These are the slides of the lecture

Pattern Recognition
Winter term 2020/21
Friedrich-Alexander University of Erlangen-Nuremberg.

Erlangen, January 8, 2021
Prof. Dr.-Ing. Andreas Maier

Pattern Recognition (PR)

Prof. Dr.-Ing. Andreas Maier
Pattern Recognition Lab (CS 5), Friedrich-Alexander-Universität Erlangen-Nürnberg
Winter Term 2020/21

Support Vector Machines II

Hard Margin Problem

The hard margin SVM optimization problem is formulated as:

$$\text{minimize} \quad \frac{1}{2}\|\boldsymbol{\alpha}\|_2^2$$

$$\text{subject to} \quad \forall i: \quad y_i \cdot (\boldsymbol{\alpha}^T \boldsymbol{x}_i + \alpha_0) - 1 \geq 0$$

Soft Margin Problem

The soft margin SVM optimization problem is formulated as:

$$\text{minimize} \quad \frac{1}{2}\|\boldsymbol{\alpha}\|_2^2 + \mu \sum_i \xi_i$$

$$\text{subject to} \quad \forall i: \quad -(y_i \cdot (\boldsymbol{\alpha}^T \boldsymbol{x}_i + \alpha_0) - 1 + \xi_i) \leq 0 \,,$$

$$\forall i: \quad -\xi_i \leq 0$$

Lagrangian

The solution of the constrained convex optimization problem requires the Lagrangian:

$$\begin{aligned} L(\boldsymbol{\alpha}, \alpha_0, \boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\mu}) &= \frac{1}{2}\|\boldsymbol{\alpha}\|_2^2 + \mu \sum_i \xi_i - \sum_i \mu_i \xi_i \\ &\quad - \sum_i \lambda_i (y_i \cdot (\boldsymbol{\alpha}^T \boldsymbol{x}_i + \alpha_0) - 1 + \xi_i) \end{aligned}$$

Lagrangian

The solution of the constrained convex optimization problem requires the Lagrangian:

meta-parameter **Lagrangian multiplier**

$$
\begin{aligned}
L(\boldsymbol{\alpha}, \alpha_0, \boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \quad = \quad & \frac{1}{2}\|\boldsymbol{\alpha}\|_2^2 + \mu \sum_i \xi_i - \sum_i \mu_i \xi_i \\
& - \sum_i \lambda_i (y_i \cdot (\boldsymbol{\alpha}^T \boldsymbol{x}_i + \alpha_0) - 1 + \xi_i)
\end{aligned}
$$

Lagrangian

The solution of the constrained convex optimization problem requires the Lagrangian:

meta-parameter **Lagrangian multiplier**

$$
\begin{aligned}
L(\boldsymbol{\alpha}, \alpha_0, \boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \quad = \quad & \frac{1}{2}\|\boldsymbol{\alpha}\|_2^2 + c \sum_i \xi_i - \sum_i \mu_i \xi_i \\
& - \sum_i \lambda_i (y_i \cdot (\boldsymbol{\alpha}^T \boldsymbol{x}_i + \alpha_0) - 1 + \xi_i)
\end{aligned}
$$

Lagrangian (cont.)

Partial derivatives I:

$$\frac{\partial L(\boldsymbol{\alpha}, \alpha_0, \boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\mu})}{\partial \boldsymbol{\alpha}} = \boldsymbol{\alpha} - \sum_i \lambda_i y_i \boldsymbol{x}_i \overset{!}{=} 0.$$

Thus we have:

$$\boldsymbol{\alpha} = \sum_i \lambda_i y_i \boldsymbol{x}_i \; .$$

Lagrangian (cont.)

Partial derivatives II:

$$\frac{\partial L(\boldsymbol{\alpha}, \alpha_0, \boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\mu})}{\partial \alpha_0} = -\sum_i \lambda_i y_i \stackrel{!}{=} 0$$

Lagrangian (cont.)

