Machine Learning

Support Vector Machines

Fabio Vandin

November 22^{nd}, 2021

Classification and Margin

Consider a classification problem with two classes:

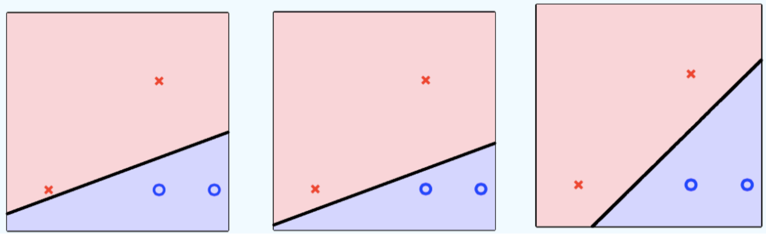
- instance set $\mathcal{X} = \mathbb{R}^d$
- label set $\mathcal{Y} = \{-1, 1\}$.

Training data: $S = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m))$

Hypothesis set $\mathcal{H}$ = halfspaces

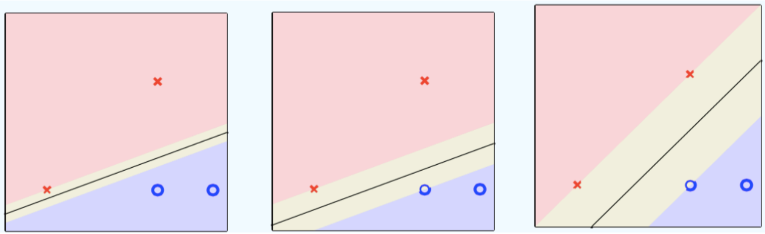
Assumption: data is linearly separable $\Rightarrow$ there exist a halfspace that perfectly classify the training set

In general: multiple separating hyperplanes: $\Rightarrow$ which one is the best choice?



Classification and Margin

The last one seems the best choice, since it can tolerate more "noise".



Informally, for a given separating halfspace we define its *margin* as its minimum distance to an example in the training set S.

Intuition: best separating hyperplane is the one with largest margin.

How do we find it?

Linearly Separable Training Set

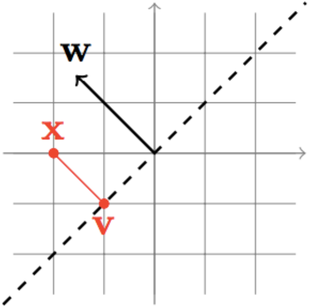
Training set $S = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m))$ is *linearly separable* if there exists a halfspace $(\mathbf{w}, b)$ such that $y_i = \text{sign}(\langle \mathbf{w}, \mathbf{x}_i \rangle + b)$ for all $i = 1, \ldots, m$.

Equivalent to:

$$\forall i = 1, \ldots, m : y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) > 0$$

Informally: *margin* of a separating hyperplane is its minimum distance to an example in the training set S

Separating Hyperplane and Margin



Given hyperplane defined by $L = \{\mathbf{v} : \langle \mathbf{w}, \mathbf{v} \rangle + b = 0\}$, and given $\mathbf{x}$, the distance of $\mathbf{x}$ to L is

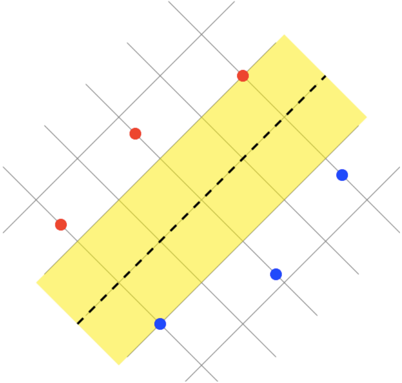
$$d(\mathbf{x}, L) = \min\{||\mathbf{x} - \mathbf{v}|| : \mathbf{v} \in L\}$$

Claim: if $||\mathbf{w}|| = 1$ then $d(\mathbf{x}, L) = |\langle \mathbf{w}, \mathbf{x} \rangle + b|$ (Proof: Claim 15.1 [UML])

Margin and Support Vectors

The *margin* of a separating hyperplane is the distance of the closest example in training set to it. If $||\mathbf{w}|| = 1$ the margin is:

$$\min_{i \in \{1,\ldots,m\}} |\langle \mathbf{w}, \mathbf{x}_i \rangle + b|$$



The closest examples are called *support vectors*

Support Vector Machine (SVM)

Hard-SVM: seek for the separating hyperplane with largest margin (only for linearly separable data)

Computational problem:

$$\arg \max_{(\mathbf{w},b):||\mathbf{w}||=1} \min_{i\in\{1,\ldots,m\}} |\langle \mathbf{w}, \mathbf{x}_i \rangle + b|$$

subject to $\forall i : y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) > 0$

Equivalent formulation (due to separability assumption):

$$\arg \max_{(\mathbf{w},b):||\mathbf{w}||=1} \min_{i\in\{1,\ldots,m\}} y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b)$$

Hard-SVM: Quadratic Programming Formulation

- **input**: $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$
- **solve**:

$$(\mathbf{w}_0, b_0) = \arg\min_{(\mathbf{w}, b)} ||\mathbf{w}||^2$$

subject to $\forall i : y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1$

- **output**: $\hat{\mathbf{w}} = \frac{\mathbf{w}_0}{||\mathbf{w}_0||}, \hat{b} = \frac{b_0}{||\mathbf{w}_0||}$

Proposition

The output of algorithm above is a solution to the *Equivalent Formulation* in the previous slide.

