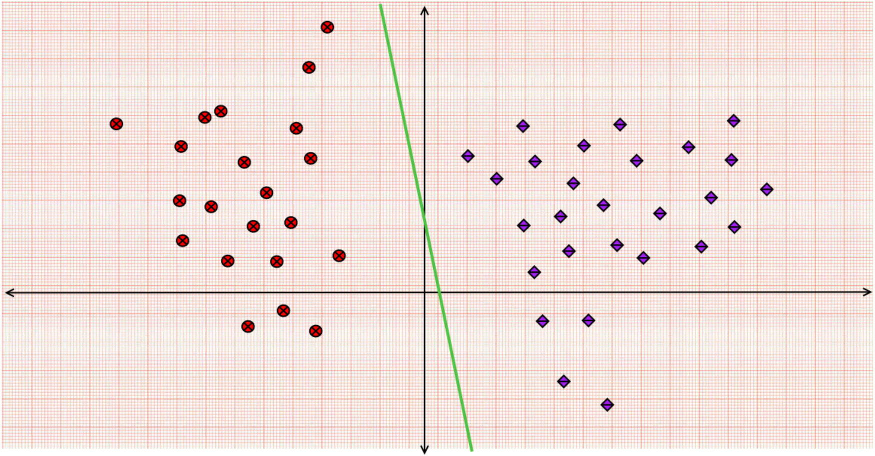
- **Note:** There may be multiple such separating lines (or hyperplanes).
- **Question:** How to choose one of these separating hyperplanes?
- **Related question:** Can we choose hyperplanes which are close to the samples?

SVMs: Geometric idea



- **Observation:** Choosing hyperplanes very close to the samples might lead to misclassification of new points even when they are close to the points with a particular label.

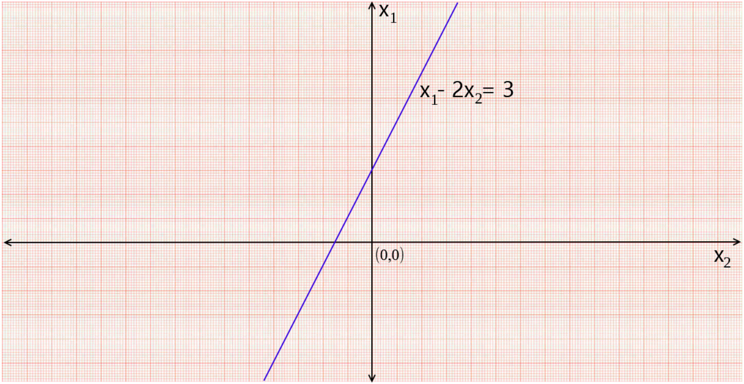
SVMs: Geometric idea



- **Motivation:** Hence we consider hyperplanes whose orientation is in the middle.

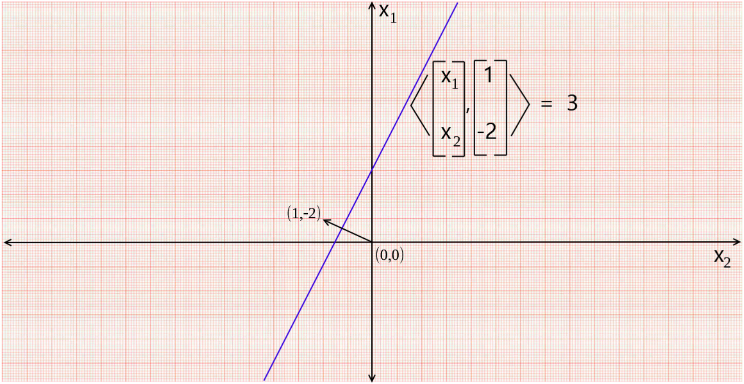
SVMs: Geometric idea

Representation of a hyperplane in 2 dimensions:



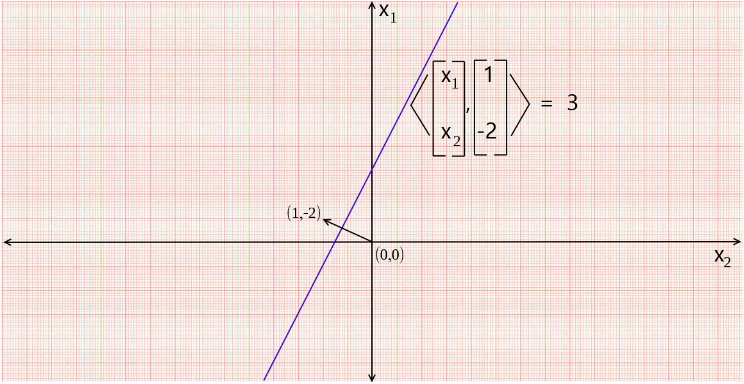
SVMs: Geometric idea

Representation of a hyperplane in 2 dimensions:



SVMs: Geometric idea

Representation of a hyperplane in 2 dimensions:

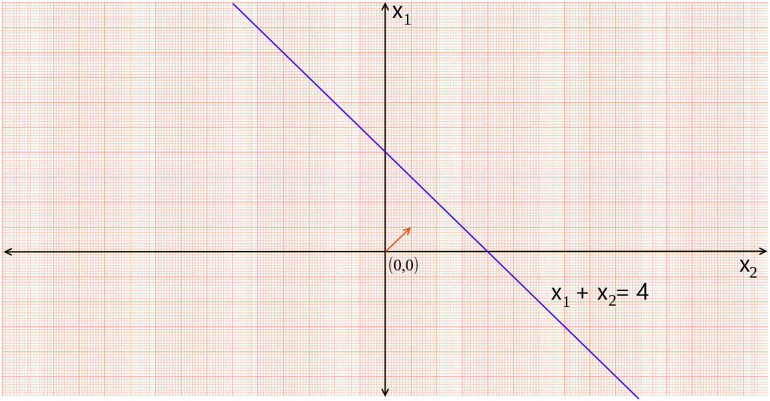


Hyperplane in $\mathbb{R}^d$

$\mathcal{H} = \{x \in \mathbb{R}^d : \exists w \neq 0 \text{ and } b \in \mathbb{R} \text{ s.t. } \langle w, x \rangle = b\}$.

