Kernel Learning And Its Application In Nonlinear Support Vector Machines

Sven Nekula, Joshua Simon

Otto-Friedrich-University Bamberg

sven.nekula@stud.uni-bamberg.de
joshua-guenter.simon@stud.uni-bamberg.de

May 16, 2021

Overview

1. Introduction
   - Linearly separable data classes
   - Hyperplane classifiers - Solving an optimization problem
2. Linear SVMs
   - Maximum margin separator
   - Limitations
3. Nonlinear SVMs
   - The kernel trick
   - Solving nonlinear separable classification problemes

Linearly separable data classes

First, let's consider a given data set $\mathcal{X}$ of labeled points with individual labels $y_i \in \{-1, 1\}$. Our goal is to implement a classification method, which is able to classify new and unlabeld data points with the right or "best" label.

- Plot of data points -

Linearly separable data classes

In machine learning, an well established classification method are the so called **Support Vector Machines** (SVM). Developed by Vladimir Vapnik and his coworkers in the 1990s, SVMs are still a relevent topic and an even more powerfull tool for **classification** and **regression**.

Hyperplane classifiers

The underlying learning algorithm of SVMs yields to find a hyperplane in some dot product space $\mathcal{H}$, which separates the data. A hyperplane of the form

$$\langle w, x \rangle + b = 0 \tag{1}$$

where $w \in \mathcal{H}, b \in \mathbb{R}$ shall be considered [p.11 Kernel Learning Book]. Futhermore decision functions

$$f(x) = sgn\left(\langle w, x \rangle + b\right) \tag{2}$$

can be used as similarity measures for unlabeld data.

Hyperplane classifiers - A constrained optimization problem

The optimal hyperplane can be calculated by finding the normal vector that leads to the largest margin. Thus we need to solve the optimization problem

$$\begin{aligned} \min_{w \in \mathcal{H}, b \in \mathbb{R}} \quad & \tau(w) = \frac{1}{2}\|w\|^2 \\ \text{subject to} \quad & y_i\left(\langle w, x \rangle + b\right) \geq 1 \ \forall i = 1, \ldots, m. \end{aligned} \tag{3}$$

The constraints in (3) ensure that $f(x_i)$ will be $+1$ for $y_i = +1$ and -1 for $y_i = -1$. The ≥ 1 on the right hand side of the constraints effectively fixes the scaling of w. This leads to the maximum margin hyperplane. A detailed explanation can be found in [Chap 7, Kernel Learning Book].

Hyperplane classifiers - Lagrangian

The constrained optimization problem in (3) can be re-written using the method of Lagrange multipliers. This leads to the Lagrangian

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{m} \alpha_i \left(y_i \left(\langle w, x \rangle + b\right) - 1\right) \tag{4}$$

subject to $\alpha_i \geq 0 \ \forall i = 1, \ldots, m$. Here, α_i are the Lagrange multipliers. The Lagrangian L has to be minimized with respect to the primal variables w and b and maximized with respect to the dual variables α_i (in other words, a saddle point has to be found).

Hyperplane classifiers - KKT conditions

The Karush-Kuhn-Tucker (KKT) complementarity conditions of optimization theory state, that at the saddle point, the derivatives of L with respect to the primal variables must vanish, since

$$\frac{\partial}{\partial b} L(w, b, \alpha) = 0 \text{ and } \frac{\partial}{\partial w} L(w, b, \alpha) = 0 \tag{5}$$

leads to

$$\sum_{i=1}^{m} \alpha_i y_i = 0 \text{ and } w = \sum_{i=1}^{m} \alpha_i y_i x_i . \tag{6}$$

The solution vector w thus has an expansion in terms of a subset of the training patterns, namely those patterns with non-zero α_i, called Support Vectors (SVs).

