Lecture 14: Support Vector Machines

Isabel Valera

Machine Learning Group
Department of Mathematics and Computer Science
Saarland University, Saarbrücken, Germany

07.06.2021

Outline

Main references

- Learning with Kernels - Chapter 7
- Bishop - Chapter 7

Outline

Linear Classification

Let $\mathcal{X} = \mathbb{R}^d$ be the input space, then the classifier $\hat{y} : \mathbb{R}^d \to \{-1, 1\}$ has the form

$$\hat{y}(x) = \operatorname{sign}(f\mathbf{x}) = \operatorname{sign}(\langle \mathbf{w}, \mathbf{x} \rangle + b) = \begin{cases} 1 & \text{if } \langle \mathbf{w}, \mathbf{x} \rangle + b > 0, \\ -1 & \text{if } \langle \mathbf{w}, \mathbf{x} \rangle + b \leq 0. \end{cases}$$

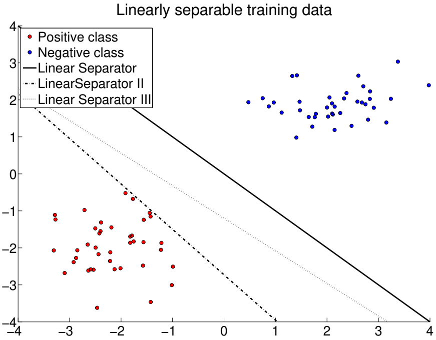
Separation of the input space $\mathbb{R}^d$ into two half spaces.

A training set $D = (\mathbf{x}_i, y_i)_{i=1}^n$ is **linearly separable** if there exists a weight vector $\mathbf{w}$ and an offset b such that,

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

$\Rightarrow$ There exists a **hyperplane** $\{x \in \mathbb{R}^d \mid \langle \mathbf{w}, \mathbf{x} \rangle + b = 0\}$ which each separates the sets $\mathbf{X}_+ = \{\mathbf{x}_i \in D \mid y_i = 1\}$ and $\mathbf{X}_- = \{\mathbf{x}_i \in D \mid y_i = -1\}$.

Example



A training sample of a two-class problem in $\mathbb{R}^2$. The two classes are linearly separable and three different decision hyperplanes are shown which separate the two classes. (Image by Prof. Hein)

Basis functions

No distinction between the original input space $\mathcal{X} = \mathbb{R}^d$ and a possibly larger **feature space**, where we use basis functions/feature maps ϕ_i

$$\mathbf{x} \in \mathbb{R}^d \longrightarrow (\phi_1(\mathbf{x}), \ldots, \phi_m(\mathbf{x})),$$

to the feature space $\mathbb{R}^m$.

Functions are linear in the parameters but not necessarily linear in the input space!

Basis functions

No distinction between the original input space $\mathcal{X} = \mathbb{R}^d$ and a possibly larger **feature space**, where we use basis functions/feature maps ϕ_i

$$\mathbf{x} \in \mathbb{R}^d \longrightarrow (\phi_1(\mathbf{x}), \ldots, \phi_m(\mathbf{x})),$$

to the feature space $\mathbb{R}^m$.

Functions are linear in the parameters but not necessarily linear in the input space!

Definition

Let $g : \mathcal{X} \to \mathbb{R}$ be a function and $\hat{y}(\mathbf{x}) = \mathrm{sign}(f(\mathbf{x}))$ be the resulting classifier with output in $\mathcal{Y} = \{-1, 1\}$, then we call the set

$$\{\mathbf{x} \in \mathcal{X} \mid f(\mathbf{x}) = 0\},$$

the **decision boundary** of the classifier $\hat{y}$.

Methods for linear classification

Three linear methods: $\hat{y}(x) = \text{sign}(f(\mathbf{x})) = \text{sign}(\langle \mathbf{w}, \Phi(\mathbf{x}) \rangle)$.

- **Linear Discriminant Analysis**:
  - Loss: Squared loss, $L(y, f(\mathbf{x})) = (y - f(\mathbf{x}))^2$
  - Regularization: none
- **Logistic Regression**:
  - Loss: Logistic loss, $L(y, f(\mathbf{x})) = \log(1 + \exp(-y\, f(\mathbf{x})))$
  - Regularization: usually none, but there exist regularized versions.
- **Support Vector Machines** (Lecture 14).
  - Loss: hinge loss, $L(y, f(\mathbf{x})) = \max(0, 1 - y\, f(x))$
  - Regularization: L2-regularization, i.e., $\Omega(\mathbf{w}) = \|\mathbf{w}\|_2^2$

All three methods construct a **linear** classifier but all three have different **objectives**.

Outline

Motivation

The linear **support vector machine** (SVM) can be motivated from different perspectives.