SVMs: Geometric idea

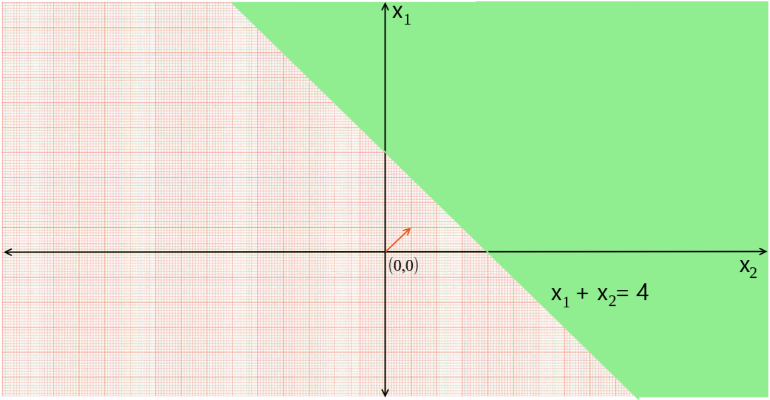
Another example of a hyperplane in 2 dimensions:



Hyperplane in $\mathbb{R}^d$

$\mathcal{H} = \{x \in \mathbb{R}^d : \exists w \neq 0$ and $b \in \mathbb{R}$ s.t. $\langle w, x \rangle = b\}$.

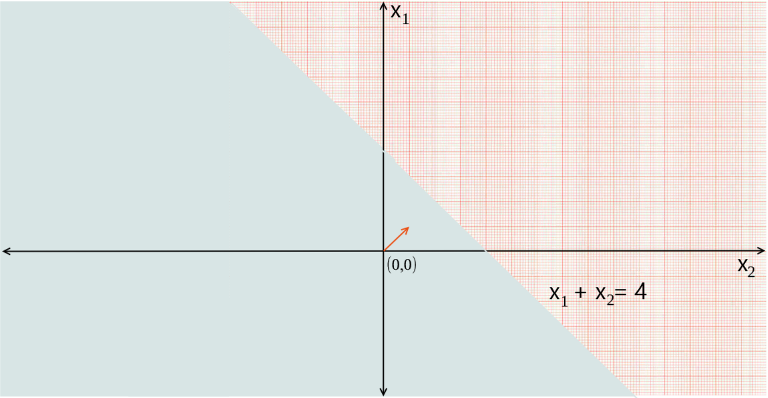
SVMs: Geometric idea



Closed Halfspaces associated with a hyperplane $\langle w, x\rangle = b$

$\mathcal{S}_1 = \{x \in \mathbb{R}^d : \langle w, x\rangle \geq b\}$ and $\mathcal{S}_2 = \{x \in \mathbb{R}^d : \langle w, x\rangle \leq b\}$.

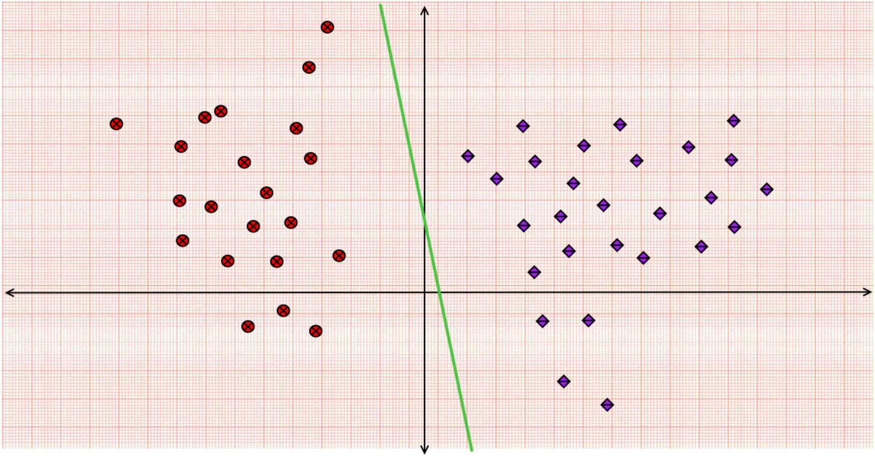
SVMs: Geometric idea



Closed Halfspaces associated with a hyperplane $\langle w, x \rangle = b$

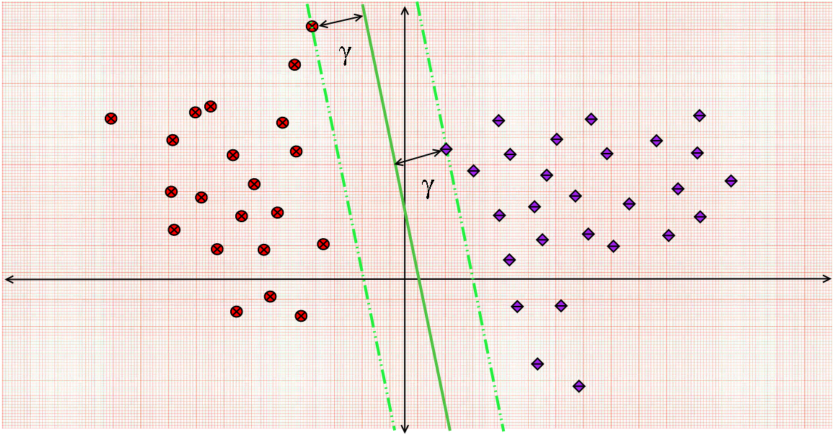
$\mathcal{S}_1 = \{x \in \mathbb{R}^d : \langle w, x \rangle \geq b\}$ and $\mathcal{S}_2 = \{x \in \mathbb{R}^d : \langle w, x \rangle \leq b\}$.

SVMs: Geometric idea



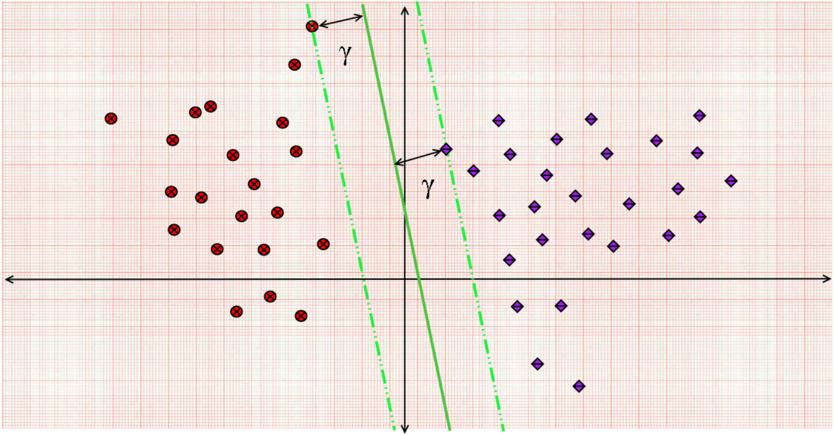
- **Recall our motivation:** We consider hyperplanes whose orientation is in the middle.

