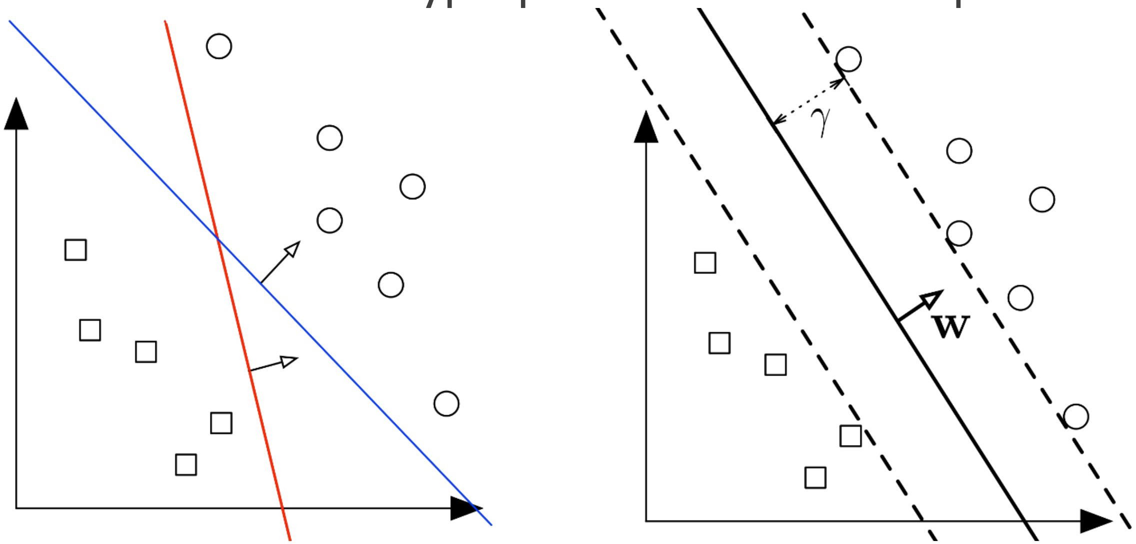
Revision

SVM HYPERPLANE

Support Vector Machines

- **Setting:** We define a linear classifier $h(x) = sign(W^T x + b)$ or and we assume a binary classification setting with labels $y \in \{+1, -1\}$
- **If data is linearly separable,** typically there are infinitely many separating hyperplanes. What is the best separating hyperplane?
- **SVM Answer:** The hyperplane that maximizes the distance to the closest data points from both classes – the hyperplane with **<u>maximum margin</u>**
- A hyperplane is defined as a set of points such that $H = \{x | W^T x + b = 0\}$. The margin γ is the distance from the hyperplane to the closest point across both classes.



Credit: Machine Learning for Intelligent Systems, Kilian Weinberger, Cornell, Lecture 9

Support Vector Machines

- Remember that the shortest distance of a point x from a hyperplane H defined by the normal vector $\vec{w}$

$$distance\ (\vec{w}.\vec{x}) = |\vec{w}.\vec{x}| \quad \text{(if } \vec{w} \text{ is a unit vector)}$$

Absolute value of a dot product between vector w and vector x

- **Side notes:**
  - A unit vector is a vector that has a magnitude of 1.
  - In other words, it is a vector that has been normalized to have a length of 1.
  - A unit vector describe direction independently of magnitude.

$$\boldsymbol{||\vec{w}|| = \sqrt{\textstyle\sum_i^n (w_i)^2} = 1 \text{ (if } \vec{w} \text{ is a unit vector)}}$$

Also known as L_2 norm and represented as: $||\vec{w}||_2$

  - For a vector $\vec{w} = [3, 4]$, unit vector would be:

$$\boldsymbol{||\vec{w}|| = \sqrt{3^2 + 4^2} = \sqrt{25} = 5}$$

$$\boldsymbol{Unit\ Vector\ (\vec{w}) = \left[\frac{3}{5}, \frac{4}{5}\right] = [0.6, 0.8]}$$

