An Introduction to Machine Learning
with Kernels
Lecture 4

Alexander J. Smola
Alex.Smola@nicta.com.au

Statistical Machine Learning Program
National ICT Australia, Canberra
Day 1

**Machine learning and probability theory**
Introduction to pattern recognition, classification, regression, novelty detection, probability theory, Bayes rule, inference

**Density estimation and Parzen windows**
Kernels and density estimation, Silverman’s rule, Watson Nadaraya estimator, crossvalidation

**Perceptron and kernels**
Hebb’s rule, perceptron algorithm, convergence, feature maps, kernel trick, examples

**Support Vector classification**
Geometrical view, dual problem, convex optimization, kernels and SVM
Support Vector Machine
- Problem definition
- Geometrical picture
- Optimization problem

Optimization Problem
- Hard margin
- Convexity
- Dual problem
- Soft margin problem
Classification

Data
Pairs of observations \((x_i, y_i)\) generated from some distribution \(P(x, y)\), e.g., (blood status, cancer), (credit transaction, fraud), (profile of jet engine, defect)

Task
- Estimate \(y\) given \(x\) at a new location.
- Modification: find a function \(f(x)\) that does the task.
One to rule them all . . .
Optimal Separating Hyperplane

\[ \{ x \mid \langle w, x \rangle + b = -1 \} \]

\[ \{ x \mid \langle w, x \rangle + b = +1 \} \]

\[ y_i = -1 \]

\[ x_1 \]

\[ x_2 \]

\[ w \]

\[ \langle w, (x_1 - x_2) \rangle = 2 \]

\[ \Rightarrow \frac{w}{\|w\|}, (x_1 - x_2) = \frac{2}{\|w\|} \]

Note:
Optimization Problem

**Margin to Norm**
- Separation of sets is given by $\frac{2}{\|w\|}$ so maximize that.
- Equivalently minimize $\frac{1}{2}||w||$.
- Equivalently minimize $\frac{1}{2}||w||^2$.

**Constraints**
- Separation with margin, i.e.
  \[
  \langle w, x_i \rangle + b \geq 1 \quad \text{if } y_i = 1 \\
  \langle w, x_i \rangle + b \leq -1 \quad \text{if } y_i = -1
  \]
- Equivalent constraint
  \[
  y_i(\langle w, x_i \rangle + b) \geq 1
  \]
Mathematical Programming Setting

Combining the above requirements we obtain

$$\minimize \frac{1}{2} \|w\|^2$$

subject to

$$y_i (\langle w, x_i \rangle + b) - 1 \geq 0 \text{ for all } 1 \leq i \leq m$$

Properties

- Problem is convex
- Hence it has unique minimum
- Efficient algorithms for solving it exist
Lagrange Function

Objective Function
We have $\frac{1}{2}\|w\|^2$.

Constraints

\[ c_i(w, b) := 1 - y_i(\langle w, x_i \rangle + b) \leq 0 \]

Lagrange Function

\[ L(w, b, \alpha) = \text{PrimalObjective} + \sum_i \alpha_i c_i \]

\[ = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{m} \alpha_i (1 - y_i(\langle w, x_i \rangle + b)) \]

Saddle Point Condition
Partial derivatives of $L$ with respect to $w$ and $b$ need to vanish.
Solving the Equations

**Lagrange Function**

\[
L(w, b, \alpha) = \frac{1}{2} \|w\|^2 + \sum_{i=1}^{m} \alpha_i (1 - y_i (\langle w, x_i \rangle + b))
\]

**Saddlepoint condition**

\[
\partial_w L(w, b, \alpha) = w - \sum_{i=1}^{m} \alpha_i y_i x_i = 0 \iff w = \sum_{i=1}^{m} \alpha_i y_i x_i
\]

\[
\partial_b L(w, b, \alpha) = - \sum_{i=1}^{m} \alpha_i y_i x_i = 0 \iff \sum_{i=1}^{m} \alpha_i y_i = 0
\]

To obtain the dual optimization problem we have to substitute the values of \(w\) and \(b\) into \(L\). Note that the dual variables \(\alpha_i\) have the constraint \(\alpha_i \geq 0\).
Dual Optimization Problem

After substituting in terms for $b, w$ the Lagrange function becomes

$$-\frac{1}{2} \sum_{i,j=1}^{m} y_i y_j \langle x_i, x_j \rangle + \sum_{i=1}^{m} \alpha_i$$

subject to $\sum_{i=1}^{m} \alpha_i y_i = 0$ and $\alpha_i \geq 0$ for all $1 \leq i \leq m$

Practical Modification

Need to **maximize** dual objective function. Rewrite as

$$\text{minimize} \quad \frac{1}{2} \sum_{i,j=1}^{m} y_i y_j \langle x_i, x_j \rangle - \sum_{i=1}^{m} \alpha_i$$

subject to the above constraints.
Support Vector Expansion

Solution in
\[ w = \sum_{i=1}^{m} \alpha_i y_i x_i \]

\( w \) is given by a linear combination of training patterns \( x_i \). **Independent of the dimensionality of \( x \).**

\( w \) depends on the Lagrange multipliers \( \alpha_i \).