SVMs: Geometric idea



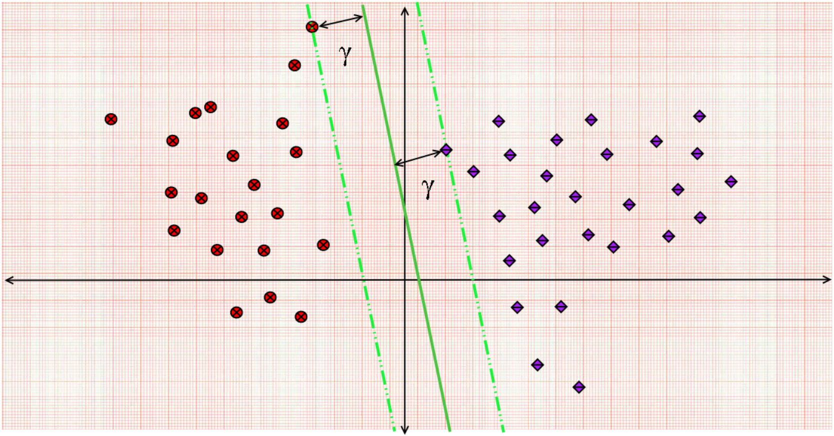
- **Motivation:** In particular, we consider the hyperplane whose distance from the closest points from both classes is maximal.

SVMs: Geometric idea



- **Motivation:** In particular, we consider the hyperplane whose distance from the closest points from both classes is maximal.
- This distance of the separating hyperplane from the closest points of either of the two classes is called **margin** (denoted by γ).

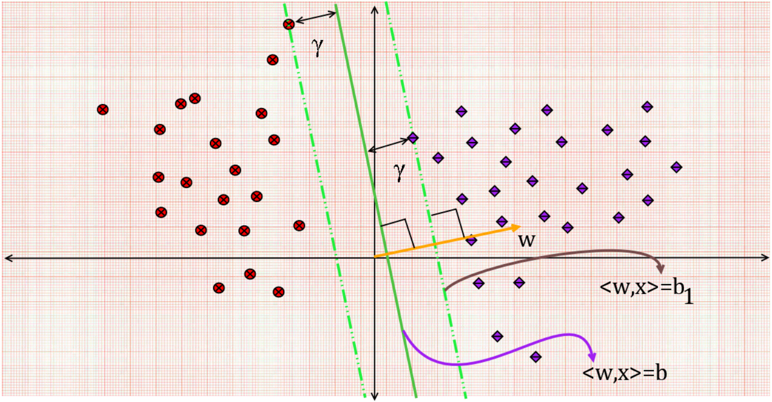
SVMs: Geometric idea



- **Formalized motivation:** Thus we seek a separating hyperplane which has the **maximum margin**.

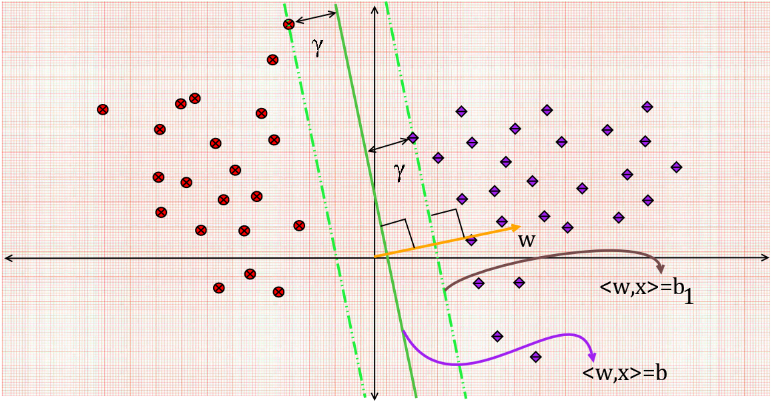
SVMs: Geometric idea

Computing the margin:



SVMs: Geometric idea

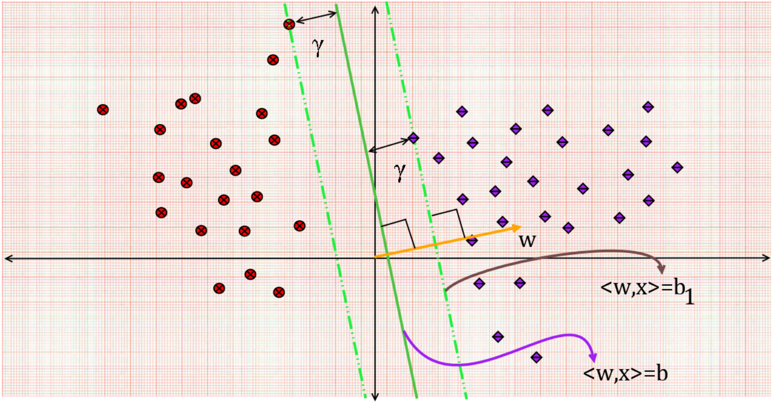
Computing the margin:



- Let the equation of the separating hyperplane be $\langle w, x \rangle = b$.

SVMs: Geometric idea

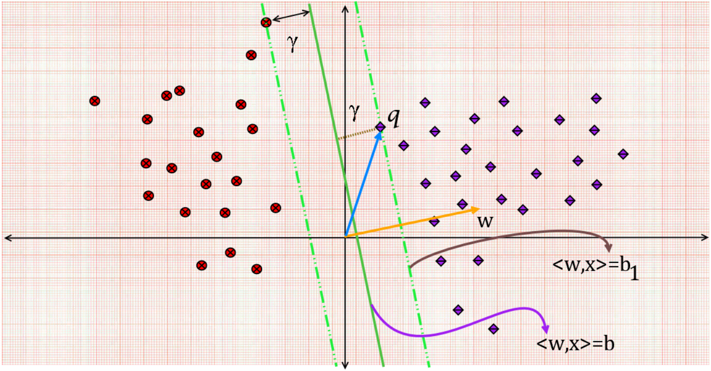
Computing the margin:



- Let the equation of the separating hyperplane be $\langle w, x \rangle = b$.
- Consider the hyperplane parallel to the separating hyperplane passing through the closest point of a particular class to be $\langle w, x \rangle = b_1$. (Check this claim!)

