DM825

Introduction to Machine Learning

Lecture 8

Support Vector Machines

Marco Chiarandini

Department of Mathematics & Computer Science
University of Southern Denmark

Overview

Support Vector Machines:

1. Functional and Geometric Margins
2. Optimal Margin Classifier
3. Lagrange Duality
4. Karush Kuhn Tucker Conditions
5. Solving the Optimal Margin
6. Kernels
7. Soft margins
8. SMO Algorithm

In This Lecture

1. Functional and Geometric Margins
2. Optimal Margin Classifier
3. Lagrange Duality
4. Karush Kuhn Tucker Conditions
5. Solving the Optimal Margin

Introduction

- Binary classification.
- $y \in \{-1, 1\}$ (instead of $\{0, 1\}$ like in GLM)
- Let's have $h(\vec{\theta}, \vec{x})$ output values $\{-1, 1\}$:

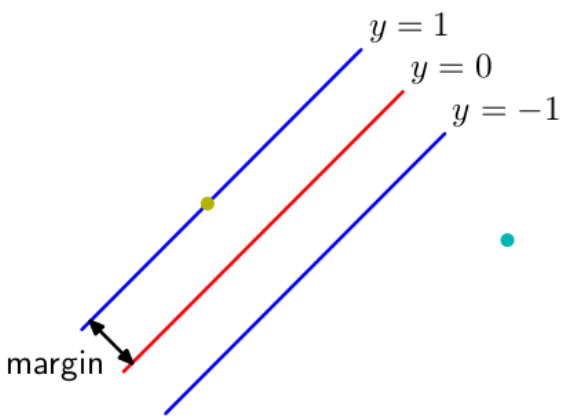
$$f(z) = \text{sign}(z) \begin{cases} 1 & if z \geq 0 \\ -1 & if z < 0 \end{cases}$$

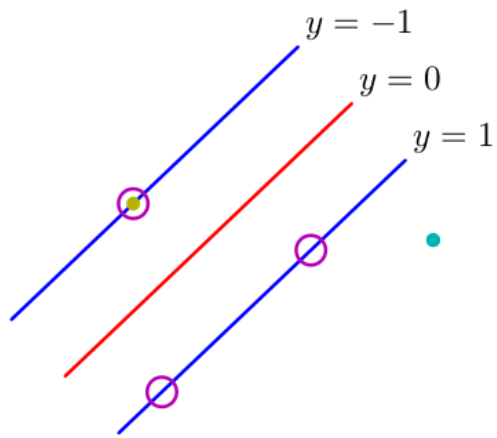
  (hence no probabilities like in logistic regression)

- $h(\vec{\theta}, \vec{x}) = f(\vec{\theta}\vec{x} + \theta_0), \vec{x} \in \mathbb{R}^n, \vec{\theta} \in \mathbb{R}^n, \theta_0 \in \mathbb{R}$
- Assume for now training set is linearly separable

SVM determine model parameters by solving a convex optimization problem and hence a local optimal solution is also global optimal.

Margin: smallest distance between the decision boundary and any of the samples.





The location of the boundary is determined by a subset of the data points, known as support vectors

Outline

Resume

- functional margin:

$$\hat{\gamma}^i = y^i(\vec{\theta}^T \vec{x}^i + \theta_0) \qquad \Longrightarrow \hat{\gamma} = \min_i \hat{\gamma}^i$$

  requires a normalization condition

- geometric margin:

$$\gamma^i = y^i \left(\frac{\vec{\theta}}{\| \vec{\theta} \|}^T x^i + \frac{\theta_0}{\| \vec{\theta} \|} \right) \qquad \Longrightarrow \gamma = \min_i \gamma^i$$

  scale invariant

- $\gamma = \frac{\hat{\gamma}}{\|\vec{\theta}\|}$
- if $\| \vec{\theta} \| = 1$ then $\hat{\gamma}^i = \gamma^i$ the two margins correspond