Credit:Machine Learning for Intelligent Systems, Kilian Weinberger, Cornell, Lecture 9

Support Vector Machines

- Remember that the shortest distance of a point x from a hyperplane H defined by the normal vector $\vec{w}$

$$distance\ (\vec{w}.\vec{x}) = |\vec{w}.\vec{x}| \quad \text{(if } \vec{w} \text{ is a unit vector)}$$

Absolute value of a dot product between vector w and vector x

- More generally

$$distance\ (\vec{w}.\vec{x}) = \frac{|\vec{w}.\vec{x}|}{||\vec{w}||_2} = \frac{|\vec{w}.\vec{x} + b|}{||\vec{w}||_2} = \frac{|\vec{w}^T.\vec{x} + b|}{||\vec{w}||_2} \quad \text{(if } \vec{w} \text{ is not a unit vector)}$$

- Here for the margin of H will be defined as:

$$\gamma\ (w.b) = \min_{x \in D} \frac{|\vec{w}^T.\vec{x} + b|}{||\vec{w}||_2}$$

Margin is the distance to the closest data points i.e., smallest distance in the classes of both sides.

Support Vector Machines

- **Note:** When γ is maximized, the hyperplane must lie right in the middle of the two classes.
  - Like a two-lane road!

- γ must be the distance to the closest point within both classes.
  - If not, the hyperplane could be moved towards data points of the class, that is further away and increase γ, which contradicts that γ is maximized!

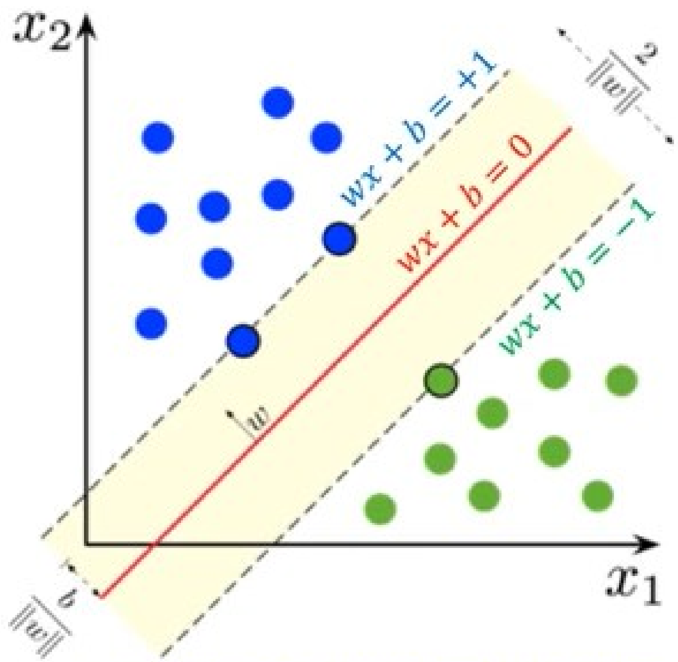


Image ref: https://en.wikipedia.org/wiki/Support-vector_machine

Formal Derivation

- We can formulate our search for maximum margin separating hyperplane as a constrained optimization problem
- The object is to maximize the margin under the constraints that all data points must lie on the correct side of the hyperplane

$$\underbrace{\max_{w,b} \gamma(w,b)}_{\text{Maximum margin}} \quad \textit{such that} \quad \underbrace{\forall i \; y_i(W^T x_i + b) \geq 0}_{\text{Separating hyperplane}}$$

- If we plug in the definition of γ, we obtain:

$$\max_{w,b}\left(\min_{x \in D} \frac{|\vec{w}^T . \vec{x} + b|}{||\vec{w}||_2}\right) \quad \textit{such that} \quad \forall i \; y_i(W^T x_i + b) \geq 0$$

Maximize over normal vector w and b, over the minimum distance that you have from any point x for all points belonging to D

i.e., what could be the width of the "road" at max!