SVMs: Geometric idea

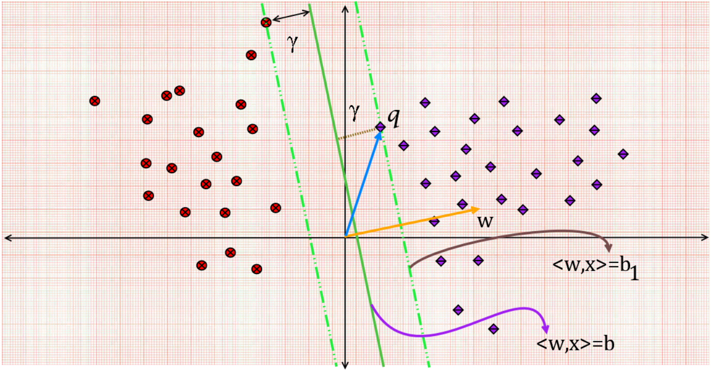
Computing the margin:



- Consider q to be a sample of a particular class closest to the separating hyperplane $\langle w, x \rangle = b$.

SVMs: Geometric idea

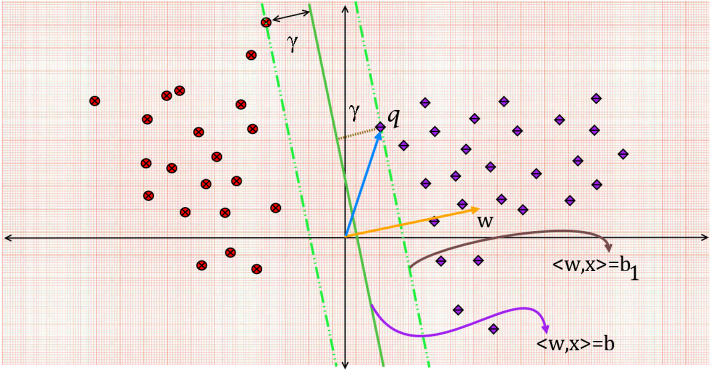
Computing the margin:



- Consider q to be a sample of a particular class closest to the separating hyperplane $\langle w, x \rangle = b$.

SVMs: Geometric idea

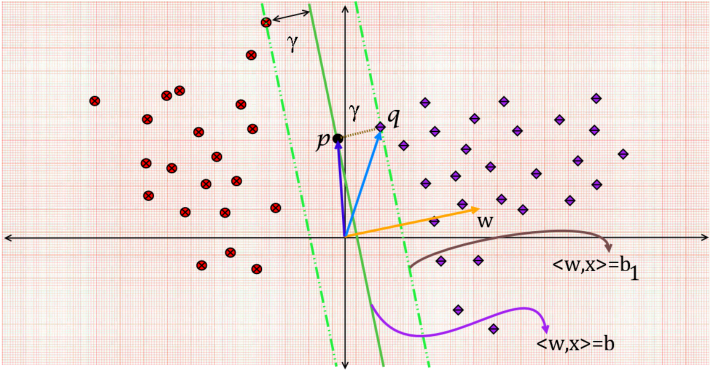
Computing the margin:



- Consider q to be a sample of a particular class closest to the separating hyperplane $\langle w, x \rangle = b$.
- The closeness is determined by the distance of q from the hyperplane $\langle w, x \rangle = b$.

SVMs: Geometric idea

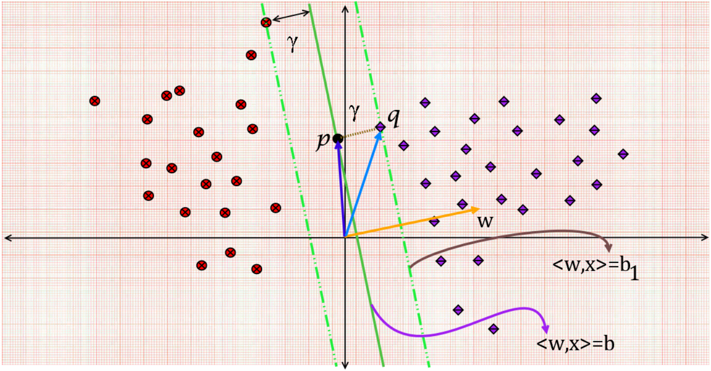
Computing the margin:



- Consider q to be a sample of a particular class closest to the separating hyperplane $\langle w, x\rangle = b$. Hence $\langle w, q\rangle = b_1$.
- The closeness is determined by the distance of q from the hyperplane $\langle w, x\rangle = b$.
- Note that distance from q to the hyerplane $\langle w, x\rangle = b$ is measured as distance between a point p on the hyperplane $\langle w, x\rangle = b$ which is the closest point to q.

SVMs: Geometric idea

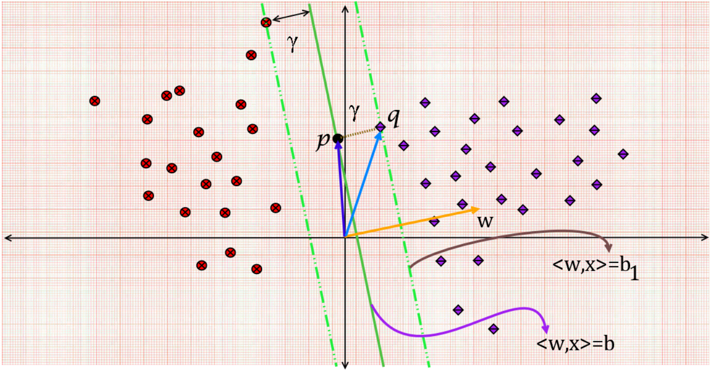
Computing the margin:



- A point p on the hyperplane $\langle w, x\rangle = b$ is closest to q when q lies along the direction of a vector v from p and v is parallel to the normal vector w to the hyerplane. (Try to prove this claim!)

SVMs: Geometric idea

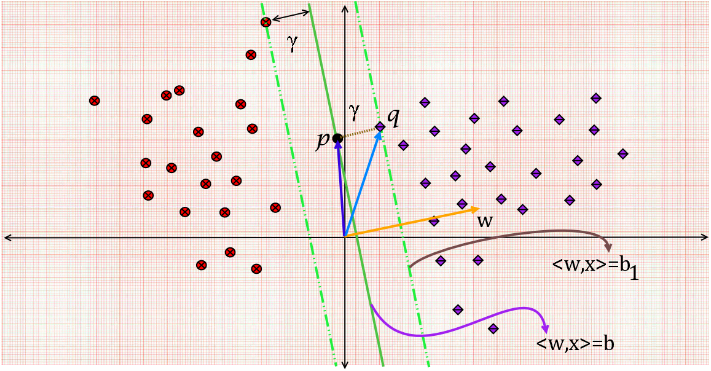
Computing the margin:



- Since q is along direction v parallel to w from the point p, we can also write q as: $q = p + a\frac{w}{\|w\|_2}$ for some scalar $a \in \mathbb{R}$. (Check this claim!)

