MACHINE LEARNING
(16) Support vector machines
• Introduction

• Geometry of linear discriminant functions

• Support vector machine training

• Foundations of constrained optimization

• Quadratic program formulations of SVM training

• Kernelizing the maximum margin SVM
• Introduction

• Geometry of linear discriminant functions

• Support vector machine training

• Foundations of constrained optimization

• Quadratic program formulations of SVM training

• Kernelizing the maximum margin SVM
Bibliographic remarks

Introductions to support vector classification can be found in all standard machine learning textbooks such as Hastie et al. (2001), Bishop (2006), ?, Barber (2012), Duda et al. (2012), and Murphy (2012). The primary literature includes Vapnik (1982), Boser et al. (1992), and Cortes and Vapnik (1995). An introduction to support vector classification from the perspective of kernel functions is provided by Scholkopf and Smola (2001).
Introduction

Fundamental question

• Given a labeled training data, what is a good classification boundary?

Fundamental answer

• A classification boundary with a large margin is good.
Vladimir N. Vapnik was born in Russia and received the Ph.D. degree in statistics from the Institute of Control Sciences, Academy of Science of the USSR, Moscow, Russia, in 1964.

Since 1991, he has been working for AT&T Bell Laboratories (since 1996, AT&T Labs Research), Red Bank, NJ. His research interests include statistical learning theory, theoretical and applied statistics, theory and methods for solving stochastic ill-posed problems, and methods of multidimensional function approximation. His main results in the last three years are related to the development of the support vector method. He is author of many publications, including seven monographs on various problems of statistical learning theory.
Introduction

Vapnik’s Statistical Learning Theory

- A mathematical framework for finding good predictive functions based on data
- Predictive performance is valued higher than correctness of the approximation
- SVMs were developed in the context of statistical learning theory

Historical overview

- 1964: Chervonenkis & Vapnik develop maximum margin classification.
- 1992: Boser, Guyon, & Vapnik suggest using kernel functions
- 1995: Cortex & Vapnik suggest soft margin classification
- Late 1990’s: SVM toolboxes, websites, and summer schools
- 2000’s: SVMs hype in the machine learning community
- Since 2010: Neural networks hype in the machine learning community
Introduction

What we will cover:

• Geometric foundations of maximum and soft margin classification
• SVM training as a constrained quadratic programming problem
• Motivation for kernel methods

What we will not cover:

• Statistical learning theory
• Optimality questions of maximum margin classification
• Multiclass classification
• Kernel methods
Teaching aim

• Training and testing an SVM using a Python quadratic programming routine.

*Programming exercise*

1. Create a training set by sampling from two Gaussian distributions.
2. Train a maximum margin SVM using the training set and cvxopt.solvers qp.
3. Test the generalization accuracy of the trained SVM.

cvxopt is available from cvxopt.org
• Introduction

• Geometry of linear discriminant functions

• Support vector machine training

• Foundations of constrained optimization

• Quadratic program formulations of SVM training

• Kernelizing the maximum margin SVM
Definition (Training set)

A *training set*

\[ D := \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(n)}, y^{(n)})\} \]  \hspace{1cm} (1)

is a set of *training examples*

\[(x^{(i)}, y^{(i)}) \text{ with } x^{(i)} \in \mathbb{R}^m, y^{(i)} \in \{-1, 1\}, i = 1, \ldots, n, \]  \hspace{1cm} (2)

where \(x^{(i)}\) is referred to as *m-dimensional feature vector* and \(y^{(i)}\) is referred to as *target variable*.

Remarks

- \(y^{(i)} \in \{-1, 1\}\) signifies the class membership of \(x^{(i)} \in \mathbb{R}^m\).
- \(x^{(i)} \in \mathbb{R}^m\) may denote the values of \(m\) clinical markers of the \(i\)th of \(n\) patients in one of two diagnoses groups \(y^{(i)}\) denoted by \(-1\) and \(1\), respectively.
Geometry of linear discriminant functions

Definition (Linear discriminant function)

A linear discriminant function is a multivariate real-valued function of the form

\[ h : \mathbb{R}^m \rightarrow \{0, 1\}, x \mapsto h(x) := g(f(x)), \]  

(3)

where

- \( f \) is a multivariate real-valued, parameter-dependent affine function

\[ f : \mathbb{R}^m \rightarrow \mathbb{R}, x \mapsto f(x) := w^T x + w_0, \]  

(4)

where \( w \in \mathbb{R}^m \) is a parameter vector and \( w_0 \in \mathbb{R} \) is a bias parameter,