Outline

Optimization Problem

Looking at the geometric margin:

$$\text{(OPT1)} : \max_{\gamma,\vec{\theta},\theta_0} \quad \gamma$$
$$\gamma \leq y^i(\vec{\theta}^T\vec{x}^i + \theta_0) \quad \forall i = 1,\ldots,m$$
$$\| \vec{\theta} \| = 1$$

Alternatively, looking at functional margins and recalling that $\gamma = \frac{\hat{\gamma}}{\|\vec{\theta}\|}$:

$$\text{(OPT2)} : \max_{\hat{\gamma},\vec{\theta},\theta_0} \quad \frac{\hat{\gamma}}{\| \vec{\theta} \|}$$
$$\hat{\gamma} \leq y^i(\vec{\theta}^T\vec{x}^i + \theta_0) \quad \forall i = 1,\ldots,m$$

For the functional margins we can fix the scale, for the geometric margin no scaling problem. Then we can arbitrary fix $\hat{\gamma} = 1$

$$\text{(OPT3)} : \min_{\vec{\theta},\theta_0} \quad \frac{1}{2} \parallel \vec{\theta} \parallel^2$$
$$1 \leq y^i(\vec{\theta}^T \vec{x}^i + \theta_0) \quad \forall i = 1, \ldots, m$$

where we used that:

$$\max 1/\parallel \vec{\theta} \parallel = \min \parallel \vec{\theta} \parallel$$

and removed the square root because monotonous in $\parallel \vec{\theta} \parallel = \sqrt{\vec{\theta}^T \vec{\theta}}$.
This problem is a convex optimization problem, it has convex quadratic objective function and linear constraints, hence it can be solved optimally and efficiently

Convex optimization problem

$$\begin{aligned} &\text{minimize } f_0(x) \\ &\text{subject to } f_i(x) \leq b_i, i = 1, \ldots, m \end{aligned}$$

objective and constraint functions are convex:

$$f_i(\alpha x + \beta y) \leq \alpha f_i(x) + \beta f_i(y)$$

if $\alpha + \beta = 1, \alpha \geq 0, \beta \geq 0$

Outline

Lagrangian

standard form problem (not necessarily convex)

$$\begin{aligned} \text{minimize } & f_0(x) \\ \text{subject to } & f_i(x) \leq 0, i = 1, ..., m \\ & h_i(x) = 0, i = 1, ..., p \end{aligned}$$

variable $x \in \mathbb{R}^n$, domain $\mathcal{D}$, optimal value p^*

Lagrangian: $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$, with $\text{dom } \mathcal{L} = \mathcal{D} \times \mathbb{R}^m \times \mathbb{R}^p$,

$$\mathcal{L}(x, \alpha, \beta) = f_0(x) + \sum_{i=1}^{m} \alpha_i f_i(x) + \sum_{i=1}^{p} \beta h_i(x)$$

- weighted sum of objective and constraint functions
- α_i is Lagrange multiplier associated with $f_i(x) \leq 0$
- β_i is Lagrange multiplier associated with $h_i(x) = 0$
- $\vec{\alpha}$ and $\vec{\beta}$ are dual or Lagrangian variables

Lagrange dual function

Lagrange dual function: $\mathcal{L}_D : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$

$$\mathcal{L}_D(\alpha, \beta) = \min_{x \in \mathcal{D}} \mathcal{L}(x, \alpha, \beta) = \min_{x \in \mathcal{D}} \left(f_0(x) + \sum_{i=1}^{m} \alpha_i f_i(x) + \sum_{i=1}^{p} \beta h_i(x) \right)$$

$\mathcal{L}_D$ is concave, can be $-\infty$ for some α and β

Lower bound property: for a feasible $\tilde{x}$

1. $\forall \alpha \geq 0, \beta \quad \mathcal{L}_D(\alpha, \beta) \leq p^*$
2. $\mathcal{L}_P(x) = \max_{\alpha \geq 0, \beta} (\mathcal{L}_D(\alpha, \beta)) \leq p^*$ (best lower bound, it may be $= p^*$)