SVMs: Geometric idea

Computing the margin:

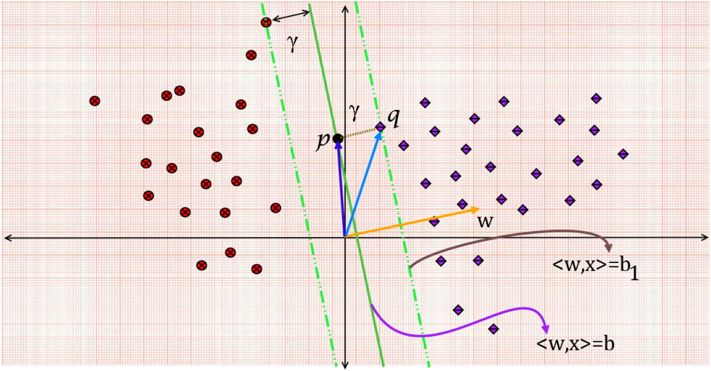


- Since q is along direction w from the point p, we can also write q as: $q = p + a\frac{w}{\|w\|_2}$ for some scalar $a \in \mathbb{R}$.
- Hence we have after multiplying on both sides by $w^\top$:

$$w^\top q = w^\top p + a\frac{w^\top w}{\|w\|_2} \iff \langle w, q\rangle = \langle w, p\rangle + a\frac{\langle w, w\rangle}{\|w\|_2}.$$

SVMs: Geometric idea

Computing the margin:

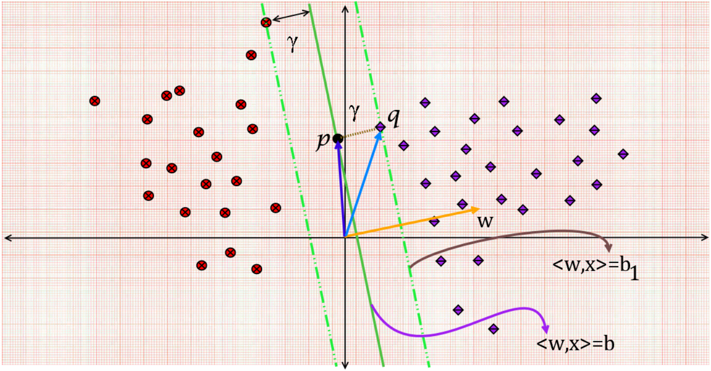


- Since q is along direction w from the point p, we can also write q as: $q = p + a\frac{w}{\|w\|_2}$ for some scalar $a \in \mathbb{R}$.
- Hence we have after multiplying on both sides by $w^\top$:

$$\langle w, q \rangle = \langle w, p \rangle + a\|w\|_2 \implies a = \frac{\langle w, q \rangle - b}{\|w\|_2}.$$

SVMs: Geometric idea

Computing the margin:



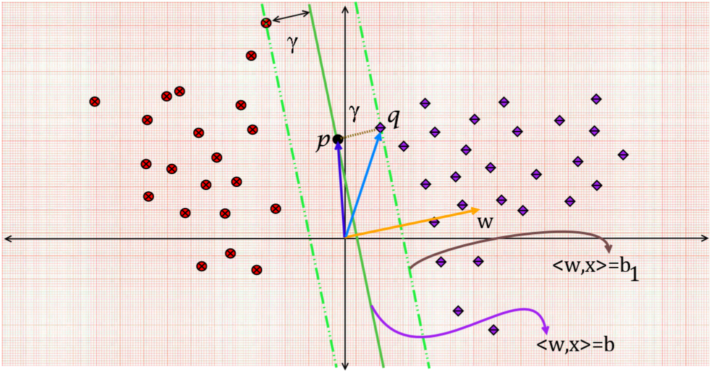
- Hence we have after multiplying on both sides by $w^\top$:

$$\langle w, q\rangle = \langle w, p\rangle + a\|w\|_2 \implies a = \frac{\langle w, q\rangle - b}{\|w\|_2}.$$

- Thus the distance between q and separating hyperplane $\langle w, x\rangle = b$ is obtained by $\gamma = |a| = \frac{|\langle w,q\rangle - b|}{\|w\|_2}$.

SVMs: Geometric idea

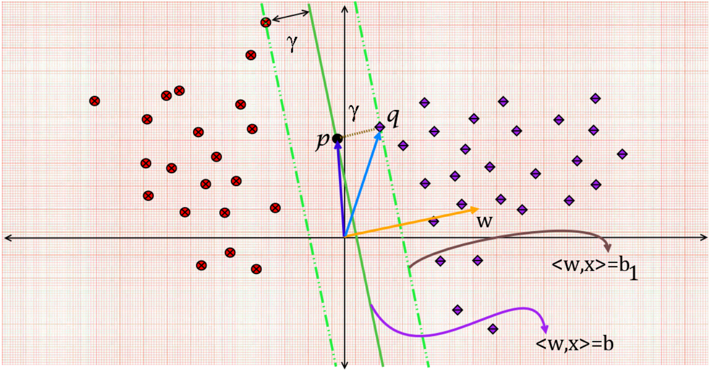
Computing the margin:



- Thus we have the margin as $\gamma = |a| = \frac{|\langle w, q\rangle - b|}{\|w\|_2}$.

SVMs: Geometric idea

Computing the margin:



- Note that the margin computation has been done assuming that the hyperplane separates the points belonging to different classes and using the knowledge of a point q closest to the hyperplane.
- However finding the margin can be cast as a general optimization problem.

SVMs: Geometric idea

Separating hyerplane:

Separating hyerplane

Given a data set $D = \{(x^i, y^i)\}_{i=1}^n$ where $x^i \in \mathbb{R}^d$ and $y^i \in \{+1, -1\}$, a hyperplane $\langle w, x \rangle = b$ is a separating hyperplane for D if:

$$\left\langle w, x^i \right\rangle - b > 0 \text{ when } y^i = +1 \text{ and}$$
$$\left\langle w, x^i \right\rangle - b < 0 \text{ when } y^i = -1$$

SVMs: Geometric idea

Separating hyerplane:

Separating hyerplane

Given a data set $D = \{(x^i, y^i)\}_{i=1}^n$ where $x^i \in \mathbb{R}^d$ and $y^i \in \{+1, -1\}$, a hyperplane $\langle w, x\rangle = b$ is a separating hyperplane for D if for every $i \in \{1, 2, \ldots, n\}$ the following holds:

$$\left\langle w, x^i \right\rangle - b > 0 \text{ when } y^i = +1 \text{ and}$$
$$\left\langle w, x^i \right\rangle - b < 0 \text{ when } y^i = -1$$