- and \( g \) is the univariate real-valued, parameter-independent classification function

\[ g : \mathbb{R} \rightarrow \{0, 1\}, f(x) \mapsto g(f(x)) := \begin{cases} -1, & f(x) < 0 \\ +1, & f(x) \geq 0 \end{cases} \]  

(5)

In the feature space \( \mathbb{R}^m \), a linear discriminant function induces

- a decision boundary \( H := \{ x \in \mathbb{R}^m | f(x) = 0 \} \), referred to as hyperplane,

- a decision region \( D_{-1} := \{ x \in \mathbb{R}^m | f(x) < 0 \} \), and

- a decision region \( D_{+1} := \{ x \in \mathbb{R}^m | f(x) \geq 0 \} \).
Geometry of linear discriminant functions

Line equation for hyperplanes in $\mathbb{R}^2$

$$f(x) = 0 \Leftrightarrow w^T x + w_0 = 0 \Leftrightarrow w_1 x_1 + w_2 x_2 + w_0 = 0 \Leftrightarrow x_2 = -\frac{w_1}{w_2} x_1 - \frac{w_0}{w_2} \quad (6)$$
Theorem (Geometry of linear discriminant functions)

Let
\[ f : \mathbb{R}^m \to \mathbb{R}, x \mapsto f(x) := w^T x + w_0, \]
\[ H := \{ x \in \mathbb{R}^m | f(x) = 0 \} \subset \mathbb{R}^m \]
denote a multivariate real-valued, parameter-dependent affine function and let denote a hyperplane. Then the following geometric relationships hold

1. \( w \) is orthogonal to any vector pointing into the direction of \( H \).
2. The minimal Euclidean distance \( d \) between \( x \in \mathbb{R}^m \) of and a point on \( H \) is
   \[ d = \frac{1}{||w||_2} f(x). \]
3. The minimal Euclidean distance \( d_0 \) between the origin and a point on \( H \) is
   \[ d_0 = \frac{w_0}{||w||_2}. \]
Geometry of linear discriminant functions

Geometry of linear discriminant functions

\[ w \]
\[ \|w\|_2 \]
\[ x_2 \]
\[ x_1 \]
\[ w_0/\|w\|_2 \]
\[ x_p \]
\[ d \frac{w}{\|w\|_2} \]
\[ d := \frac{1}{\|w\|_2} f(x) \]
\[ H := \{ x \in \mathbb{R}^2 | f(x) = 0 \} \]
Proof

Proof of (1)

To see that (1) holds, let $x_a \in H_w$ and $x_b \in H_w$ be two arbitrary points on the hyperplane. Then the following system of affine-linear equations holds by definition of the hyperplane:

$$w^T x_a + w_0 = 0 \quad (11)$$
$$w^T x_b + w_0 = 0 \quad (12)$$

Subtracting (12) from (11) yields

$$w^T x_a - w^T x_b = 0 \Leftrightarrow w^T (x_a - x_b) = 0 \quad (13)$$

Thus the weight vector is orthogonal to the vector $y := (x_a - x_b)$, which points in the direction of the hyperplane.

□
Proof

Proof of (2)

Consider the decomposition of a point \( x \in \mathbb{R}^m \) into its orthogonal projection onto a hyperplane \( x_p \in \mathbb{R}^m \) and its distance from the hyperplane \( d \frac{w}{\|w\|_2} \)

\[
x = x_p + d \frac{w}{\|w\|_2}.
\] (14)

Note that this decomposition is possible, because \( w \) is orthogonal to any vector pointing in the direction of the hyperplane and \( \frac{\|w\|_2}{\|w\|_2} = 1 \). Consider now the evaluation of the thus decomposed \( x \) under the linear discriminant function:

\[
f(x) = w^T x + w_0 = w^T \left( x_p + d \frac{w}{\|w\|_2} \right) + w_0 = w^T x_p + w_0 + d \frac{w^T w}{\|w\|_2}.
\] (15)

Then, because \( x_p \in H_w \) and hence \( w^T x_p + w_0 = 0 \), we have

\[
f(x) = d \frac{w^T w}{\|w\|_2} = d \frac{\|w\|_2^2}{\|w\|_2} = d \|w\|_2
\] (16)

and hence

\[
d = \frac{1}{\|w\|_2} f(x).
\] (17)
Proof

*Proof of (3)*

For the minimal distance of the origin \( x_0 = (0, ..., 0)^T \in \mathbb{R}^m \) to points on the hyperplane, we have

\[
d_0 = \frac{1}{\|w\|_2} f(x_0) = \frac{1}{\|w\|_2} (w^T x_0 + w_0) = \frac{1}{\|w\|_2} w^T \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} + \frac{w_0}{\|w\|_2} = \frac{w_0}{\|w\|_2}. \tag{18}
\]
Geometry of linear discriminant functions