**Kuhn-Tucker-Conditions**

At optimal solution Constraint \( \cdot \) Lagrange Multiplier = 0

In our context this means

\[ \alpha_i (1 - y_i (\langle w, x_i \rangle + b)) = 0. \]

Equivalently we have

\[ \alpha_i \neq 0 \iff y_i (\langle w, x_i \rangle + b) = 1. \]

Only points at the decision boundary can contribute to the solution.
Linear Classification
- Many solutions
- Optimal separating hyperplane
- Optimization problem

Support Vector Machines
- Quadratic problem
- Lagrange function
- Dual problem

Interpretation
- Dual variables and SVs
- SV expansion
- Hard margin and infinite weights
Nonlinearity via Feature Maps

Replace $x_i$ by $\Phi(x_i)$ in the optimization problem.

Equivalent optimization problem

$$
\begin{align*}
\text{minimize} & \quad \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j k(x_i, x_j) - \sum_{i=1}^{m} \alpha_i \\
\text{subject to} & \quad \sum_{i=1}^{m} \alpha_i y_i = 0 \text{ and } \alpha_i \geq 0 \text{ for all } 1 \leq i \leq m
\end{align*}
$$

Decision Function

From $w = \sum_{i=1}^{m} \alpha_i y_i \Phi(x_i)$ we conclude

$$f(x) = \langle w, \Phi(x) \rangle + b = \sum_{i=1}^{m} \alpha_i y_i k(x_i, x) + b.$$
Examples and Problems

**Advantage**
Works well when the data is noise free.

**Problem**
Already a single wrong observation can ruin everything — we require $y_i f(x_i) \geq 1$ for all $i$.

**Idea**
Limit the influence of individual observations by making the constraints less stringent (introduce slacks).
Optimization Problem (Soft Margin)

Recall: Hard Margin Problem

\[
\text{minimize} \quad \frac{1}{2} \|w\|^2 \\
\text{subject to} \quad y_i(\langle w, x_i \rangle + b) - 1 \geq 0
\]

Softening the Constraints

\[
\text{minimize} \quad \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{m} \xi_i \\
\text{subject to} \quad y_i(\langle w, x_i \rangle + b) - 1 + \xi_i \geq 0 \text{ and } \xi_i \geq 0
\]
Linear SVM $C = 1$
Linear SVM $C = 2$
Linear SVM $C = 5$
Linear SVM $C = 10$
Linear SVM $C = 20$
Linear SVM $C = 50$
Linear SVM $C = 100$
Linear SVM $C = 2$
Linear SVM $C = 10$
Linear SVM $C = 20$
Linear SVM $C = 50$
Linear SVM $C = 100$
Linear SVM $C = 1$
Linear SVM $C = 2$
Linear SVM $C = 5$
Linear SVM $C = 10$
Linear SVM $C = 20$
Linear SVM $C = 50$
Linear SVM $C = 100$
Linear SVM $C = 1$
Linear SVM $C = 2$
Linear SVM $C = 5$
Linear SVM $C = 10$
Linear SVM $C = 20$
Linear SVM $C = 50$
Linear SVM $C = 100$
Changing $C$

- For clean data $C$ doesn’t matter much.
- For noisy data, large $C$ leads to narrow margin (SVM tries to do a good job at separating, even though it isn’t possible)

Noisy data

- Clean data has few support vectors
- Noisy data leads to data in the margins
- More support vectors for noisy data
Lagrange Function

We have \( m \) more constraints, namely those on the \( \xi_i \), for which we will use \( \eta_i \) as Lagrange multipliers.

\[
L(w, b, \xi, \alpha, \eta) = \frac{1}{2}||w||^2 + C \sum_{i=1}^{m} \xi_i + \sum_{i=1}^{m} \alpha_i \left(1 - \xi_i - y_i \langle w, x_i \rangle + b \right)
\]

Saddle Point Conditions

\[
\partial_w L(w, b, \xi, \alpha, \eta) = w - \sum_{i=1}^{m} \alpha_i y_i x_i = 0 \iff w = \sum_{i=1}^{m} \alpha_i y_i x_i.
\]

\[
\partial_b L(w, b, \xi, \alpha, \eta) = \sum_{i=1}^{m} -\alpha_i y_i = 0 \iff \sum_{i=1}^{m} \alpha_i y_i = 0.
\]

\[
C - \alpha_i - \eta_i = 0 \iff \alpha_i \in [0, C]
\]
Dual Optimization Problem

