Support Vector Machine

CSC 461: Machine Learning

Fall 2020

Prof. Marco Alvarez
University of Rhode Island

Quick Notes

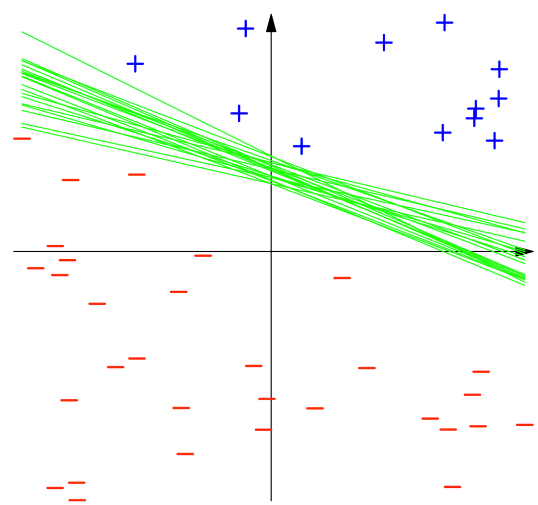
- Assignment
  - ✓ can plot the images with PIL or Matplotlib

Binary classification

- Data instance
  - ✓ $x \in \mathcal{X}, \mathcal{X} = \mathbb{R}^d$
  - ✓ $y \in \mathcal{Y}, \mathcal{Y} = \{-1, +1\}$
- Hypothesis
  - ✓ each hypothesis **g** is a classifier

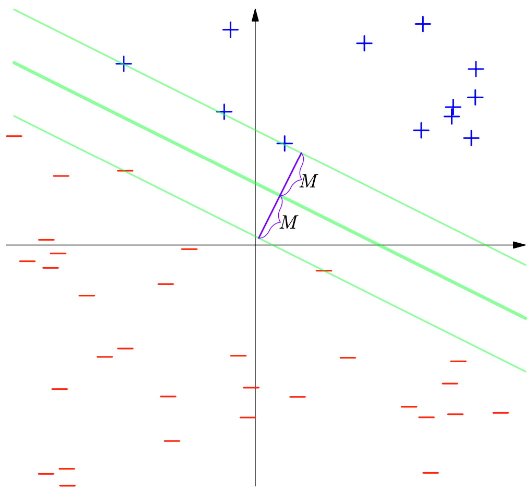
$$g : \mathcal{X} \mapsto \mathcal{Y}, g \in \mathcal{H}$$

Which hyperplane is best?



https://davidrosenberg.github.io/mlcourse/Labs/3-SVM-Notes_sol.pdf

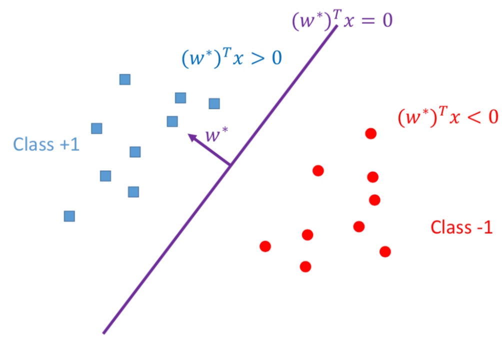
Idea: maximize the margin



https://davidrosenberg.github.io/mlcourse/Labs/3-SVM-Notes_sol.pdf

Support vector machine

- Linear classifier
  - ✓ similar formulation to the perceptron



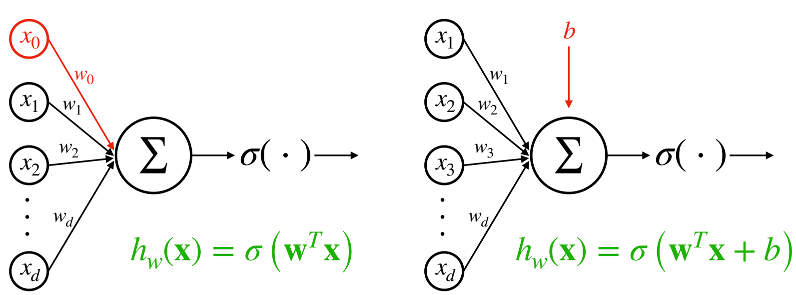
$$h_w(\mathbf{x}) = \sigma\left(\mathbf{w}^T\mathbf{x}\right)$$

$$\sigma(z) = \begin{cases} +1 & \text{if } z \geq 0 \\ -1 & \text{if } z < 0 \end{cases}$$

credit: yingyu liang, cos 495, princeton

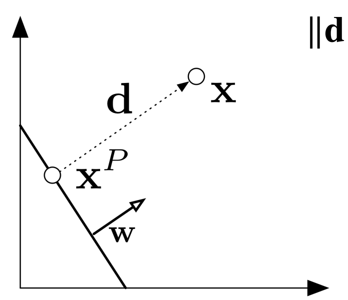
The hyperplane

- A hyperplane H is the set of points such that $H = \{\mathbf{x} \,|\, \mathbf{w}^T\mathbf{x} + b = 0\}$
- Remember from the perceptron:



The margin

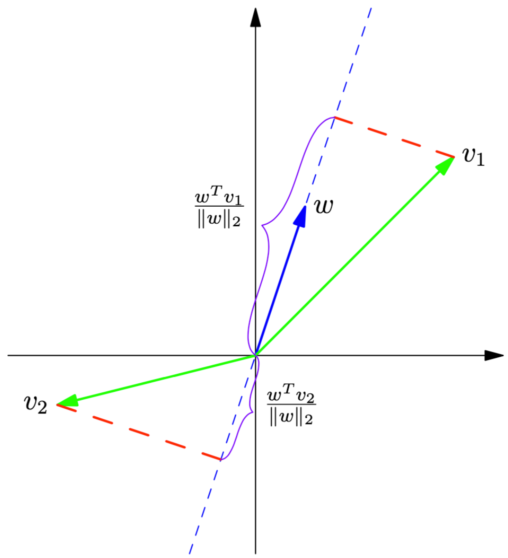
- The margin γ can be defined as the distance from H to the closest point



$$\|\mathbf{d}\|_2 = \frac{|\,\mathbf{w}^T\mathbf{x} + b\,|}{\sqrt{\mathbf{w}^T\mathbf{w}}} = \frac{|\,\mathbf{w}^T\mathbf{x} + b\,|}{\|\mathbf{w}\|_2}$$

$$\gamma = \min_{\mathbf{x}^{(i)} \in \mathcal{D}} \frac{|\,\mathbf{w}^T\mathbf{x}^{(i)} + b\,|}{\|\mathbf{w}\|_2}$$

http://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecturenote09.html



https://davidrosenberg.github.io/mlcourse/Labs/3-SVM-Notes_sol.pdf

Maximizing the margin

- Now we want to maximize the margin and, at the same time, classify all instances correctly

$$\underbrace{\arg\max_{\mathbf{w},b} \gamma}_{\text{maximize margin}} \quad \text{s.t.} \quad \underbrace{\forall i \; y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq 0}_{\text{separating hyperplane}}$$

$$\underbrace{\arg\max_{\mathbf{w},b} \min_{\mathbf{x}^{(i)} \in \mathscr{D}} \frac{|\mathbf{w}^T\mathbf{x}^{(i)} + b|}{\|\mathbf{w}\|_2}}_{\text{maximize margin}} \quad \text{s.t.} \quad \underbrace{\forall i \; y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq 0}_{\text{separating hyperplane}}$$

Scaling w and b

- Previous formulation is not so friendly
- Note that H and γ are **scale invariant**
  - ✓ i.e. rescaling **w** and **b** by the same constant does not change the margin
- We can (conveniently) pick a rescaling constant such that:

$$\min_{\mathbf{x}^{(i)} \in \mathscr{D}} |\mathbf{w}^T\mathbf{x}^{(i)} + b| = 1$$

Simplifying

$$\arg\max_{\mathbf{w},b} \min_{\mathbf{x}^{(i)} \in \mathscr{D}} \frac{|\mathbf{w}^T\mathbf{x}^{(i)} + b|}{\|\mathbf{w}\|_2}$$

$$\arg\max_{\mathbf{w},b} \frac{1}{\|\mathbf{w}\|_2} \min_{\mathbf{x}^{(i)} \in \mathscr{D}} |\mathbf{w}^T\mathbf{x}^{(i)} + b|$$

$$\arg\max_{\mathbf{w},b} \frac{1}{\|\mathbf{w}\|_2} = \arg\min_{\mathbf{w},b} \|\mathbf{w}\|_2 = \arg\min_{\mathbf{w},b} \mathbf{w}^T\mathbf{w}$$

Finally

$$\underset{\mathbf{w},b}{\arg\min} \quad \mathbf{w}^T\mathbf{w}$$

$$\text{s.t.} \quad \forall i \; y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq 0$$

$$\min_{\mathbf{x}^{(i)}\in\mathcal{D}} \; |\mathbf{w}^T\mathbf{x}^{(i)} + b| = 1$$

$$\underset{\mathbf{w},b}{\arg\min} \quad \mathbf{w}^T\mathbf{w}$$

$$\text{s.t.} \quad \forall i \; y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq 1$$