Proof of (1): for any $\tilde{x}$ feasible and $\alpha \geq 0$:

$$\mathcal{L}(\tilde{x}, \alpha, \beta) = f_0(\tilde{x}) + \sum_{i=1}^{m} \alpha_i f_i(\tilde{x}) + \sum_{i=1}^{p} \beta h_i(\tilde{x}) \leq f_0(\tilde{x})$$

hence

$$\mathcal{L}_D(\alpha, \beta) = \min_{x \in \mathcal{D}} \mathcal{L}(x, \alpha, \beta) \leq \mathcal{L}(\tilde{x}, \alpha, \beta) \leq f_0(\tilde{x})$$

(2) is true because (1) true for any α, β.

If f_0 and g_i are convex and h_i affine,

$$d^* = \max_{\alpha \geq 0, \beta} (\mathcal{L}_D(\alpha, \beta)) = p^*$$

so we can solve the dual in place of the primal.

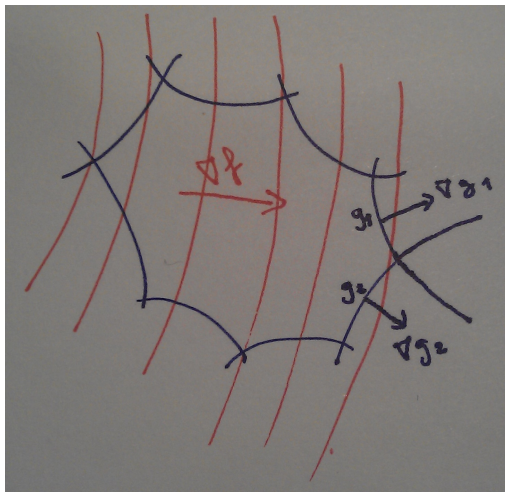
Outline

Karush Kuhn Tucker Conditions

standard form problem
(not necessarily convex)

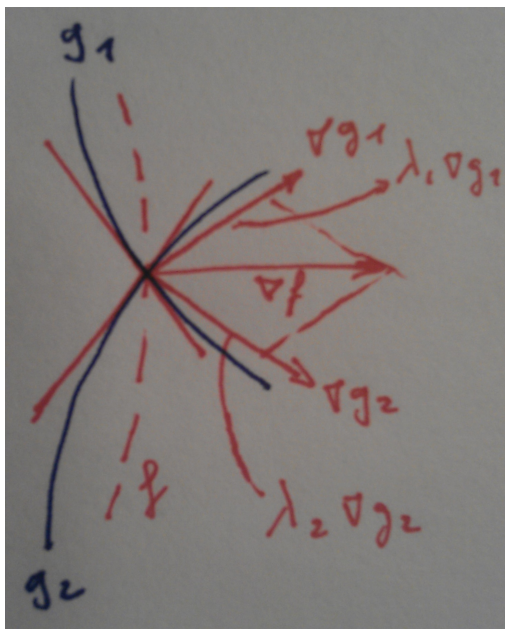
$$\text{minimize } f_0(x)$$
$$\text{subject to } g_i(x) \leq b_i, i = 1, ..., m$$

variable $x \in \mathbb{R}^n$, f, g nonlinear, $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}^m$



Necessary conditions for optimality (local validity):

$$\begin{cases} \nabla f(x_0) = \sum_{i=1}^{m} \lambda_i \nabla g_i(x_0) \\ \lambda_i \geq 0 \forall i \\ \sum_{i=1}^{m} \lambda_i (g_i(x_0) - b_i) = 0 \\ g_i(x_0) - b_i \leq 0 \end{cases}$$