### Definition (Hyperplane margin, support vector)

Let \( D \) denote a training set, let \( f \) denote a multivariate real-valued affine function, and let \( H \) denote a hyperplane induced by \( f \). Furthermore, let

\[
|d(i)| := \left| \frac{1}{\|w\|_2} f\left(x(i)\right) \right| = \frac{y(i)}{\|w\|_2} f\left(x(i)\right) = \frac{y(i)(w^T x(i) + w_0)}{\|w\|_2} \geq 0 \quad (19)
\]

denote the absolute minimal Euclidean distance of the training feature vector \( x(i) \), \( i = 1, ..., n \) from \( H \). Then the margin \( d^* \) of \( H \) with respect to \( D \) is defined as the minimum of the absolute minimal Euclidean distances of training feature vectors to the hyperplane,

\[
d^* := \min_{i=1,\ldots,n} \left\{ |d(i)| \right\} = \min_{i=1,\ldots,n} \left\{ \frac{y(i)(w^T x(i) + w_0)}{\|w\|_2} \right\} \quad (20)
\]

A training set feature vector \( x(i) \) is referred to as a support vector, if \( |d(i)| = d^* \), i.e., if \( x(i) \) is located on the margin of the hyperplane.
Geometry of linear discriminant functions
Definition (Equivalent hyperplanes, canonical hyperplane)

Let $f$ denote a multivariate real-valued affine function and let

$$H := \{ x \in \mathbb{R}^m | f(x) = 0 \}$$  \hspace{1cm} (21)

denote a hyperplane induced by $f$. Then all scalar multiples of $f$ (and thus of $w$ and $w_0$) induce the identical hyperplane, because from $f(x) = 0$ it follows that $af(x) = 0$ for $a \in \mathbb{R} \setminus \{0\}$. The hyperplanes

$$H_a := \{ x \in \mathbb{R}^m | af(x) = 0, a \in \mathbb{R} \setminus \{0\} \}$$  \hspace{1cm} (22)

are referred to as **equivalent hyperplanes** to $H$. Given a support vector $x^*$ and a set of equivalent hyperplanes (and thus a set of $w$ and $w_0$ inducing equivalent hyperplanes), the **canonical hyperplane** is defined as that hyperplane (and thus the specific $w$ and $w_0$) for which

$$|f(x^*)| = y^* (w^T x^* + w_0) = 1.$$  \hspace{1cm} (23)

From the definition of the canonical hyperplane, it follows immediately that the margin of the canonical hyperplane is given by

$$d^* = \frac{1}{\|w\|_2}.$$  \hspace{1cm} (24)
• Introduction

• Geometry of linear discriminant functions

• Support vector machine training

• Foundations of constrained optimization

• Quadratic program formulations of SVM training

• Kernelizing the maximum margin SVM
Definition (Linearly separable and non-linear separable training set)

A training set $\mathcal{D}$ is called *linearly separable* training set, if a linear discriminant function can be found such that all training data points are classified correctly. A training set $\mathcal{D}$ *non-linearly separable* training set, if no linear discriminant function can be found such that all training data point are classified correctly.
Support vector machine training

**Definition (Maximum margin classification in the linearly separable case)**

Let $\mathcal{D}$ denote a linearly separable training set. Then support vector machine training for maximum margin classification corresponds to the constrained optimization problem

$$\min_{w, w_0} \frac{1}{2} \|w\|^2_2 \quad \text{subject to } y^{(i)}(w^T x^{(i)} + w_0) \geq 1 \text{ for } i = 1, \ldots, n$$

(25)

Here,

- the objective

  $$\min_{w, w_0} \frac{1}{2} \|w\|^2_2 \Leftrightarrow \max_{w, w_0} \frac{1}{\|w\|_2}$$

  (26)

  corresponds to maximizing the hyperplane margin, and

- the constraints

  $$y^{(i)}(w^T x^{(i)} + w_0) \geq 1 \text{ for } i = 1, \ldots, n$$

  (27)

  correspond to the aim of all feature vectors to either be support vectors or to lie on the correct side of the hyperplane margin.
Support vector machine training

Maximum margin classification in the linearly separable case

\[
\min_{w, w_0} \frac{1}{2} \|w\|_2^2 \text{ subject to } y^{(i)} (w^T x^{(i)} + w_0) \geq 1
\]
Support vector machine training

Definition (Soft margin classification)

Let $\mathcal{D}$ denote a not necessarily linearly separable training set. Then support vector machine training for soft margin classification corresponds to the constrained optimization problem

$$
\min_{w, w_0, \xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i^k \quad \text{subject to } y^{(i)}(w^T x^{(i)} + w_0) \geq 1 - \xi_i, \xi \geq 0 \quad (28)
$$

where $\xi := (\xi_1, ..., x_n)$ is a vector of slack variables $\xi_i, 1, ..., n$, the term $\sum_{i=1}^{n} \xi_i^k$ is referred to as loss, $k \in \mathbb{N}$ is a constant determining the particular kind of loss (e.g., hinge loss for $k = 1$, quadratic loss for $k = 2$), and $C \in \mathbb{R}$ is an empirically chosen constant.