Equivalently, a hyperplane $\langle w, x\rangle = b$ is a separating hyperplane for D if:

$$y^i(\left\langle w, x^i \right\rangle - b) > 0, \ \forall i \in \{1, 2, 3, \ldots, n\}.$$

SVMs: Geometric idea

Separating hyerplane:

Separating hyerplane

Given a data set $D = \{(x^i, y^i)\}_{i=1}^n$ where $x^i \in \mathbb{R}^d$ and $y^i \in \{+1, -1\}$, a hyperplane $\langle w, x \rangle = b$ is a separating hyperplane for D if for every $i \in \{1, 2, \dots, n\}$ the following holds:

$$\left\langle w, x^i \right\rangle - b > 0 \text{ when } y^i = +1 \text{ and}$$
$$\left\langle w, x^i \right\rangle - b < 0 \text{ when } y^i = -1$$

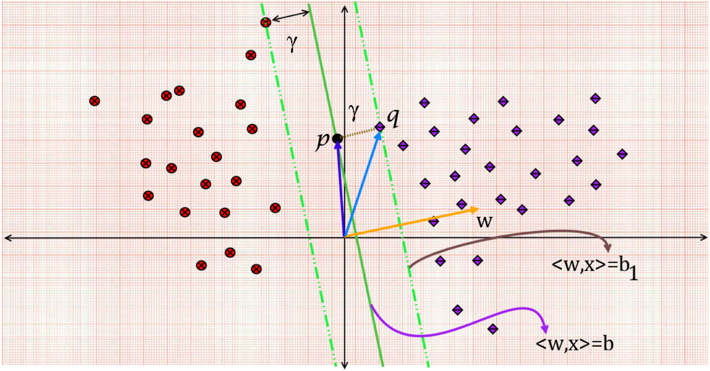
Equivalently, a hyperlane $\langle w, x \rangle = b$ is a separating hyperplane for D if:

$$y^i(\left\langle w, x^i \right\rangle - b) > 0, \ \forall i \in \{1, 2, 3, \dots, n\}.$$

Linear Separability: When a separating hyperplane (w, b) exists for D, then D is said to be Linearly Separable by the hyperplane (w, b).

SVMs: Geometric idea

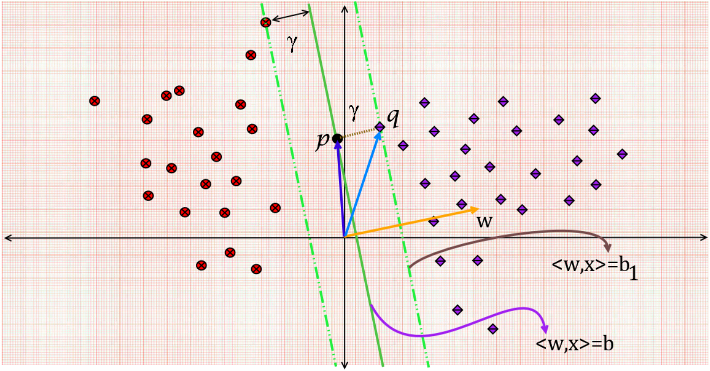
Computing the margin:



- Given a data set $D = \{(x^i, y^i)\}_{i=1}^n$ where $x^i \in \mathbb{R}^d$ and $y^i \in \{+1, -1\}$, we have the distance associated with each sample x^i from the separating hyerplane as $\gamma_i = \frac{|\langle w, x^i \rangle - b|}{\|w\|_2}$.

SVMs: Geometric idea

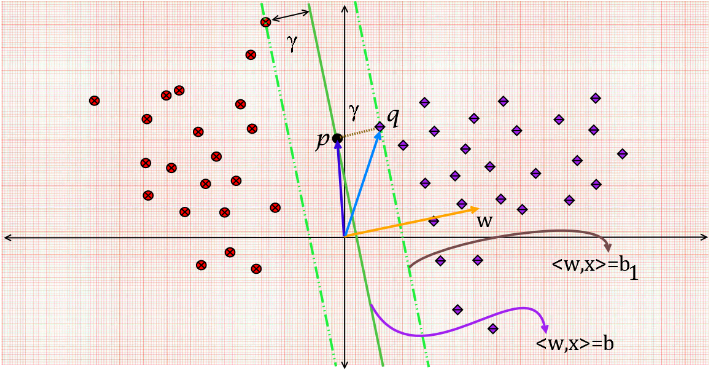
Computing the margin:



- Recall that $\langle w, x^i \rangle - b$ denotes the (signed) displacement of x^i from its closest point on the hyerplane.

SVMs: Geometric idea

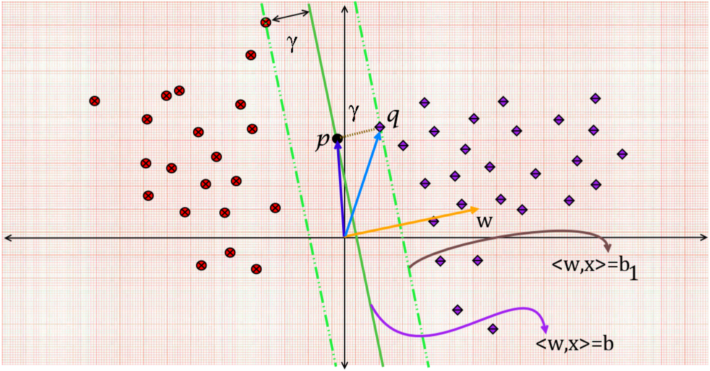
Computing the margin:



- Recall that $\langle w, x^i \rangle - b$ denotes the (signed) displacement of x^i from its closest point on the hyerplane.
- Thus to enforce the orientation of the hyperplane so that the samples with class label $+1$ are on one side of it and samples with class label -1 are on the other side of it, we can write the distance as: $\gamma_i = \frac{y^i\left(\left\langle w, x^i \right\rangle - b\right)}{\|w\|_2}$. **(why?)**

SVMs: Geometric idea

Computing the margin:

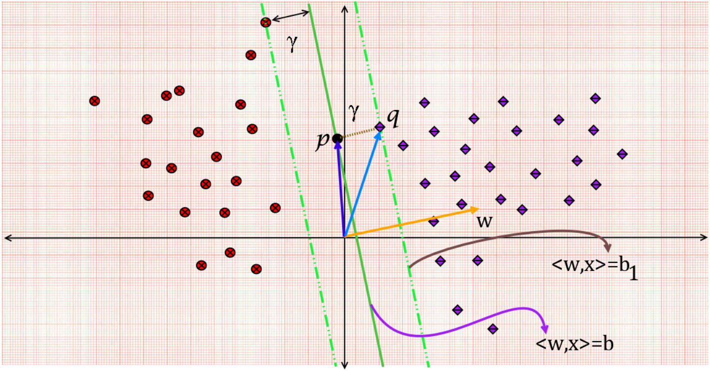


- Given a data set $D = \{(x^i, y^i)\}_{i=1}^n$ where $x^i \in \mathbb{R}^d$ and $y^i \in \{+1, -1\}$, computing the margin γ can be hence cast as the optimization problem:

$$\min_{i\in\{1,2,\ldots,n\}} \gamma_i \iff \min_{i\in\{1,2,\ldots,n\}} \frac{y^i\left(\left\langle w, x^i\right\rangle - b\right)}{\|w\|_2}$$

SVMs: Geometric idea

Computing the margin:



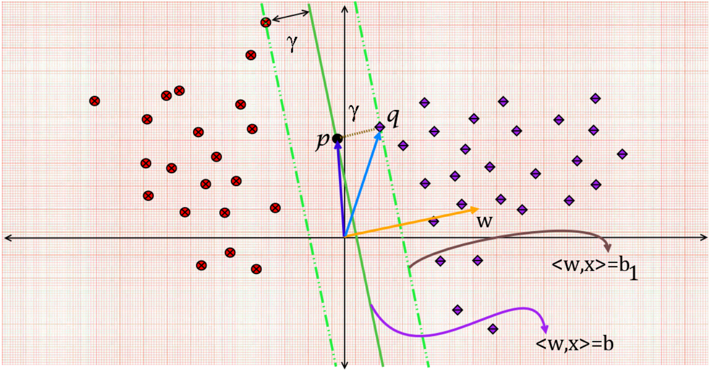
- Given a data set $D = \{(x^i, y^i)\}_{i=1}^n$ where $x^i \in \mathbb{R}^d$ and $y^i \in \{+1, -1\}$, computing the margin γ can be hence cast as the optimization problem:

  $\min_{i \in \{1,2,\ldots,n\}} \gamma_i \iff \min_{i \in \{1,2,\ldots,n\}} \frac{y^i\left(\langle w, x^i \rangle - b\right)}{\|w\|_2}$

- Note however that this assumes knowledge of the hyperplane $\langle w, x \rangle = b$.

SVMs: Geometric idea

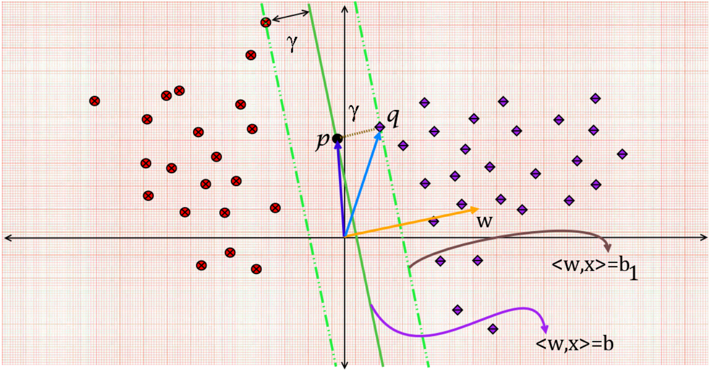
Computing the margin:



- Hence when do not know the separating hyerplane, we wish to find the parameters w, b of the hyperplane which would help to **maximize** the margin γ.

SVMs: Geometric idea

Computing the margin:



- Hence when do not know the separating hyerplane, we wish to find the parameters w, b of the hyperplane which would help to **maximize** the margin γ.
- Thus the overall optimization problem becomes:

$$\max_{w\in\mathbb{R}^d, b\in\mathbb{R}} \min_{i\in\{1,2,\ldots,n\}} \gamma_i \iff \max_{w\in\mathbb{R}^d, b\in\mathbb{R}} \min_{i\in\{1,2,\ldots,n\}} \frac{y^i\left(\left\langle w, x^i\right\rangle - b\right)}{\|w\|_2}$$

SVMs: Geometric idea

A first version of optimization problem for SVM

Given a data set $\{(x^i, y^i)\}_{i=1}^n$ where $x^i \in \mathbb{R}^d$ and $y^i \in \{+1, -1\}$, SVM solves:

$$\max_{w \in \mathbb{R}^d, b \in \mathbb{R}} \min_{i \in \{1,2,\ldots,n\}} \frac{y^i \left(\langle w, x^i \rangle - b \right)}{\|w\|_2}.$$

SVMs: Geometric idea

A first version of optimization problem for SVM

Given a data set $\{(x^i, y^i)\}_{i=1}^n$ where $x^i \in \mathbb{R}^d$ and $y^i \in \{+1, -1\}$, SVM solves:

$$\max_{w \in \mathbb{R}^d, b \in \mathbb{R}} \min_{i \in \{1,2,\dots,n\}} \frac{y^i \left(\langle w, x^i \rangle - b \right)}{\|w\|_2}.$$

This problem can be equivalently written as:

$$\begin{aligned} &\max_{w \in \mathbb{R}^d, b \in \mathbb{R}, \mu \in \mathbb{R}} \mu \\ &\text{such that } \frac{y^i \left(\langle w, x^i \rangle - b \right)}{\|w\|_2} \geq \mu, \ \forall i \in \{1, 2, \dots, n\}. \end{aligned}$$

SVMs: Geometric idea

Considering the problem:

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

$$\text{such that } \frac{y^i\left(\langle w, x^i \rangle - b\right)}{\|w\|_2} \geq \mu, \ \forall i \in \{1, 2, \ldots, n\}.$$

we note that $\|w\|_2$ can be chosen to be any positive quantity by choosing an appropriate w and b without changing the solution of the problem. (Check!)

SVMs: Geometric idea

Considering the problem:

$$\max_{w \in \mathbb{R}^d, b \in \mathbb{R}, \mu \in \mathbb{R}} \mu$$
$$\text{such that} \quad \frac{y^i \left(\langle w, x^i \rangle - b \right)}{\|w\|_2} \geq \mu, \ \forall i \in \{1, 2, \ldots, n\}.$$

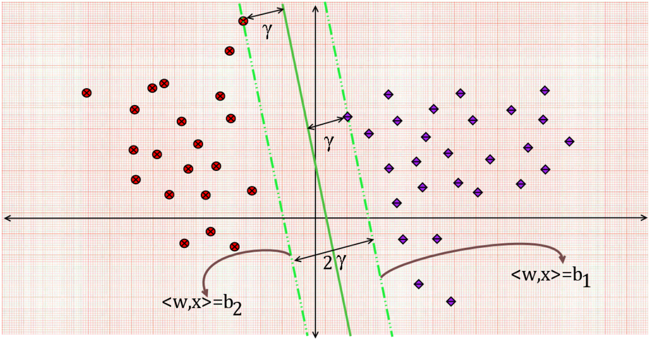
we note that $\|w\|_2$ can be chosen to be any positive quantity by choosing an appropriate w and b without changing the solution of the problem. (Check!)