Outline

Let's go back to our problem:

$$\text{(OPT3)} : \min_{\vec{\theta},\theta_0} \quad \frac{1}{2} \parallel \vec{\theta} \parallel^2$$

$$1 \leq y^i(\vec{\theta}^T \vec{x}^i + \theta_0) \qquad \forall i = 1, \ldots, m$$

$$\mathcal{L}(\vec{\theta}, \theta_0, \vec{\alpha}) = \frac{1}{2} \parallel \vec{\theta} \parallel^2 - \sum_{i=1}^{m} \alpha_i \left(y^i(\vec{\theta}^T \vec{x}^i + \theta_0) - 1 \right)$$

we find the dual form by solving in $\vec{\theta}, \theta_0$

$$\mathcal{L}_D(\vec{\alpha}) = \min_{\vec{\theta},\theta_0} \mathcal{L}(\vec{\theta}, \theta_0, \vec{\alpha})$$

$$\nabla_{\vec{\theta}} \mathcal{L}(\vec{\theta}, \theta_0, \vec{\alpha}) = \vec{\theta} - \sum_{i=1}^{m} \alpha_i y^i \vec{x}^i = 0 \qquad \Longrightarrow \boxed{\vec{\theta} = \sum_{i=1}^{m} \alpha_i y^i \vec{x}^i}$$

$$\frac{\partial \mathcal{L}(\vec{\theta}, \theta_0, \vec{\alpha})}{\partial \theta_0} = - \sum_{i=1}^{m} \alpha_i y^i \alpha^i = 0 \qquad \Longrightarrow \boxed{\sum_{i=1}^{m} \alpha_i y^i \alpha^i = 0}$$

Substituting in $\mathcal{L}(\vec{\theta}, \theta_0, \vec{\alpha})$:

$$\begin{aligned}
\mathcal{L}_D(\vec{\theta}) &= \frac{1}{2}\left(\sum_{i=1}^{m} \alpha_i y^i \vec{x}^i\right)\left(\sum_{j=1}^{m} \alpha_j y^j \vec{x}^j\right) \\
&\quad - \sum_{i=1}^{m} \alpha_i \left(y^i \left((\sum_{j=1}^{m} \alpha_j y^j \vec{x}^j)\vec{x}^i + \theta_0\right) - 1\right) \\
&= \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} y^i y^j \alpha_i \alpha_j \langle \vec{x}^i \vec{x}^j \rangle
\end{aligned}$$

We are left with the dual problem:

$$\max_{\vec{\alpha}} \; W(\vec{\alpha}) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} y^i y^j \alpha_i \alpha_j \langle \vec{x}^i \vec{x}^j \rangle$$

$$\text{s.t.} \quad \alpha_i \geq 0 \qquad \forall i = 1 \dots m$$

$$\sum_{i=1}^{m} \alpha_i y^i = 0$$

- This problem is in m variables. Problem (OPT3) has D variables and quandratic programming can be solved in $O(D^3)$. If $D \ll n$ then it seems we did not earned a lot
- the form above allows us to use kernel trick and have even infinite dimensions ($D \gg m$)
- the use of the kernel and its constraint of being positive semidefinite ensures that the problem is bounded from below.