Here,

• the objective corresponds to maximizing the hyperplane margin while also minimizing the loss, where the relative weighting of each objective is given by $C$,

• the constraints correspond to (1) correct classification and margin maximization for $\xi_i = 0$, (2) correct classification for $0 < \xi < 1$, and (3) misclassification for $\xi > 1$ of training set feature vectors.
Support vector machine training

Soft margin classification in a non-linearly separable case

\[
\begin{align*}
    & y^{(i)}(w^T x^{(i)} + w_0) \geq 1 \\
    \downarrow & \\
    & y^{(i)}(w^T x^{(i)} + w_0) \geq 1 - \xi_i
\end{align*}
\]

\[
\min_{w,w_0,\xi} \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^{n} \xi_i 
\text{subject to } y^{(i)}(w^T x^{(i)} + w_0) \geq 1 - \xi_i, \xi \geq 0
\]
• Introduction

• Geometry of linear discriminant functions

• Support vector machine training

• Foundations of constrained optimization

• Quadratic program formulations of SVM training

• Kernelizing the maximum margin SVM
Foundations of constrained optimization

Definition (Constrained optimization problem)

A *constrained optimization problem* is of the general form

\[
\min_{x \in \mathbb{R}^n} f(x) \text{ subject to } c_i(x) = 0, i \in E, c_i(x) \geq 0, i \in I,
\]

(29)

where \( f : \mathbb{R}^n \to \mathbb{R} \) and \( c_i : \mathbb{R}^n \to \mathbb{R}, i \in E \cup I \) are twice-differentiable multivariate real-valued functions and \( E, I \) are finite index sets. \( f \) is referred to as *objective function*, \( c_i, i \in E \) are *equality constraints*, and \( c_i, i \in I \) are *inequality constraints*. The set

\[
\mathcal{X} := \{x \in \mathbb{R}^n | c_i(x) = 0, i \in E \text{ and } c_i(x) \geq 0, i \in I\}
\]

(30)

is referred to as *feasible set*.

Remarks

- Necessary conditions for unconstrained minimizers \( x^* \) are
  - for \( n = 1 \): \( f'(x^*) = 0 \) and \( f''(x^*) > 0 \).
  - for \( n > 1 \): \( \nabla f(x^*) \) and \( H^f(x^*) \) positive-semidefinite.

- Here, we will introduce analogue conditions for constrained minimizers.
Example (Quadratic program)

A *quadratic program* is a constrained convex optimization problem of the form

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} x^T P x + q^T x \text{ subject to } A x = b, \text{ and } -G x + h \geq 0 \text{ and }$$

where $P \in \mathbb{R}^{n \times n}$ is a strictly positive definite matrix, $x, q \in \mathbb{R}^n$, $G \in \mathbb{R}^{m \times n}$, $h \in \mathbb{R}^m$, $A \in \mathbb{R}^{p \times n}$, and $b \in \mathbb{R}^p$.

Remarks

- Standard optimization libraries contain quadratic programming solvers (cvxopt.org)
- The quadratic programming formulation of SVM training motivated *kernel methods*.
Definition (Local, strict local, and isolated local solutions)

A vector $x^*$ is

- a *local solution* of the constrained optimization problem, if $x^* \in \mathcal{X}$ and if there is a neighborhood $\mathcal{N}$ of $x^*$, such that $f(x) \geq f(x^*)$ for $x \in \mathcal{N} \cap \mathcal{X}$,

- a *strict local solution* of the constrained optimization problem, if $x^* \in \mathcal{X}$ and if there is a neighborhood $\mathcal{N}$ of $x^*$, such that $f(x) > f(x^*)$ for $x \in \mathcal{N} \cap \mathcal{X}$, and

- is an *isolated local solution* of the constrained optimization problem, if $x^* \in \mathcal{X}$ and if there is a neighborhood $\mathcal{N}$ of $x^*$, such that $x^*$ is the only local solution in $\mathcal{N} \cap \mathcal{X}$.