Formal Derivation

$$\max_{w,b}\left(\min_{x\in D}\frac{\left|\vec{w}^T.\vec{x}+b\right|}{\left|\left|\vec{w}\right|\right|_2}\right) \quad \text{such that} \quad \forall i\ y_i\left(W^T x_i+b\right)\geq 0$$

- As $\left|\left|\vec{w}\right|\right|_2$ will not be impacted by the $\min_{x\in D}$, taking it outside:

$$\max_{w,b}\frac{1}{\left|\left|\vec{w}\right|\right|_2}\left(\min_{x\in D}\left|\vec{w}^T.\vec{x}+b\right|\right) \quad \text{such that} \quad \forall i\ y_i\left(W^T x_i+b\right)\geq 0$$

Maximum margin

Separating hyperplane

- Remember we said the hyperplane is scale invariant, we can fix the scan of w, b anyway we want:

$$\min_{x\in D}\left|\vec{w}^T.\vec{x}+b\right|=1$$

- We can add this re-scaling as an equality constraint. Then our objective becomes:

$$\max_{w,b}\frac{1}{\left|\left|\vec{w}\right|\right|_2}.1=\min_{w,b}\left|\left|w\right|\right|_2=\min_{w,b}w^T w$$

How?

Formal Derivation

$$\max_{w,b} \frac{1}{||\vec{w}||_2} . 1 = \min_{w,b} ||w||_2$$

Maximizing left hand side is same as minimizing right hand side

$$\min_{w,b} ||w||_2 = \min_{w,b} w^T w$$

As L_2 norm of w is nothing but dot product with it self, adding all results and then taking square root, thus, $\sqrt{w.w} = \sqrt{w^T w}$

- **Recall: For a vector $\vec{w} = [3, 4]$, unit vector would be:**

$$||\vec{w}|| = \sqrt{\sum_{i}^{n} (w_i)^2}$$

$$||\vec{w}|| = \sqrt{3^2 + 4^2} = \sqrt{25} = 5$$

Formal Derivation

$$\boxed{\max_{w,b} \frac{1}{||\vec{w}||_2} . 1 = \min_{w,b} ||w||_2}$$

Maximizing left hand side is same as minimizing right hand side

$$\boxed{\min_{w,b} ||w||_2 = \min_{w,b} w^T w}$$

As L_2 norm of w is nothing but dot product with it self, adding all results and then taking square root, thus, $\sqrt{w.w} = \sqrt{w^T w}$

- We have ignored the square root in $\sqrt{w^T w}$ as $f(z) = z^2$ is a monotonically increasing function for $z \geq 0$ and $||w|| \geq 0$; i.e., the w that maximizes $||w||_2$ also maximizes $w^T w$

Formal Derivation

- Our new optimization problem becomes:

$$\min_{w,b} w^T w$$
$$\textit{such that} \quad \forall i \; y_i(W^T x_i + b) \geq 0$$
$$\min_i \left| w^T x_i + b \right| = 1$$

- These constraints are still hard to deal with, however, luckily we can show that for the optimal solution, these constraints are equivalent to a much simpler formulation:

$$\min_{w,b} w^T w$$
$$\textit{such that} \quad \forall i \; y_i(W^T x_i + b) \geq 1$$

Can you figure out how above expression is same as bottom expression?

- Second constraint in above expression says in all the dataset, the minimum distance to any point x_i should be 1. If minimum distance is 1, then all other distances are greater than or equal to 1.
- Bottom expression is also equal to above expression, even if the distance is 10, we can rescale w and b

Formal Derivation

- That's the final version of SVM equation!

$$\boxed{\begin{array}{c} \min_{w,b} w^T w \\ \textit{such that} \quad \forall i \; y_i\left(W^T x_i + b\right) \geq 1 \end{array}}$$

- Find the simplest hyperplane, where simpler means smaller $w^T w$, such that all inputs lie at least 1 unit away from the hyperplane on the correct side!

- This is a linearly constrained quadratic optimization problem, that could be solved using quadratic programming, e.g., using Lagrange Multipliers.
  - Recall differential equation

- That's one way of solving it. There are other ways of solving it as well.