- In particular, we can set $\|w\|_2 = \frac{1}{\mu}$ to get:

$$\max_{w \in \mathbb{R}^d, b \in \mathbb{R}} \frac{1}{\|w\|_2}$$
$$\text{such that} \quad y^i \left(\langle w, x^i \rangle - b \right) \geq 1, \ \forall i \in \{1, 2, \ldots, n\}.$$

SVMs: Geometric idea

SVM Optimization problem:



- Thus given a data set $\{(x^i, y^i)\}_{i=1}^n$ where $x^i \in \mathbb{R}^d$ and $y^i \in \{+1, -1\}$, SVM solves:

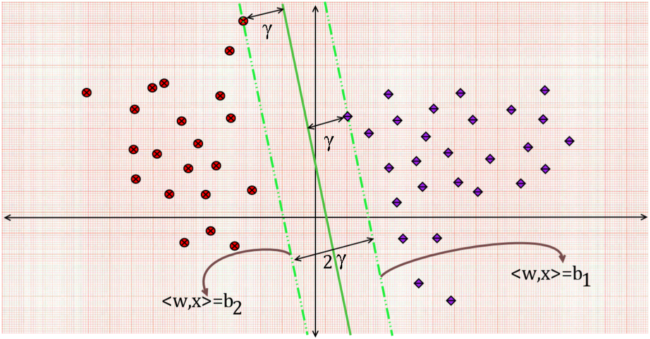
$$\max_{w,b} \frac{1}{\|w\|_2} \text{ s.t.}$$

$$\left\langle w, x^i \right\rangle \geq 1 + b, \text{ if } y^i = +1,$$

$$\left\langle w, x^i \right\rangle \leq b - 1, \text{ if } y^i = -1.$$

SVMs: Geometric idea

SVM Optimization problem:



- Equivalently, given a data set $\{(x^i, y^i)\}_{i=1}^n$ where $x^i \in \mathbb{R}^d$ and $y^i \in \{+1, -1\}$, hard-margin SVM solves:

$$\min_{w,b} \|w\|_2^2 \text{ s.t.}$$

$$y^i \left(\left\langle w, x^i \right\rangle - b \right) \geq 1, \forall i \in \{1, 2, \ldots, n\}.$$

SVMs: Geometric idea

Hard-margin SVM

Given a data set $\{(x^i, y^i)\}_{i=1}^n$ where $x^i \in \mathbb{R}^d$ and $y^i \in \{+1, -1\}$, hard-margin SVM solves:

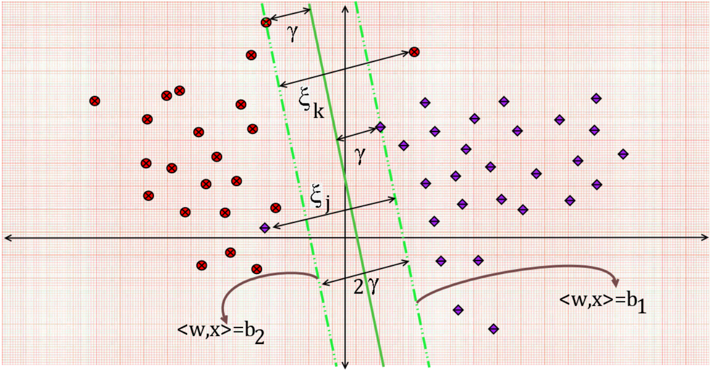
$$\min_{w,b} \frac{1}{2}\|w\|_2^2 \text{ s.t.}$$

$$y^i(\langle w, x^i \rangle - b) \geq 1, \forall i \in \{1, 2, \ldots, n\}.$$

Note the presence of a constant $\frac{1}{2}$ in the objective function of this optimization problem!

SVMs: Geometric idea

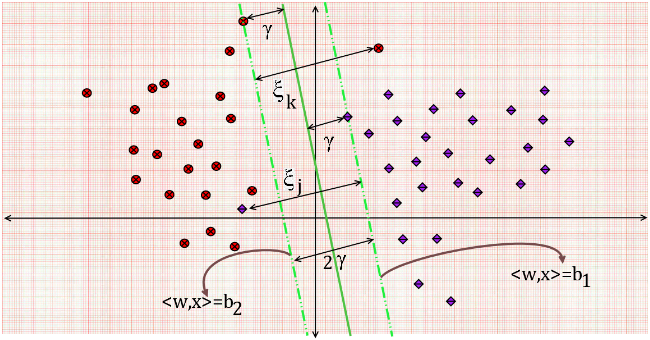
SVM Optimization problem:



- We associate a penalty (or) slack $\xi_i > 0$ whenever i-th sample is mislabeled.
- These variables are also included in the optimization problem.

SVMs: Geometric idea

SVM Optimization problem:



- Given a data set $\{(x^i, y^i)\}_{i=1}^n$ where $x^i \in \mathbb{R}^d$ and $y^i \in \{+1, -1\}$, soft-margin SVM solves:

$$\min_{w,b,\xi_i \geq 0} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n}\xi_i \text{ s.t.}$$

$$y^i(\left\langle w, x^i \right\rangle - b) \geq 1 - \xi_i \forall i \in \{1, 2, \dots, n\}.$$

SVMs: Geometric idea

Soft-margin SVM

Given a data set $\{(x^i, y^i)\}_{i=1}^n$ where $x^i \in \mathbb{R}^d$ and $y^i \in \{+1, -1\}$, soft-margin SVM solves:

$$\min_{w,b,\xi_i \geq 0} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n}\xi_i \text{ s.t.}$$

$$y^i(\langle w, x^i \rangle - b) \geq 1 - \xi_i \forall i \in \{1, 2, \ldots, n\}.$$

References:

- An Introduction to Support Vector Machines and Other Kernel-based Learning Methods by Nello Cristianini and John Shawe Taylor, Cambridge University Press.
- Chapter 7 in Pattern Recognition and Machine Learning by Chris Bishop (Springer publication).