Remarks

- Isolated local solutions are strict local solutions but not vice versa.
Foundations of constrained optimization

Definition (Lagrangian function)

Let

\[
\min_{x \in \mathbb{R}^n} f(x) \text{ subject to } c_i(x) = 0, i \in E, c_i(x) \geq 0, i \in I,
\]

(32)
denote a constrained optimization problem. Then the Lagrangian function of
this problem is defined as

\[
L : \mathbb{R}^n \times \mathbb{R}^{|E \cup I|} \rightarrow \mathbb{R}, (x, \lambda) \mapsto L(x, \lambda) := f(x) - \sum_{i \in E \cup I} \lambda_i c_i(x)
\]

(33)
Definition (First-order necessary conditions)

Suppose that $x^*$ is a local solution of the constrained optimization problem

$$\min_{x \in \mathbb{R}^n} f(x) \text{ subject to } c_i(x) = 0, i \in E, c_i(x) \geq 0, i \in I,$$  \hfill (34)

Then there is a Lagrange multiplier vector $\lambda^*$ with components $\lambda^*_i, i \in E \cup I$, such that the following conditions are satisfied at $(x^*, \lambda^*)$

$$\nabla_x L(x^*, \lambda^*) = 0 \hfill (35)$$

$$c_i(x^*) = 0 \text{ for all } i \in E \hfill (36)$$

$$c_i(x^*) \geq 0 \text{ for all } i \in I \hfill (37)$$

$$\lambda^*_i \geq 0 \text{ for all } i \in I \hfill (38)$$

$$\lambda^*_i c_i(x^*) = 0 \text{ for all } i \in E \cup I \hfill (39)$$

Remarks

- The conditions only hold under certain regularity conditions omitted here.
- The conditions are also known as Karush-Kuhn-Tucker (KKT) conditions.
- A proof is given in Nocedal and Wright (2006, Section 12.4).
- Note that the last condition implies $\lambda^*_i > 0 \Rightarrow c_i(x^*) = 0$. 
Definition (The dual problem)

Consider the constrained nonlinear optimization problem

\[
\min_{x \in \mathbb{R}^n} f(x) \text{ subject to } c(x) \geq 0,
\]  

where there are no equality constraints, \( c(x) := (c_1(x), c_2(x), ..., c_m(x))^T \) denotes the vector (function) of inequality constraints, and for which the Lagrangian function with Lagrange multiplier vector \( \lambda \in \mathbb{R}^m \) is given by

\[
L : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}, (x, \lambda) \mapsto L(x, \lambda) := f(x) - \lambda^T c(x).
\]

Then the dual objective function (also known as dual Lagrangian function) is defined as

\[
q : \mathbb{R}^m \to \mathbb{R}, \lambda \mapsto q(\lambda) := \inf_x L(x, \lambda)
\]

and the dual problem is defined as

\[
\max_{\lambda \in \mathbb{R}^m} q(\lambda) \text{ subject to } \lambda \geq 0.
\]
Foundations of constrained optimization

**Theorem (Weak duality)**

For any $\bar{x}$ feasible for

$$\min_{x \in \mathbb{R}^n} f(x) \text{ subject to } c(x) \geq 0,$$

and any $\bar{\lambda} \geq 0$, it holds that

$$q(\bar{\lambda}) \leq f(\bar{x}) \quad (45)$$

**Proof**

With the definition of $q$, $\bar{\lambda} \geq 0$, and $c(\bar{x}) \geq 0$, we have

$$q(\bar{\lambda}) = \inf_x f(x) - \bar{\lambda}^T c(x) \leq f(\bar{x}) - \bar{\lambda}^T c(\bar{x}) \leq f(\bar{x}) \quad (46)$$

**Remarks**

- The theorem states that the optimal value of the dual problem is a lower bound on the optimal objective value for the primal problem.
Theorem (Strong duality I)

Consider the constrained nonlinear optimization problem

$$\min_{x \in \mathbb{R}^n} f(x) \text{ subject to } c(x) \geq 0$$  \hspace{1cm} (47)

and its associated first-order necessary conditions

$$\nabla f(\bar{x}) - \nabla c(\bar{x})\bar{\lambda} = 0$$
$$c(\bar{x}) \geq 0$$
$$\bar{\lambda} \geq 0$$
$$\bar{\lambda}_i c_i(\bar{x}) = 0, \ i = 1, 2, \ldots, m,$$  \hspace{1cm} (48)

where $\nabla c(x) = (\nabla c_1(x), \nabla c_2(x), \ldots, \nabla c_m(x)) \in \mathbb{R}^{n \times m}$. Suppose that $\bar{x}$ is a solution of the primal problem and that $f$ and $-c_i, \ i = 1, 2, \ldots, m$ are convex functions on $\mathbb{R}^n$ that are differentiable at $\bar{x}$. Then any $\bar{\lambda}$ for which $(\bar{x}, \bar{\lambda})$ satisfies the first-order necessary conditions of the primal problem is a solution of the dual problem.