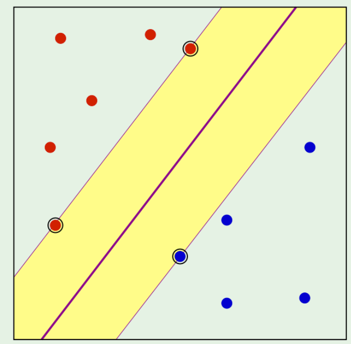
Linear SVM (hard margin)

Quadratic optimization problem (quadratic function with linear constraints)

Unique solution (convex) as long as data is linearly separable

Support vectors? hard-margin

- Margin is defined by special training points, called **support vectors**: $y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) = 1$



http://work.caltech.edu/slides/slides14.pdf

Relaxing the constraints

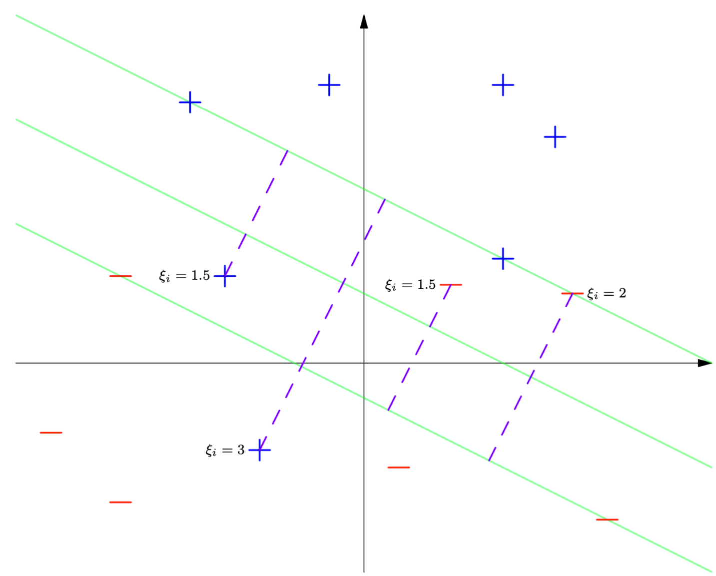
- What if the data is not linearly separable?
  - ✓ we can allow violations to the constraints and introduce some penalties (using slack variables)

hyperparameter controlling the penalty

$$\underset{\mathbf{w},b}{\arg\min} \quad \mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n}\xi_i$$

$$\text{s.t.} \quad \forall i \; y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq 1 - \xi_i$$

$$\forall i \; \xi_i \geq 0$$



https://davidrosenberg.github.io/mlcourse/Labs/3-SVM-Notes_sol.pdf

Removing the constraints

- As the objective function is minimized, for points outside the margin the slack should be zero, and greater than zero otherwise

$$\xi_i = \begin{cases} 1 - y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) & \text{if } y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) < 1 \\ 0 & \text{if } y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq 1 \end{cases}$$

$$\xi_i = \max(1 - y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b), 0)$$

Removing the constraints

$$\begin{aligned} \arg\min_{\mathbf{w},b} \quad & \mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n} \xi_i \\ \text{s.t.} \quad & \forall i \; y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq 1 - \xi_i \\ & \forall i \; \xi_i \geq 0 \end{aligned}$$

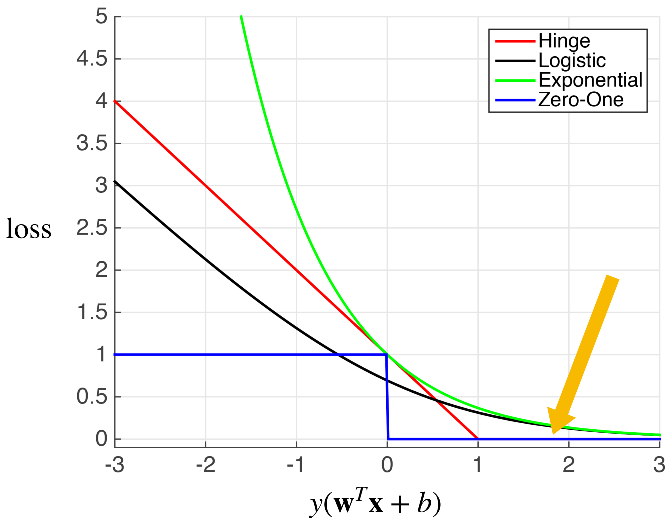
l_2 regularization   hinge loss

$$\arg\min_{\mathbf{w},b} \; \boxed{\mathbf{w}^T\mathbf{w}} + C\sum_{i=1}^{n} \boxed{\max\left(1 - y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b), 0\right)}$$