Geometric Perspective: Maximum margin hyperplane
Unique hyperplane which correctly classifies the data and has maximal distance/margin to the training data.

- **hard margin** case: linearly separable data.
- **soft margin** case: all kind of data allowed.

Maximum margin hyperplane

Maximum margin hyperplane: a hyperplane which correctly classifies the data and has maximum distance/margin to the data.

Definition

A **maximum margin hyperplane** $(\mathbf{w}, b)$ for a **linearly separable** set of training data $(\mathbf{x}_i, y_i)_{i=1}^n$ is defined as

$$\max_{\mathbf{w}\in\mathbb{R}^d,\, b\in\mathbb{R}} \min\{\|\mathbf{x} - \mathbf{x}_i\| \mid \langle \mathbf{w}, \mathbf{x}\rangle + b = 0,\ \mathbf{x} \in \mathbb{R}^d,\ i = 1, \dots, n\},$$

where we optimize over all $(\mathbf{w}, b)$ such that $y_i\left(\langle \mathbf{w}, \mathbf{x}_i\rangle + b\right) > 0$.

- Linear classifier is determined by the weight vector $\mathbf{w}$ and the offset b.

$$\hat{y}(\mathbf{x}) = \mathrm{sign}(\langle \mathbf{w}, \mathbf{x}\rangle + b).$$

- classifier and the decision boundary are not unique. For $\gamma > 0$, $\tilde{\mathbf{w}} = \gamma\mathbf{w}$ and $\tilde{b} = \gamma b$ gives the same classifier.

Geometrical margin and canonical hyperplane

Definition (Geometrical margin)

For a hyperplane $\{\mathbf{x} | \langle \mathbf{w}, \mathbf{x} \rangle + b = 0\}$, the **geometrical margin** of a point $(\mathbf{x}, y)$ is:

$$\rho_{\mathbf{w},b}(\mathbf{x}, y) = y(\langle \mathbf{w}, \mathbf{x} \rangle + b) / \|\mathbf{w}\| .$$

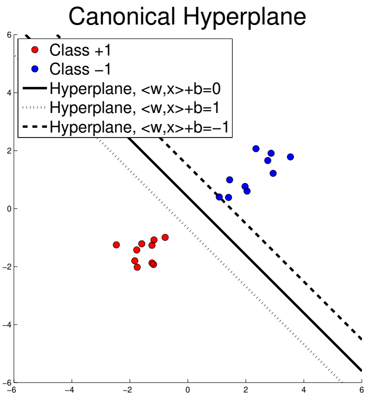
Definition (Canonincal hyperplane)

The pair $(\mathbf{w}, b) \in \mathbb{R}^d \times \mathbb{R}$ is said to be in **canonical** form with respect to $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$, if it is scaled such that

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

which implies that the point closest to the hyperplane $h = \{\mathbf{x} | \langle \mathbf{w}, \mathbf{x} \rangle + b = 0\}$ has distance $\rho = \frac{1}{\|\mathbf{w}\|}$.

Illustration



Canonical hyperplane for a set of training points $(\mathbf{x}_i)_{i=1}^n$.
(Image by Prof. Hein)

SVM formulation

Formulation:

$$\max_{\mathbf{w}\in\mathbb{R}^d,\ b\in\mathbb{R}} \frac{1}{\|\mathbf{w}\|}$$
$$\text{subject to: } y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) \geq 1, \quad \forall\, i = 1, \dots, n$$

Second equivalent formulation:

$$\min_{\mathbf{w}\in\mathbb{R}^d,\ b\in\mathbb{R}} \frac{1}{2}\|\mathbf{w}\|^2$$
$$\text{subject to: } y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) \geq 1, \quad \forall\, i = 1, \dots, n$$

Observation: convex optimization problem – quadratic program

Outline

Lagranange function

Lagrange function: Let $\mathbf{w} \in \mathbb{R}^d$ and $\alpha \in \mathbb{R}^n$

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2} \|\mathbf{w}\|^2 + \sum_{i=1}^{n} \alpha_i \Big[1 - y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b)\Big],$$

where $\alpha_i \geq 0, \quad \forall\, i = 1, \dots, n$, are the **Lagrange multipliers**.

Dual Lagrange function:

$$q(\alpha) = \inf_{\mathbf{w} \in \mathbb{R}^d,\ b \in \mathbb{R}} L(\mathbf{w}, b, \alpha).$$

Observations:

- L is convex!
- Slater condition fulfilled if data is linearly separable $\Rightarrow$ strong duality
- We can solve primal problem via the dual problem.