Hyperplane classifiers - Dual optimization problem

We can again re-write our optimization problem by substituting (6) into the Lagrangian (4) to eliminate the primal variables. This yields the dual optimization problem, which is usually solved in practice

$$\begin{aligned}
\max_{\alpha \in \mathbb{R}^m} \quad & W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \\
\text{subject to} \quad & \alpha_i \geq 0 \; \forall i = 1, \ldots, m \text{ and } \sum_{i=1}^{m} \alpha_i y_i = 0.
\end{aligned} \tag{7}$$

Hyperplane classifiers - Dual optimization problem

Finally, the decision function can be re-written using (6) as

$$f(x) = sgn\left(\sum_{i=1}^{m} \alpha_i y_i \langle x, x_i \rangle + b\right), \tag{8}$$

where b can be computed by exploiting $\alpha_i \left[y_i \left(\langle x_i, w \rangle + b \right) - 1 \right] = 0$, which follows from the KKT conditions.

Maximum margin separator

We now have all the theoretical background to go back to our inital classification problem. We can implement a SVM as an maximum margin separator for the given data set $\mathcal{X}$.

- Python code, plots, etc. -

Limitations

Let's consider the following data set.

- Plot of nonlinear separable data -

Paragraphs of Text

Sed iaculis dapibus gravida. Morbi sed tortor erat, nec interdum arcu. Sed id lorem lectus. Quisque viverra augue id sem ornare non aliquam nibh tristique. Aenean in ligula nisl. Nulla sed tellus ipsum. Donec vestibulum ligula non lorem vulputate fermentum accumsan neque mollis.

Sed diam enim, sagittis nec condimentum sit amet, ullamcorper sit amet libero. Aliquam vel dui orci, a porta odio. Nullam id suscipit ipsum. Aenean lobortis commodo sem, ut commodo leo gravida vitae. Pellentesque vehicula ante iaculis arcu pretium rutrum eget sit amet purus. Integer ornare nulla quis neque ultrices lobortis. Vestibulum ultrices tincidunt libero, quis commodo erat ullamcorper id.

Bullet Points

- Lorem ipsum dolor sit amet, consectetur adipiscing elit
- Aliquam blandit faucibus nisi, sit amet dapibus enim tempus eu
- Nulla commodo, erat quis gravida posuere, elit lacus lobortis est, quis porttitor odio mauris at libero
- Nam cursus est eget velit posuere pellentesque
- Vestibulum faucibus velit a augue condimentum quis convallis nulla gravida

Blocks of Highlighted Text

Block 1

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Integer lectus nisl, ultricies in feugiat rutrum, porttitor sit amet augue. Aliquam ut tortor mauris. Sed volutpat ante purus, quis accumsan dolor.

Block 2

Pellentesque sed tellus purus. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Vestibulum quis magna at risus dictum tempor eu vitae velit.

Block 3

Suspendisse tincidunt sagittis gravida. Curabitur condimentum, enim sed venenatis rutrum, ipsum neque consectetur orci, sed blandit justo nisi ac lacus.

Multiple Columns

Heading

1. Statement
2. Explanation
3. Example

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Integer lectus nisl, ultricies in feugiat rutrum, porttitor sit amet augue. Aliquam ut tortor mauris. Sed volutpat ante purus, quis accumsan dolor.

Table

Treatments	Response 1	Response 2
Treatment 1	0.0003262	0.562
Treatment 2	0.0015681	0.910
Treatment 3	0.0009271	0.296

Table: Table caption

Theorem

Theorem (Mass–energy equivalence)

$E = mc^2$

Verbatim

Example (Theorem Slide Code)

```
\begin{frame}
\frametitle{Theorem}
\begin{theorem}[Mass--energy equivalence]
$E = mc^2$
\end{theorem}
\end{frame}
```

Figure

Uncomment the code on this slide to include your own image from the same directory as the template .TeX file.

Citation

An example of the `\cite` command to cite within the presentation:

This statement requires citation [Smith, 2012].

References

John Smith (2012)
Title of the publication
Journal Name 12(3), 45 – 678.

The End