Remarks

- The theorem shows that optimal Lagrange multipliers for the primal problem are solutions of the dual problem.
The dual problem

Proof
Assume that \((\bar{x}, \bar{\lambda})\) satisfies the first-order necessary conditions for a minimum of the primal problem and assume that \(L(\cdot, \bar{\lambda})\) is convex and differentiable. Then, for any \(x\), it holds that

\[
L(x, \bar{\lambda}) \geq L(\bar{x}, \bar{\lambda}) + \nabla_x L(\bar{x}, \bar{\lambda})(x - \bar{x}) = L(\bar{x}, \bar{\lambda}),
\]

because \(\nabla_x L(\bar{x}, \bar{\lambda}) = 0\). Hence, we have for the dual objective function

\[
q(\bar{\lambda}) = \inf_x L(x, \bar{\lambda}) = L(\bar{x}, \bar{\lambda})
\]

With the last first-order necessary condition it follows in addition that

\[
q(\bar{\lambda}) = L(\bar{x}, \bar{\lambda}) = f(\bar{x}) - \bar{\lambda}^T c(\bar{x}) = f(\bar{x})
\]

Finally, with the weak duality theorem, we have \(q(\lambda) \leq f(\bar{x})\) for all \(\lambda \geq 0\). Hence with \(q(\bar{\lambda}) = f(\bar{x})\), it follows that \(\bar{\lambda}\) is a solution of the dual problem.

\[\square\]
• Introduction

• Geometry of linear discriminant functions

• Support vector machine training

• Foundations of constrained optimization

• **Quadratic program formulations of SVM training**

• Kernelizing the maximum margin SVM
Theorem (Maximum margin training (1))

The dual problem of the maximum margin SVM training problem

$$\min_{w, w_0} \frac{1}{2} \|w\|^2 \text{ subject to } y(i)(w^T x(i) + w_0) \geq 1 \text{ for } i = 1, ..., n$$

is given by

$$\max_{\lambda \in \mathbb{R}^n} q(\lambda) := \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y(i) y(j) x(i)^T x(j)$$

subject to

$$\lambda \geq 0 \text{ and } \sum_{i=1}^{n} \lambda_i y(i) = 0.$$  

Given a solution \( \bar{\lambda} \) of the dual problem, all \( x^{(k)} \) with \( \lambda_k > 0, k = 1, ..., K \) are support vectors and the solutions for the weight and bias parameters of the primal problem are given by

$$\bar{w} = \sum_{i=1}^{n} \bar{\lambda}_i y(i) x(i) \text{ and } \bar{w}_0 = \frac{1}{K} \sum_{k=1}^{K} \left( y^{(k)} - \bar{w}^T x^{(k)} \right),$$

respectively.
Quadratic program formulations of SVM training

Proof

(1) Lagrangian function of the primal problem

The Lagrangian function of the primal problem

\[
\min_{w,w_0} \frac{1}{2} \|w\|^2 \quad \text{subject to} \quad y^{(i)}(w^T x^{(i)} + w_0) \geq 1 \quad \text{for } i = 1, \ldots, n
\]  

is given by

\[
L(w, w_0, \lambda) := \frac{1}{2} w^T w - \sum_{i=1}^{n} \lambda_i (y^{(i)}(w^T x^{(i)} + w_0) - 1)
\]  

(2) The dual objective function

The dual objective is here defined as

\[
q : \mathbb{R}^n \to \mathbb{R}, \lambda \mapsto q(\lambda) := \min_{w,w_0} L(w, w_0, \lambda)
\]  

Analytically evaluating the minimum of the Lagrangian with respect to \(w\) and \(w_0\) entails evaluating the derivatives of \(L\) with respect to \(w\) and \(w_0\) and setting to zero. To this end, let

\[
\bar{w} := \arg \min_{w \in \mathbb{R}^m} L(w, w_0, \lambda) \quad \text{and} \quad \bar{w}_0 := \arg \min_{w_0 \in \mathbb{R}} L(w, w_0, \lambda)
\]
Proof (cont.)

For the minimization with respect to $w$, we have

$$\nabla_w L(w, w_0, \lambda) = \nabla_w \left( \frac{1}{2} w^T w - \sum_{i=1}^{n} \lambda_i (y^{(i)} (w^T x^{(i)} + w_0) - 1) \right)$$

$$= w - \nabla_w \left( \sum_{i=1}^{n} \lambda_i y^{(i)} w^T x^{(i)} + \lambda_i y^{(i)} w_0 - \lambda_i \right)$$

$$= w - \sum_{i=1}^{n} \lambda_i y^{(i)} x^{(i)}$$

(60)

For the minimum of $L$ with respect to $w$, we thus have

$$\bar{w} = \sum_{i=1}^{n} \lambda_i y^{(i)} x^{(i)}$$

(61)
Proof (cont.)