Optimality conditions

Derivatives:

$$\nabla_{\mathbf{w}} L(\mathbf{w}, b, \alpha) = \mathbf{w} - \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i, \qquad \frac{\partial L(\mathbf{w}, b, \alpha)}{\partial b} = -\sum_{i=1}^{n} \alpha_i y_i.$$

Conditions for global minimum:

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i, \qquad \sum_{i=1}^{n} \alpha_i y_i = 0.$$

Plugging these expressions into $L(\mathbf{w}, b, \alpha)$ we get **the dual Lagrangian**:

$$q(\alpha) = -\frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle + \sum_{i=1}^{n} \alpha_i,$$

where $\alpha_i \geq 0, \quad \forall\, i = 1, \ldots, n.$

SVM dual formulation

Dual problem:

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^n} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i\alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle ,$$

$$\text{subject to: } \alpha_i \geq 0, \quad i = 1, \ldots, n,$$

$$\sum_{i=1}^{n} y_i\alpha_i = 0.$$

Observations:

- The dual problem is solved in practice using SMO (Sequential minimal optimization) method.
- Complexity is in the worst case cubic in n but often much faster.

KKT conditions

Karush-Kuhn-Tucker (KKT) conditions: The most important one is the complementary slackness condition:

$$\Big[1 - y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b)\Big] = 0 \quad \text{if} \quad \alpha_i > 0$$
$$\text{and} \qquad \alpha_i = 0 \quad \text{if} \quad \Big[1 - y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b)\Big] < 0.$$

or more compactly

$$\alpha_i \Big[1 - y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b)\Big] = 0.$$

The offset b can thus be determined by averaging the value $y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle$ over all points with $\alpha_i > 0$:

$$b = \frac{1}{\sum_{i=1}^n \mathbb{1}_{\alpha_i > 0}} \sum_{i=1}^n \mathbb{1}_{\alpha_i > 0} \left(y_i - \sum_{j=1}^n \alpha_j \, y_j \, \langle \mathbf{x}_i, \mathbf{x}_j \rangle \right).$$

Support Vectors

Final weight vector:

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i.$$

Only the closest points to the decision boundary contribute to solution, i.e.,

$$\alpha_i > 0 \quad \Leftrightarrow \quad \Big[1 - y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b)\Big] = 0,$$

The points $\mathbf{x}_i$ for which $\alpha_i > 0$ are called **support vectors**. The area between the two supporting hyperplanes $\{\mathbf{x} \mid \langle \mathbf{w}, \mathbf{x} \rangle + b = 1\}$ and $\{\mathbf{x} \mid \langle \mathbf{w}, \mathbf{x} \rangle + b = -1\}$ is called the **margin**.

Observations:

1. The weight vector of the support vector machine is typically **sparse** in terms of $\boldsymbol{\alpha}$.
2. Modifications of the training points matter only if they move into the margin.

Convex hull formulation

Equivalent reformulation of the dual problem:

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^n} \left\| \sum_{i=1, y_i=1}^{n} \alpha_i \mathbf{x}_i - \sum_{j=1, y_j=-1}^{n} \alpha_j \mathbf{x}_j \right\|^2,$$

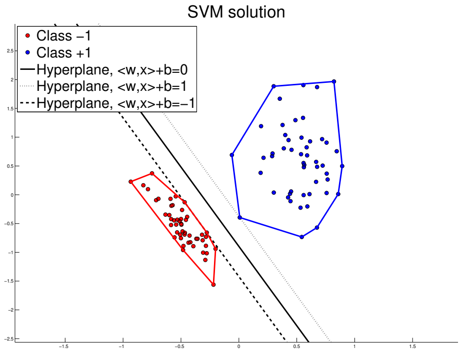
$$\text{subject to: } \alpha_i \geq 0, \quad i = 1, \ldots, n,$$

$$\sum_{i=1, y_i=1}^{n} \alpha_i = \sum_{j=1, y_j=-1}^{n} \alpha_j = 1.$$

Observations:

- It can be shown that the above problem maximizes the distance between the convex hulls of the positive and negative class.
- The maximum margin hyperplane is the one bisecting the shortest line orthogonally connecting both hulls.

Example: linearly separable case



A linearly separable problem. The hard margin solution of the SVM is shown together with the convex hulls of the positive and negative class. The points on the margin, that is $\langle \mathbf{w}, \mathbf{x} \rangle + b = \pm 1$, are called **support vectors**. (Image by Prof. Hein)

Outline

Transition to soft-margin

Problems of the hard margin case:

- in general, data is not linearly separable,
- the **hard margin** case is often too strict since it is sensitive to outliers.

Relaxation of the constraints:

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

where $\xi_i \geq 0$ are the **slack variables**.