Optimization Problem

minimize \[ \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j k(x_i, x_j) - \sum_{i=1}^{m} \alpha_i \]

subject to \[ \sum_{i=1}^{m} \alpha_i y_i = 0 \] and \[ C \geq \alpha_i \geq 0 \] for all \[ 1 \leq i \leq m \]

Interpretation

- Almost same optimization problem as before
- Constraint on weight of each \( \alpha_i \) (bounds influence of pattern).
- Efficient solvers exist (more about that tomorrow).
SV Classification Machine

output \( \sigma \left( \sum v_i k(x, x_i) \right) \)

weights

dot product \( \langle \Phi(x), \Phi(x_i) \rangle = k(x, x_i) \)

mapped vectors \( \Phi(x_i), \Phi(x) \)

support vectors \( x_1 \ldots x_n \)

test vector \( x \)
Gaussian RBF with $C = 1$
Gaussian RBF with $C = 2$
Gaussian RBF with $C = 5$
Gaussian RBF with $C = 10$
Gaussian RBF with $C = 20$
Gaussian RBF with $C = 50$
Gaussian RBF with $C = 100$
Gaussian RBF with $C = 1$
Gaussian RBF with $C = 2$
Gaussian RBF with $C = 5$
Gaussian RBF with $C = 10$
Gaussian RBF with $C = 20$
Gaussian RBF with $C = 50$
Gaussian RBF with $C = 100$
Gaussian RBF with $C = 1$
Gaussian RBF with $C = 2$
Gaussian RBF with $C = 5$
Gaussian RBF with $C = 10$
Gaussian RBF with $C = 20$
Gaussian RBF with $C = 50$
Gaussian RBF with $C = 100$
Gaussian RBF with $C = 1$
Gaussian RBF with $C = 2$
Gaussian RBF with $C = 5$
Gaussian RBF with $C = 10$
Gaussian RBF with $C = 20$
Gaussian RBF with $C = 50$
Gaussian RBF with \( C = 100 \)
Insights

Changing $C$

- For clean data $C$ doesn’t matter much.
- For noisy data, large $C$ leads to more complicated margin (SVM tries to do a good job at separating, even though it isn’t possible)
- Overfitting for large $C$

Noisy data

- Clean data has few support vectors
- Noisy data leads to data in the margins
- More support vectors for noisy data
Gaussian RBF with $\sigma = 1$
Gaussian RBF with $\sigma = 2$
Gaussian RBF with $\sigma = 5$
Gaussian RBF with $\sigma = 10$
Gaussian RBF with $\sigma = 1$
Gaussian RBF with $\sigma = 2$
Gaussian RBF with $\sigma = 5$
Gaussian RBF with $\sigma = 10$
Gaussian RBF with $\sigma = 1$
Gaussian RBF with $\sigma = 2$
Gaussian RBF with $\sigma = 5$
Gaussian RBF with $\sigma = 10$
Gaussian RBF with $\sigma = 1$
Gaussian RBF with $\sigma = 2$
Gaussian RBF with $\sigma = 5$
Gaussian RBF with $\sigma = 10$
Insights

Changing $\sigma$

- For clean data $\sigma$ doesn’t matter much.
- For noisy data, small $\sigma$ leads to more complicated margin (SVM tries to do a good job at separating, even though it isn’t possible)
- Lots of overfitting for small $\sigma$

Noisy data

- Clean data has few support vectors
- Noisy data leads to data in the margins
- More support vectors for noisy data
Summary

Support Vector Machine
- Problem definition
- Geometrical picture
- Optimization problem

Optimization Problem
- Hard margin
- Convexity
- Dual problem
- Soft margin problem
Today’s Summary

**Machine learning and probability theory**
- Introduction to pattern recognition, classification, regression, novelty detection, probability theory, Bayes rule, inference

**Density estimation and Parzen windows**
- Kernels and density estimation, Silverman’s rule, Watson Nadaraya estimator, crossvalidation

**Perceptron and kernels**
- Hebb’s rule, perceptron algorithm, convergence, feature maps, kernel trick, examples

**Support Vector classification**
- Geometrical view, dual problem, convex optimization, kernels and SVM
We are hiring. For details contact
Alex.Smola@nicta.com.au (http://www.nicta.com.au)

Positions

- PhD scholarships
- Postdoctoral positions, Senior researchers
- Long-term visitors (sabbaticals etc.)

More details on kernels

http://www.kernel-machines.org
http://www.learning-with-kernels.org

Schölkopf and Smola: Learning with Kernels

Machine Learning Summer School

http://canberra05.mlss.cc
MLSS’05 Canberra, Australia, 23/1-5/2/2005