For the minimization with respect to $w_0$, we have

$$
\nabla_{w_0} L(w, w_0, \lambda) = \nabla_{w_0} \left( \frac{1}{2} w^T w - \sum_{i=1}^{n} \lambda_i (y^{(i)}(w^T x^{(i)} + w_0) - 1) \right)
$$

$$
= \nabla_{w_0} \left( \sum_{i=1}^{n} \lambda_i y^{(i)} w^T x^{(i)} + \lambda_i y^{(i)} w_0 - \lambda_i \right)
$$

$$
= - \sum_{i=1}^{n} \lambda_i y^{(i)}
$$

At the minimum of $L$ with respect to $w_0$, we thus have

$$
- \sum_{i=1}^{n} \lambda_i y^{(i)} = 0 \tag{63}
$$

Note that we do only obtain this condition at a minimum of $L$ with respect to $w_0$, but not a minimizer $\bar{w}_0$. 
Proof (cont.)

For the dual objective function, we thus have

\[ q(\lambda) = \min_{\bar{w}, \bar{w}_0} L(\bar{w}, \bar{w}_0, \lambda) \]

\[ = L(\bar{w}, \bar{w}_0, \lambda) \]

\[ = \frac{1}{2} \bar{w}^T \bar{w} - \sum_{i=1}^{n} \lambda_i \left( y(i)(\bar{w}^T x(i) + \bar{w}_0) - 1 \right) \]

\[ = \frac{1}{2} \left( \sum_{i=1}^{n} \lambda_i y(i) x(i) \right)^T \left( \sum_{j=1}^{n} \lambda_i y(j) x(j) \right) - \sum_{i=1}^{n} \lambda_i \left( y(i) \left( \sum_{j=1}^{n} \lambda_j y(j) x(j) \right)^T x(i) + \bar{w}_0 \right) - 1 \]

\[ = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y(i) y(j) x(i)^T x(j) - \sum_{i=1}^{n} \lambda_i y(i) \left( \sum_{j=1}^{n} \lambda_j y(j) x(j) \right)^T x(i) + \bar{w}_0 \sum_{i=1}^{n} \lambda_i y(i) + \sum_{i=1}^{n} \lambda_i \]

\[ = \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y(i) y(j) x(i)^T x(j) - \bar{w}_0 \sum_{i=1}^{n} \lambda_i y(i) \]

where the last equality follows with the fact that at \( \bar{w}_0 \) it holds that \( \sum_{i=1}^{n} \lambda_i y(i) = 0 \).
Proof (cont.)

We have thus shown that the dual objective function for maximum margin SVM training has the form

\[
q : \mathbb{R}^n \rightarrow \mathbb{R}, \lambda \rightarrow q(\lambda) = \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y^{(i)} y^{(j)} x^{(i)T} x^{(j)}. \tag{64}
\]

(3) The dual problem

The dual problem for maximum margin SVM training thus has the form

\[
\max_{\lambda \in \mathbb{R}} q(\lambda) = \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y^{(i)} y^{(j)} x^{(i)T} x^{(j)} \tag{65}
\]

subject to

\[
\lambda_i \geq 0, i = 1, \ldots, n \text{ and } \sum_{i=1}^{n} \lambda_i y^{(i)} = 0 \tag{66}
\]

where the latter constraint enforces the minimum of the Langrian function with respect to \( w_0 \).
Proof (cont.)

(4) Formulas for optimal weight and bias values

Solving the dual problem for maximum margin SVM training yields an optimal La-
grange multiplier vector

\[
\bar{\lambda} = \arg\max_{\lambda \in \mathbb{R}^n} q(\lambda) = \arg\max_{\lambda \in \mathbb{R}^n} L(\bar{w}, \bar{w}_0, \lambda) \quad (67)
\]

Based on the minimization of \( L \) with respect to \( w \), we thus have

\[
\bar{w} = \sum_{i=1}^{n} \bar{\lambda}_i y^{(i)} x^{(i)} \quad (68)
\]
Proof (cont.)