How do we get a solution? Quadratic optimization problem: objective is convex quadratic function, constraints are linear inequalities $\Rightarrow$ Quadratic Programming solvers!

Equivalent Formulation and Support Vectors

Equivalent formulation (homogeneous halfspaces): assume first component of $\mathbf{x} \in \mathcal{X}$ is 1, then

$$\mathbf{w}_0 = \min_{\mathbf{w}} ||\mathbf{w}||^2 \text{ subject to } \forall i : y_i \langle \mathbf{w}, \mathbf{x}_i \rangle \geq 1$$

"Support Vectors" = vectors at minimum distance from $\mathbf{w}_0$

The support vectors are the only ones that matter for defining $\mathbf{w}_0$!

Proposition

Let $\mathbf{w}_0$ be as above. Let $I = \{i : |\langle \mathbf{w}_0, \mathbf{x}_i \rangle| = 1\}$. Then there exist coefficients $\alpha_1, \ldots, \alpha_m$ such that

$$\mathbf{w}_0 = \sum_{i \in I} \alpha_i \mathbf{x}_i$$

"Support vectors" $= \{\mathbf{x}_i : i \in I\}$

Note: Solving Hard-SVM is equivalent to find α_i for $i = 1, \ldots, m$, and $\alpha_i \neq 0$ only for support vectors

Soft-SVM

Hard-SVM works if data is linearly separable.

What if data is not linearly separable? $\Rightarrow$ soft-SVM

Idea: modify constraints of Hard-SVM to allow for some violation, but take into account violations into objective function

Soft-SVM Constraints

Hard-SVM constraints:

$$y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1$$

Soft-SVM constraints:

- *slack variables*: $\xi_1, \ldots, \xi_m \geq 0 \Rightarrow$ vector $\boldsymbol{\xi}$
- for each $i = 1, \ldots, m$: $y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 - \xi_i$
- ξ_i: how much constraint $y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1$ is violated

Soft-SVM minimizes combinations of

- norm of $\mathbf{w}$
- average of ξ_i

Tradeoff among two terms is controlled by a parameter $\lambda \in \mathbb{R}, \lambda > 0$

Soft-SVM: Optimization Problem

- **input**: $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$, parameter $\lambda > 0$
- **solve**:

$$\min_{\mathbf{w},b,\xi} \left(\lambda ||\mathbf{w}||^2 + \frac{1}{m} \sum_{i=1}^{m} \xi_i \right)$$

  subject to $\forall i : y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 - \xi_i$ and $\xi_i \geq 0$
- **output**: $\mathbf{w}, b$

Equivalent formulation: consider the *hinge loss*

$$\ell^{\text{hinge}}((\mathbf{w}, b), (\mathbf{x}, y)) = \max\{0, 1 - y(\langle \mathbf{w}, \mathbf{x} \rangle + b)\}$$

Given $(\mathbf{w}, b)$ and a training S, the empirical risk $L_S^{\text{hinge}}((\mathbf{w}, b))$ is

$$L_S^{\text{hinge}}((\mathbf{w}, b)) = \frac{1}{m} \sum_{i=1}^{m} \ell^{\text{hinge}}((\mathbf{w}, b), (\mathbf{x}_i, y_i))$$

Soft-SVM as RLM

Soft-SVM: solve

$$\min_{\mathbf{w},b,\xi} \left(\lambda ||\mathbf{w}||^2 + \frac{1}{m} \sum_{i=1}^{m} \xi_i \right)$$

subject to $\forall i : y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 - \xi_i$ and $\xi_i \geq 0$

Equivalent formulation with hinge loss:

$$\min_{\mathbf{w},b} \left(\lambda ||\mathbf{w}||^2 + L_S^{\text{hinge}}(\mathbf{w}, b) \right)$$

that is

$$\min_{\mathbf{w},b} \left(\lambda ||\mathbf{w}||^2 + \frac{1}{m} \sum_{i=1}^{m} \ell^{\text{hinge}}((\mathbf{w}, b), (\mathbf{x}_i, y_i)) \right)$$

Note:

- $\lambda ||\mathbf{w}||^2$: ℓ_2 regularization
- $L_S^{\text{hinge}}(\mathbf{w}, b)$: empirical risk for hinge loss

Soft-SVM: Solution

We need to solve:

$$\min_{\mathbf{w},b} \left(\lambda ||\mathbf{w}||^2 + \frac{1}{m} \sum_{i=1}^{m} \ell^{\text{hinge}}((\mathbf{w}, b), (\mathbf{x}_i, y_i)) \right)$$

where

$$\ell^{\text{hinge}}((\mathbf{w}, b), (\mathbf{x}, y)) = \max\{0, 1 - y(\langle \mathbf{w}, \mathbf{x} \rangle + b)\}$$

How?

- standard solvers for optimization problems
- **Stochastic Gradient Descent**