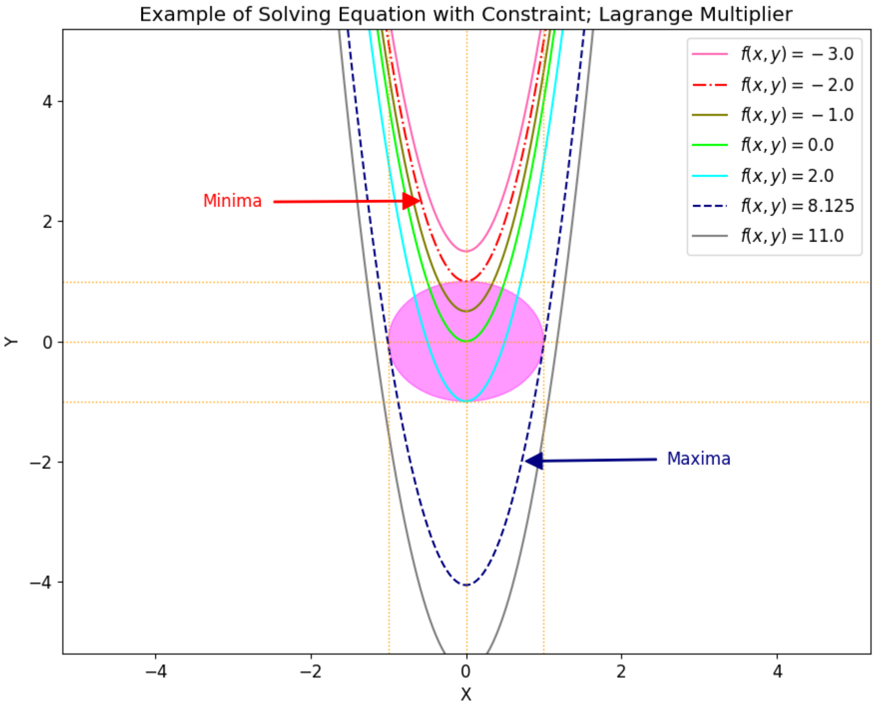
Side Note: Constrained Optimization and Lagrange Multipliers

- Suppose we are given a function $f(x, y, z, \dots)$ for which we want to find the extrema, subject to the condition $g(x, y, z, \dots) = k$

- The idea used in Langrange multiplier is that the gradient of the objective function f, lines up wither in parallel or anti-parallel direction to the gradient of the constraint g, at an optimal point. In such case, one of the gradients should be some multiple of another.

- E.g.,

$$f(X, y) = 8x^2 + 2y$$
$$g(x, y) = x^2 + y^2$$
$$\nabla f = \lambda \nabla g$$



- The extrema of constrained function f, lie on the surface of the constraint g, which is a circle of unit radius .
- It is a necessary condition
- The tangent vectors of the function and the constraint are either parallel or anti-parallel at each extremum.

Credit: https://towardsdatascience.com/understanding-support-vector-machine-part-1-lagrange-multipliers-5c24a52ffc5e

SVMs

- For the optimal w, b pair, some training points will have tight constraints i.e.,

$$\boxed{\boldsymbol{y_i\left(W^T x_i + b\right) = 1}}$$

  - This must be the case, because if for al training points, we had a strict $>$ inequality, it would be possible to scale down both parameters w, b until the constraints are tight and obtained an even lower objective value.

- We refer to these training points as **<u>support vectors</u>**
  - Support vectors are special because they are the training points that define the maximum margin of the hyperplane to the dataset and they therefore determine the shape of the hyperplane.
  - If you were to move on of them and retrain the svm, resulting hyperplane would change
  - The opposite is the case for non-support vectors (provided that you don't move them so much that they turn into support vectors themselves)
  - This will become particularly important in the dual formulation for kernel-SVMs.

Book Reading

- Murphy – Chapter 1, Chapter 14