Primal problem of the soft-margin case:

$$\begin{aligned} \min_{\mathbf{w} \in \mathbb{R}^d,\ b \in \mathbb{R},\ \boldsymbol{\xi} \in \mathbb{R}^n} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + \frac{C}{n} \sum_{i=1}^{n} \xi_i \\ \text{subject to:} \quad & y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 - \xi_i, \quad \forall\, i = 1, \ldots, n, \\ & \xi_i \geq 0, \quad \forall\, i = 1, \ldots, n \end{aligned}$$

Soft Margin as RERM

At the optimum: (note that $\xi_i \geq 0$)

$$\xi_i = \max\Big(0, 1 - y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b)\Big),$$

where we recall that $\max\Big(0, 1 - y_i f(\mathbf{x}_i)\Big)$ is the **hinge loss**.

Soft Margin SVM is RERM with Hinge loss and L_2-regularization:

$$\min_{\mathbf{w}\in\mathbb{R}^d,\ b\in\mathbb{R}} C\,\frac{1}{n}\sum_{i=1}^{n} \max\Big(0, 1 - y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b)\Big) + \|\mathbf{w}\|^2\,,$$

Error parameter C is the inverse of the regularization parameter $\lambda = \frac{1}{C}$.

Lagrangian of Soft Margin

Lagrangian of the soft margin problem:

$$L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i\Big[1 - \xi_i - y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b)\Big] - \sum_{i=1}^{n}\beta_i\xi_i$$

where $\alpha_i \geq 0,\ i = 1, \ldots, n$ and $\beta_i \geq 0,\ i = 1, \ldots, n$.

Conditions for a stationary point: ($\mathbf{1}$ is an n-dimensional vector of ones)

$$\mathbf{w} = \sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i, \qquad \sum_{i=1}^{n}\alpha_i y_i = 0, \qquad \boldsymbol{\beta} = \frac{C}{n}\mathbf{1} - \boldsymbol{\alpha}.$$

The last equation can be used to get rid of $\boldsymbol{\beta}$. Due to the positivity of $\boldsymbol{\beta}$ we get the new constraint for $\boldsymbol{\alpha}$ as

$$0 \leq \alpha_i \leq \frac{C}{n}, \quad i = 1, \ldots, n.$$

Lagrangian of Soft Margin

Dual Lagrangian of the soft margin problem:

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^n} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i\alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle ,$$

$$\text{subject to: } 0 \leq \alpha_i \leq \frac{C}{n}, \quad i = 1, \ldots, n, \qquad \sum_{i=1}^{n} y_i\alpha_i = 0.$$

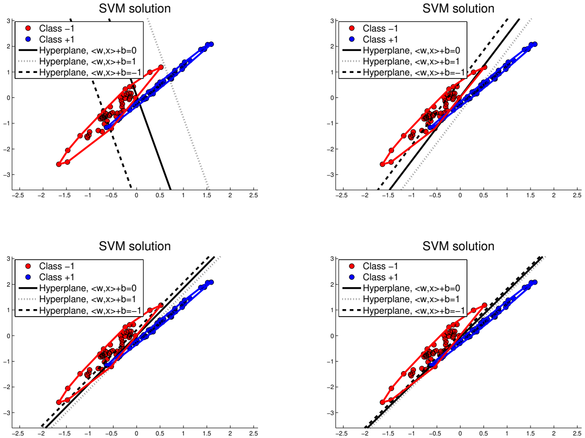
Complementary slackness conditions (part of KKT conditions) of the original problem:

$$\alpha_i \Big[1 - \xi_i - y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b)\Big] = 0 \quad \text{and} \qquad \beta_j \xi_j = 0, \quad \text{for} \quad i, j = 1, \ldots, n.$$

Three classes of points:

- $\alpha_i = 0$, outside the margin and all correctly classified.
- $0 < \alpha_i < \frac{C}{n}$, lie exactly on the margin and are all correctly classified.
- $\alpha_i = \frac{C}{n}$, inside the margin and may be misclassified.

Comparison of different C



Top row: error parameter $C = 10$ (left) and $C = 10^2$ (right)
Bottom row: error parameter $C = 10^3$ (left) and $C = 10^4$ (right).
(Image by Prof. Hein)

Outline

Summary

- Linear SVMs find the hyperplane that maximizes the margin between two classes in linearly separable datasets. A solution is found using the dual optimization problem.
- The resulting classifier (hyperplane) is computed only using the support vectors, i.e., those datapoints that lie exactly at the margin.
- Thus, the SVM classifier only varies across datasets if the support vectors change. This is due to the robust Hinge loss.
- For nonlinearly separable datasets, we relax the formulation and allow a subset of observations to lie inside the margin. The parameter C controls the proportion of observations that can lie inside the margin.
- We can generalize SVM to non-linear problems using specific basis functions that result from a similarity function over pairs of data points, known as **kernels** (next lecture!).