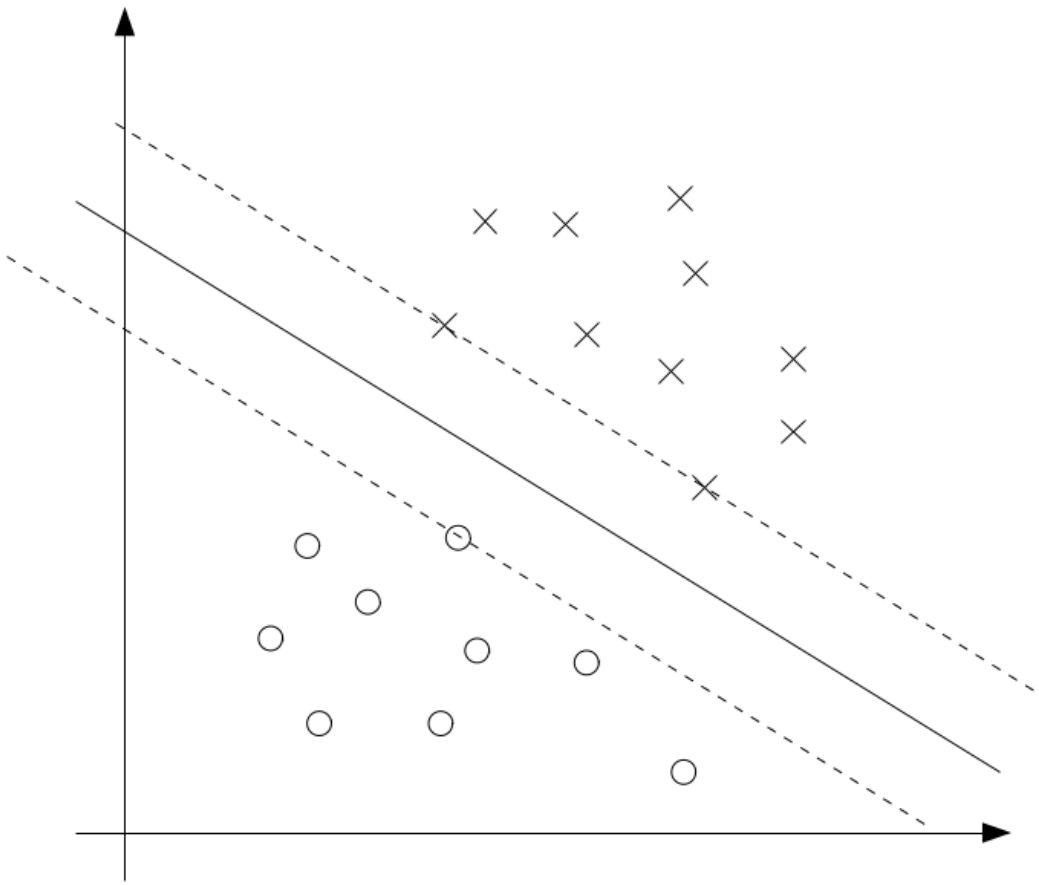
In addition, an optimal solution satisfies the KKT conditions on (OPT3):

$$y_i(\vec{\theta}^T \vec{x}^i + \theta_0) \geq 1$$
$$\alpha_i[y_i(\vec{\theta}^T \vec{x}^i + \theta_0) - 1] = 0 \quad \forall i$$

From these we can see that

- if $\alpha_i > 0$, then $y_i(\vec{\theta}^T \vec{x}^i + \theta_0) = 1$ ($\vec{x}^i$ is on the boundary)
- if $y_i(\vec{\theta}^T \vec{x}^i + \theta_0) > 1$, $\vec{x}^i$ is not on the boundary and $\alpha_i = 0$

Points where $y_i(\vec{\theta}^T \vec{x}^i + \theta_0) > 1$ are the support vectors:



Prediction

For a new pint $\vec{x}$ predict by:

$$\begin{aligned} h(\vec{\theta}, \vec{x}) &= f(\vec{\theta}\vec{x} + \theta_0) = \text{sign}(\vec{\theta}\vec{x} + \theta_0) \\ &= \text{sign}\left(\left(\sum_{i=1}^{m} \alpha_i y^i \vec{x^i}\right)^T \vec{x} + \theta_0\right) \\ &= \text{sign}\left(\sum_{i=1}^{m} \alpha_i y^i \langle \vec{x^i}, \vec{x} \rangle + \theta_0\right) \end{aligned}$$

For the KKT conditions, most trainig data can be discarded after training and only the points that are support vectors need to be retained for this computation

Intercept

We can derive θ_0 by:

$$\theta_0 = -\frac{\max_{i:y^i=-1} \vec{\theta}^T \vec{x}^i + \min_{i:y^i=1} \vec{\theta}^T \vec{x}^i}{2}$$