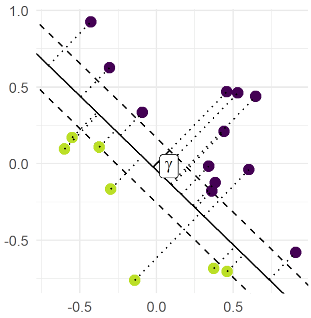
Introduction to Machine Learning

Linear Support Vector Machines

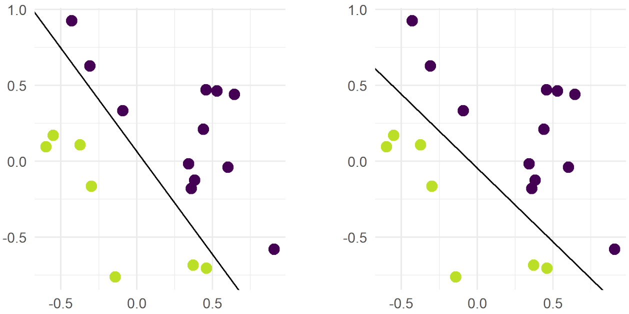
Linear Hard Margin SVM



Learning goals

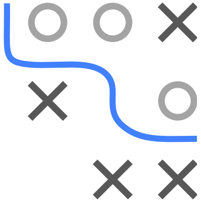
- Know that the hard-margin SVM maximizes the margin between data points and hyperplane
- Know that this is a quadratic program
- Know that support vectors are the data points closest to the separating hyperplane

LINEAR CLASSIFIERS



- We want study how to build a binary, linear classifier from solid geometrical principles.
- Which of these two classifiers is “better”?

SUPPORT VECTOR MACHINES: GEOMETRY



For labeled data $\mathcal{D} = \left(\left(\mathbf{x}^{(1)}, y^{(1)}\right), \ldots, \left(\mathbf{x}^{(n)}, y^{(n)}\right)\right)$, with $y^{(i)} \in \{-1, +1\}$:

- Assume linear separation by $f(\mathbf{x}) = \boldsymbol{\theta}^\top \mathbf{x} + \theta_0$, such that all $+$-observations are in the positive halfspace

$$\{\mathbf{x} \in \mathcal{X} : f(\mathbf{x}) > 0\}$$

  and all $-$-observations are in the negative halfspace

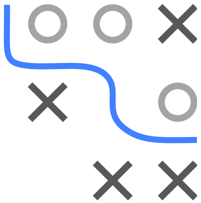
$$\{\mathbf{x} \in \mathcal{X} : f(\mathbf{x}) < 0\}.$$

- For a linear separating hyperplane, we have

$$y^{(i)} \underbrace{\left(\boldsymbol{\theta}^\top \mathbf{x}^{(i)} + \theta_0\right)}_{=f\left(\mathbf{x}^{(i)}\right)} > 0 \quad \forall i \in \{1, 2, ..., n\}.$$

MAXIMUM MARGIN SEPARATION

We formulate the desired property of a large "safety margin" as an optimization problem:

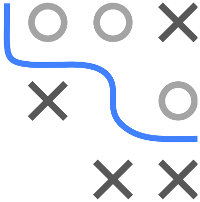


$$\max_{\boldsymbol{\theta}, \theta_0} \quad \gamma$$

$$\text{s.t.} \quad d\left(f, \mathbf{x}^{(i)}\right) \geq \gamma \quad \forall i \in \{1, \dots, n\}.$$

- The constraints mean: We require that any instance i should have a "safety" distance of at least γ from the decision boundary defined by $f (= \boldsymbol{\theta}^T \mathbf{x} + \theta_0) = 0$.
- Our objective is to maximize the "safety" distance.

MAXIMUM MARGIN SEPARATION



We reformulate the problem:

$$\max_{\boldsymbol{\theta},\theta_0} \quad \gamma$$

$$\text{s.t.} \quad \frac{y^{(i)}\left(\langle \boldsymbol{\theta}, \mathbf{x}^{(i)} \rangle + \theta_0\right)}{\|\boldsymbol{\theta}\|} \geq \gamma \quad \forall\, i \in \{1, \ldots, n\}.$$

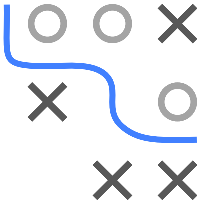
- The inequality is rearranged by multiplying both sides with $\|\boldsymbol{\theta}\|$:

$$\max_{\boldsymbol{\theta},\theta_0} \quad \gamma$$

$$\text{s.t.} \quad y^{(i)}\left(\left\langle \boldsymbol{\theta}, \mathbf{x}^{(i)} \right\rangle + \theta_0\right) \geq \|\boldsymbol{\theta}\|\gamma \quad \forall\, i \in \{1, \ldots, n\}.$$

QUADRATIC PROGRAM

We derived the following optimization problem:



$$\min_{\boldsymbol{\theta},\theta_0} \quad \frac{1}{2}\|\boldsymbol{\theta}\|^2$$

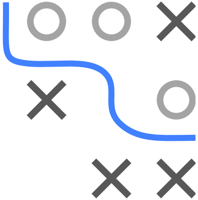
$$\text{s.t.} \quad y^{(i)}\left(\left\langle \boldsymbol{\theta}, \mathbf{x}^{(i)} \right\rangle + \theta_0\right) \geq 1 \quad \forall\, i \in \{1, \ldots, n\}.$$

This turns out to be a **convex optimization problem** – particularly, a **quadratic program**: The objective function is quadratic, and the constraints are linear inequalities.

This is called the **primal** problem. We will later show that we can also derive a dual problem from it.

We will call this the **linear hard-margin SVM**.

SUPPORT VECTORS



- Some $(\mathbf{x}^{(i)}, y^{(i)})$ will have minimal margin, $y^{(i)} f\left(\mathbf{x}^{(i)}\right) = 1$, fulfilling the inequality constraints with equality.
- Implies a distance of $\gamma = 1/\|\boldsymbol{\theta}\|$ from separating hyperplane.
- Geometrically obvious that optimal hyperplane doesn't depend on observations with larger distance.
- Hence, we call some of these minimal margin vectors (but not necessarily all) support vectors.
- More formal definition: in upcoming section on duality.