Finally, for the optimal bias parameter $\bar{w}_0$, we first note that with the KKT conditions it holds that for all $\lambda_k > 0, k = 1, \ldots, K$ that

$$y^{(k)}(\bar{w}^T x^{(k)} + w_0) - 1 = 0$$
$$\iff y^{(k)}(\bar{w}^T x^{(k)} + w_0) = 1$$
$$\iff y^{(k)} y^{(k)} (\bar{w}^T x^{(k)} + w_0) = y^{(k)}$$
$$\iff \bar{w}^T x^{(k)} + w_0 = y^{(k)} \quad (69)$$

This implies (1) that all $x^{(k)}$ with $\lambda_k > 0$ are support vectors, because their distance to the optimal hyperplane is 1 and (2) that

$$\sum_{k=1}^{K} \bar{w}^T x^{(k)} + Kw_0 = \sum_{k=1}^{K} y^{(k)} \quad (70)$$
$$w_0 = \frac{1}{K} \sum_{k=1}^{K} \left( y^{(k)} - \bar{w}^T x^{(k)} \right)$$
Quadratic program formulations of SVM training

Theorem (Maximum margin training (2))

By defining

\[ y := \left( y^{(i)} \right)_{i=1,\ldots,n} \in \mathbb{R}^n \quad \text{and} \quad K := \left( x^{(i)\,T} x^{(j)} \right)_{i,j=1,\ldots,n} \in \mathbb{R}^{n \times n}, \quad (71) \]

as well as

\[ P := yy^T K \in \mathbb{R}^{n \times n}, \quad q := -1_n, \quad G := -I_n, \quad h := 0_n, \quad A := y^T, \quad \text{and} \quad b := 0 \quad (72) \]

the dual problem of the maximum margin SVM training problem can be written as the quadratic programming problem

\[ \min_{\lambda} \frac{1}{2} \lambda^T P \lambda + q^T \lambda \quad \text{subject to} \quad G \lambda \leq h \quad \text{and} \quad A \lambda = b \]

as solved by cvxopt.solvers.qp.
Proof

The equivalences

\[
\sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y^{(i)} y^{(j)} x^{(i)T} x^{(j)} \Leftrightarrow \lambda^T y y^T K \lambda \Leftrightarrow \lambda^T P \lambda
\]

\[
\sum_{i=1}^{n} \lambda_i \Leftrightarrow 1_n^T \lambda \Leftrightarrow q^T \lambda
\]

\[
\lambda \geq 0 \Leftrightarrow -I_n \leq 0_n \Leftrightarrow G \lambda \leq h \]

\[
\sum_{i=1}^{n} \lambda_i y^{(i)} = 0 \Leftrightarrow y^T \lambda = 0 \Leftrightarrow A \lambda = b
\]

follow directly with the rules of matrix multiplication
Theorem (Soft margin training)

The dual problem of the soft margin SVM training problem

\[
\min_{w,w_0,\xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i \quad \text{subject to } y^{(i)}(w^T x^{(i)} + w_0) \geq 1 - \xi_i, \xi \geq 0 \tag{75}
\]

is given by

\[
\max_{\lambda \in \mathbb{R}^n} q(\lambda) := \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y^{(i)} y^{(j)} x^{(i)T} x^{(j)} \tag{76}
\]

subject to

\[
0 \leq \lambda_i \leq C \text{ for } i = 1, \ldots, n \geq 0 \text{ and } \sum_{i=1}^{n} \lambda_i y^{(i)} = 0. \tag{77}
\]

Given a solution \(\bar{\lambda}\) of the dual problem, all \(x^{(k)}\) with \(\lambda_k > 0, k = 1, \ldots, K\) are support vectors and the solutions for the weight and bias parameters of the primal problem are given by

\[
\bar{w} = \sum_{i=1}^{n} \bar{\lambda}_i y^{(i)} x^{(i)} \quad \text{and} \quad \bar{w}_0 = \frac{1}{K} \sum_{k=1}^{K} \left( y^{(k)} - \bar{w}^T x^{(k)} \right), \tag{78}
\]

respectively.
• Introduction

• Geometry of linear discriminant functions

• Support vector machine training

• Foundations of constrained optimization

• Quadratic program formulations of SVM training

• Kernelizing the maximum margin SVM
The dual problem formulation of SVM training gave rise to \textit{kernel methods}.

The key insight is that the objective function of maximum margin SVM training

\begin{equation}
q(\lambda) := \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y^{(i)} y^{(j)} x^{(i)T} x^{(j)}
\end{equation}

depends only on the scalar products

\begin{equation}
x^{(i)T} x^{(j)} \text{ for } i, j = 1, ..., n
\end{equation}

Projecting each training pattern into a “high-dimensional feature space” in which linear separably can be hoped for thus only necessitates the evaluation of the scalar product in this space.
Kernelizing the maximum margin SVM

\[ \phi : \mathcal{X} \to \tilde{X}, \, x \mapsto \phi(x) \text{ with } k(x, x') = \langle \phi(x), \phi(x') \rangle \] (81)
Kernelizing the maximum margin SVM
References