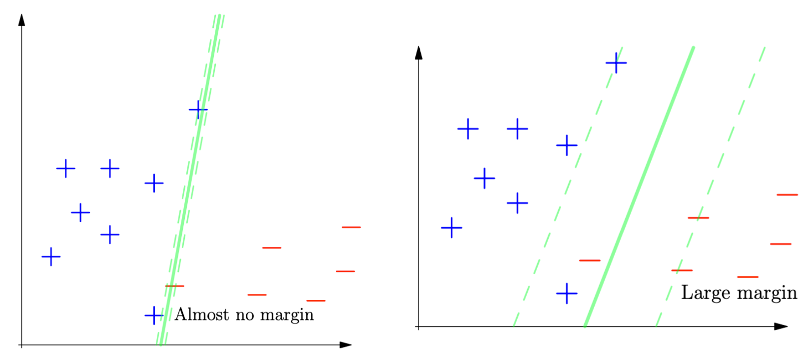
Linear SVM (soft margin)

unconstrained optimization problem

The hinge loss



Impact of C



https://davidrosenberg.github.io/mlcourse/Labs/3-SVM-Notes_sol.pdf

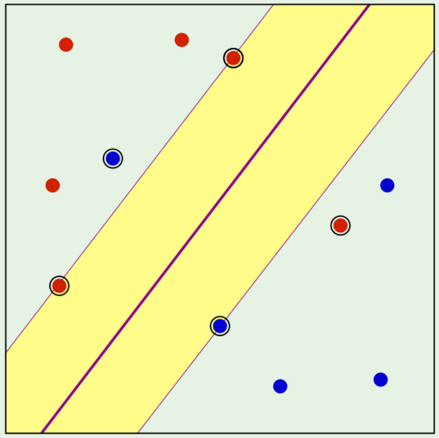
Support vectors? soft-margin

Margin SVs

$y_n(\mathbf{w}^\intercal \mathbf{x}_n + b) = 1 \qquad (\xi_n = 0)$

Non margin SVs

$y_n(\mathbf{w}^\intercal \mathbf{x}_n + b) < 1 \qquad (\xi_n > 0)$

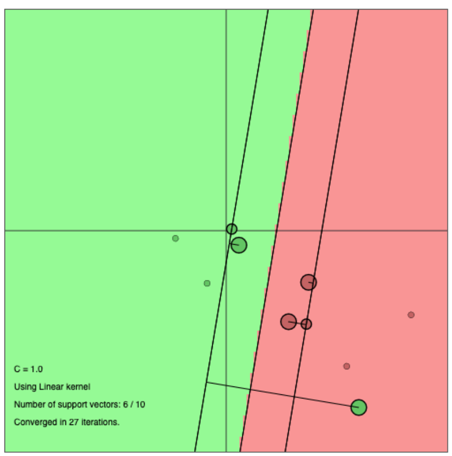


http://work.caltech.edu/slides/slides15.pdf

Solving the optimization

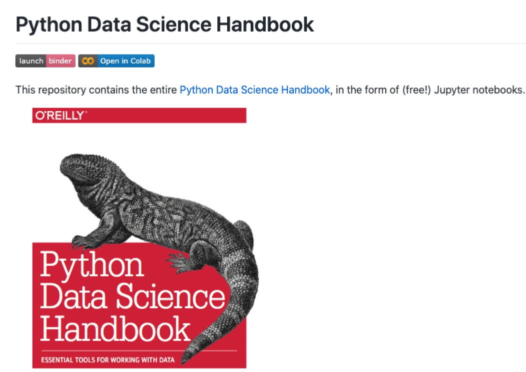
- The soft-margin loss is convex
  - ✓ can use quadratic optimization (very slow)
- Can SGD be applied?
- Yes, but requires calculating the sub-gradient
  - ✓ hinge loss is non-differentiable
  - ✓ **strategy**: if max is 0, then gradient is 0, otherwise calculate the gradient of $1 - y^{(i)}\mathbf{w}^T\mathbf{x}^{(i)}$

Demo



https://cs.stanford.edu/~karpathy/svmjs/demo/

Follow-up example



https://github.com/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/05.07-Support-Vector-Machines.ipynb

Coming up (kernels)