Partial derivatives II:

$$\frac{\partial L(\boldsymbol{\alpha}, \alpha_0, \boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\mu})}{\partial \alpha_0} = -\sum_i \lambda_i y_i \stackrel{!}{=} 0$$

Partial derivatives III:

$$\frac{\partial L(\boldsymbol{\alpha}, \alpha_0, \boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\mu})}{\partial \xi_i} = c - \mu_i - \lambda_i \stackrel{!}{=} 0$$

Lagrange Dual

Let us consider the Lagrange function for the dual problem for the hard margin case:

$$L_{\mathrm{D}} = \frac{1}{2}\boldsymbol{\alpha}^T\boldsymbol{\alpha} - \sum_i \lambda_i(y_i \cdot (\boldsymbol{\alpha}^T\boldsymbol{x}_i + \alpha_0) - 1)$$

Lagrange Dual

Let us consider the Lagrange function for the dual problem for the hard margin case:

$$
\begin{aligned}
L_{\mathrm{D}} &= \frac{1}{2}\boldsymbol{\alpha}^T\boldsymbol{\alpha} - \sum_i \lambda_i \big(y_i \cdot (\boldsymbol{\alpha}^T \boldsymbol{x}_i + \alpha_0) - 1\big) \\
&= \frac{1}{2}\boldsymbol{\alpha}^T\boldsymbol{\alpha} - \underbrace{(\sum_i \lambda_i y_i \cdot \boldsymbol{x}_i)^T}_{\boldsymbol{\alpha}^T} \boldsymbol{\alpha} - \underbrace{\sum_i \lambda_i y_i}_{=0} \; \alpha_0 + \sum_i \lambda_i
\end{aligned}
$$

Lagrange Dual

Let us consider the Lagrange function for the dual problem for the hard margin case:

$$
\begin{aligned}
L_{\mathrm{D}} &= \frac{1}{2}\boldsymbol{\alpha}^T\boldsymbol{\alpha} - \sum_i \lambda_i \left(y_i \cdot (\boldsymbol{\alpha}^T \boldsymbol{x}_i + \alpha_0) - 1\right) \\
&= \frac{1}{2}\boldsymbol{\alpha}^T\boldsymbol{\alpha} - \underbrace{(\sum_i \lambda_i y_i \cdot \boldsymbol{x}_i)^T}_{\boldsymbol{\alpha}^T} \boldsymbol{\alpha} - \underbrace{\sum_i \lambda_i y_i}_{=0} \; \alpha_0 + \sum_i \lambda_i \\
&= -\frac{1}{2}\sum_i \sum_j \lambda_i \lambda_j y_i y_j \cdot \boldsymbol{x}_i^T \boldsymbol{x}_j + \sum_i \lambda_i
\end{aligned}
$$

The Lagrange Dual Problem

The Lagrange dual problem is given by the optimization problem:

$$\begin{aligned} &\text{maximize} && -\frac{1}{2}\sum_i \sum_j \lambda_i \lambda_j y_i y_j \cdot \boldsymbol{x}_i^T \boldsymbol{x}_j + \sum_i \lambda_i \\ &\text{subject to} && \boldsymbol{\lambda} \succeq 0 \\ & && \sum_i \lambda_i y_i = 0 \end{aligned}$$

The Lagrange Dual Problem

The Lagrange dual problem is given by the optimization problem:

$$\begin{aligned} \text{maximize} \quad & -\frac{1}{2}\sum_i\sum_j \lambda_i\lambda_j y_i y_j \cdot \boldsymbol{x}_i^T \boldsymbol{x}_j + \sum_i \lambda_i \\ \text{subject to} \quad & \boldsymbol{\lambda} \succeq 0 \\ & \sum_i \lambda_i y_i = 0 \end{aligned}$$

Benefits of the dual representation

- The model can be reformulated using kernels.
- SVMs can be applied efficiently to feature spaces whose dimensionality exceeds the number of samples.

Lagrange Dual Problem (cont.)

For convex optimization problems with differentiable objective and constraint functions, the duality gap is zero, if the KKT conditions are satisfied.

Lagrange Dual Problem (cont.)

For convex optimization problems with differentiable objective and constraint functions, the duality gap is zero, if the KKT conditions are satisfied.

Especially the complementary slackness condition is interesting for us:

$$\forall i : \quad \lambda_i f_i(\boldsymbol{x}) = 0$$

Lagrange Dual Problem (cont.)

Complementary slackness for hard margin SVMs:

$$\forall i: \quad \lambda_i \left(y_i \cdot (\boldsymbol{\alpha}^T \boldsymbol{x}_i + \alpha_0) - 1 \right) = 0$$

Lagrange Dual Problem (cont.)

Complementary slackness for hard margin SVMs:

$$\forall i : \quad \lambda_i \left(y_i \cdot (\boldsymbol{\alpha}^T \boldsymbol{x}_i + \alpha_0) - 1 \right) = 0$$

Implications:

1. If $\lambda_i > 0$, then $y_i \left(\boldsymbol{\alpha}^T \boldsymbol{x}_i + \alpha_0 \right) - 1 = 0$, and thus:

$$y_i(\boldsymbol{\alpha}^T \boldsymbol{x}_i + \alpha_0) = 1 \ .$$

All $\boldsymbol{x}_i$ with $\lambda_i > 0$ are elements at the boundary of the slab; these samples are called *support vectors*.

Lagrange Dual Problem (cont.)

Complementary slackness for hard margin SVMs:

$$\forall i : \quad \lambda_i \left(y_i \cdot (\boldsymbol{\alpha}^T \boldsymbol{x}_i + \alpha_0) - 1 \right) = 0$$

Implications:

1. If $\lambda_i > 0$, then $y_i \left(\boldsymbol{\alpha}^T \boldsymbol{x}_i + \alpha_0 \right) - 1 = 0$, and thus:

   $$y_i(\boldsymbol{\alpha}^T \boldsymbol{x}_i + \alpha_0) = 1 \; .$$

   All $\boldsymbol{x}_i$ with $\lambda_i > 0$ are elements at the boundary of the slab;
   these samples are called *support vectors*.

2. We have seen that $\boldsymbol{\alpha} = \sum_i \lambda_i y_i \boldsymbol{x}_i$, thus the norm vector of the decision boundary is a linear combination of support vectors.

Dual Representation

The decision function can also be rewritten using the duality:

$$f(\boldsymbol{x}) = \boldsymbol{\alpha}^T \boldsymbol{x} + \alpha_0 = \sum_{i=1}^{m} \lambda_i y_i \boldsymbol{x}_i^T \boldsymbol{x} + \alpha_0$$

Dual Representation

The decision function can also be rewritten using the duality:

$$f(\boldsymbol{x}) = \boldsymbol{\alpha}^T \boldsymbol{x} + \alpha_0 = \sum_{i=1}^{m} \lambda_i y_i \boldsymbol{x}_i^T \boldsymbol{x} + \alpha_0$$

Conclusion:

Feature vectors only appear in inner products, both in the learning and the classification phase.

Feature Transforms

Linear decision boundaries in its current form have serious limitations:

- Non-linearly separable data cannot be classified.
- Noisy data cause problems.
- Formulation deals with vectorial data only.

Feature Transforms

Linear decision boundaries in its current form have serious limitations:

- Non-linearly separable data cannot be classified.
- Noisy data cause problems.
- Formulation deals with vectorial data only.

Possible solution:

- Map data into richer feature space using non-linear feature transform, then use a linear classifier.

Feature Transforms (cont.)

We select a feature transform

$$\phi : \mathbb{R}^d \rightarrow \mathbb{R}^D, \quad D \geq d$$

such that the resulting features

$$\phi(\boldsymbol{x}_i), \quad i = 1, 2, \ldots, m$$

are linearly separable.

Feature Transforms (cont.)

Example

Assume the decision boundary is given by the quadratic function

$$f(\boldsymbol{x}) = a_0 + a_1 x_1^2 + a_2 x_2^2 + a_3 x_1 x_2 + a_4 x_1 + a_5 x_2 .$$

Obviously this is not a linear decision boundary.

Feature Transforms (cont.)

Example

Assume the decision boundary is given by the quadratic function

$$f(\boldsymbol{x}) = a_0 + a_1 x_1^2 + a_2 x_2^2 + a_3 x_1 x_2 + a_4 x_1 + a_5 x_2 .$$

Obviously this is not a linear decision boundary.

By the following mapping, we get features that have a linear decision boundary:

$$\phi(\boldsymbol{x}) = \begin{pmatrix} 1 \\ x_1^2 \\ x_2^2 \\ x_1 x_2 \\ x_1 \\ x_2 \end{pmatrix}$$

Feature Transforms (cont.)

These feature transforms can be easily incorporated into SVMs:

- Decision boundary:

$$f(\boldsymbol{x}) = \sum_i \lambda_i y_i \cdot \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}) \rangle + \alpha_0$$

Feature Transforms (cont.)

These feature transforms can be easily incorporated into SVMs:

- Decision boundary:

$$f(\boldsymbol{x}) = \sum_i \lambda_i y_i \cdot \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}) \rangle + \alpha_0$$

- The Lagrange dual problem is given by the optimization problem:

$$\text{maximize} \quad -\frac{1}{2} \sum_i \sum_j \lambda_i \lambda_j y_i y_j \cdot \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle + \sum_i \lambda_i$$

$$\text{subject to} \quad \boldsymbol{\lambda} \succeq 0, \quad \sum_i \lambda_i y_i = 0$$

Kernel Functions

We now define kernel functions:

$$k(\boldsymbol{x}, \boldsymbol{x}') = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}') \rangle$$

Kernel Functions

We now define kernel functions:

$$k(\boldsymbol{x}, \boldsymbol{x}') = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}') \rangle$$

Typical kernel functions are:

- Linear:
$$k(\boldsymbol{x}, \boldsymbol{x}') = \langle \boldsymbol{x}, \boldsymbol{x}' \rangle$$
- Polynomial:
$$k(\boldsymbol{x}, \boldsymbol{x}') = (\langle \boldsymbol{x}, \boldsymbol{x}' \rangle + 1)^k$$
- Radial basis function:
$$k(\boldsymbol{x}, \boldsymbol{x}') = e^{-\gamma \|\boldsymbol{x} - \boldsymbol{x}'\|_2^2}$$
- Sigmoid kernel:
$$k(\boldsymbol{x}, \boldsymbol{x}') = \tanh(\alpha \langle \boldsymbol{x}, \boldsymbol{x}' \rangle + \beta)$$

Lessons Learned

- Lagrangian formulation of the hard and soft margin problems
- Langrange dual representation
- Idea of feature transforms

Pattern Recognition Lab

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

TECHNISCHE FAKULTÄT

Next Time in

Pattern Recognition

Further Readings

- Bernhard Schölkopf, Alexander J. Smola:
  Learning with Kernels,
  The MIT Press, Cambridge, 2003.
- Vladimir N. Vapnik:
  The Nature of Statistical Learning Theory,
  Information Science and Statistics, Springer, Heidelberg, 2000.
- S. Boyd, L. Vandenberghe:
  Convex Optimization,
  Cambridge University Press, 2004.
  ☞ http://www.stanford.edu/~boyd/cvxbook/
- Christopher M. Bishop:
  Pattern Recognition and Machine Learning,
  Springer, New York, 2006

Comprehensive Questions

- What is the Lagrangian of the hard margin SVM?
- What are the KKT optimality conditions for the hard margin SVM?
- How do we apply the KKT conditions to the Lagrange Dual?
- What can we conclude from this reformulated Lagrange Dual?