Invariance in Kernel Methods - Distance and Integration Kernels

PASCAL Workshop on Machine Learning, SVM and Large Scale Optimization 2005
Dipl.-Math. Bernard Haasdonk
University of Freiburg, Germany
Chair of Pattern Recognition and Image Processing

Abstract
In pattern analysis with kernel methods, it is widely accepted, that
the kernel function is the main ingredient to represent any kind
of prior knowledge. A frequently observed kind of prior
knowledge is invariance with respect to transformations of single
patterns. For this reason, we present two generic methods
of incorporating such knowledge into a kernel learning
framework. The first is based on invariant distances, the second on
invariant integration. As the former method can not guarantee to result in
positive definite kernels, we discuss the use of indefinite kernels
in machine learning. Application on several classification
problems with SVMs demonstrates the competitiveness in terms
of recognition accuracy and computational complexity with
existing methods.

Overview
- Motivation and Notions
  - Kernel Methods and Invariance
- Invariant Distance Substitution Kernels
  - Distance Substitution Kernels
  - Tangent Distance Kernels
  - Application Raman-Microspectroscopy
- Transformation Integration Kernels
  - Application OCR
- Indefinite Kernels in SVM
  - Interpretation in $p$ spaces
- Summary and Perspectives

Kernell Methods

![Kernell Methods Diagram]

- Typical tasks:
  - Classification, Regression, Clustering, Novelty Detection, ...
  - Multitude of kernel methods: SVM, SVR, KPCA, ...
- Analysis chain:
  - kernel choice
  - kernel matrices
  - model function

Importance of Prior Knowledge

- Without prior assumptions no generalization possible
  - There is no best classification method ("No Free Lunch Theorem")
  - There is no best feature representation ("Jugly Duckling Theorem")
- Equivalence of training data and model complexity

- Consequences of modelling more prior knowledge:
  - less training samples required
  - better generalisation in case of equal number of samples
  - feature representation or classification can be simplified
Invariance as Prior Knowledge

- Transformation knowledge
  - i.e. samples maintain inherent meaning under certain transformations
  - continuous and discrete
  - continuous and discrete

Transformation Examples

- Complete group
  - \( \mathcal{G} \cong \mathbb{R} \rightarrow \mathbb{R} \)
  - \( \mathbb{N} \rightarrow \mathbb{N} \rightarrow \mathbb{Z} \)
  - \( \mathbb{M} \rightarrow \mathbb{M} \rightarrow \mathbb{W} \)
- Reversible
- Inversible

Transformation Formalization

- Patterns \( x \) stem from pattern space \( \mathcal{X} \)
- Transformations \( t: \mathcal{X} \rightarrow \mathcal{X} \)
- Set of transformations \( \mathcal{T} \)
- Set of transformed patterns of \( x \)
  - \( \mathcal{T}_x = \{ t(x) \mid t \in \mathcal{T} \} \)
  - “similar” meaning as \( x \)
- Total invariance: \( k(x, x') = k(t(x), t(x')) \)

Specific Goals

- Wanted properties for general approach:
  - Support for various kernel methods, not only SVM
  - Support for many kernels, not only Gaussian
  - Support for infinite set of transformations, exponentially many transformation combinations
  - Support for discrete, continuous, group and non-group transformations
  - Applicability, good memory, time and model complexities
  - Adjustability of degree of invariance

Invariant Distances

- Distances of \( \mathcal{T}_x, \mathcal{T}_{x'} \) better than \( d(x, x') \)

<table>
<thead>
<tr>
<th>( \lambda = 0 )</th>
<th>( \lambda = \infty )</th>
</tr>
</thead>
</table>

- Invariant Distances
  - Two-sided distance
    \( d_{\Omega}(x, x') = \min \{ d(t(x), t'(x')) + \lambda \Omega(t, t') \} \)
  - Cost function
    \( \Omega(t, t') \geq 0 \quad \Omega(t, t') = 0 \iff t = t' = \text{id} \)
  - \( \lambda \) is regularization parameter
  - Similar: one-sided distance + symmetry term
**Distance Substitution Kernels**

- Distance $d$: symmetric, nonnegative, zero-diagonal
- Distance-based kernels:
  - Inner-product case: $k(x, x') = k(x, x') = k(d(x, x'))$
  - where $O$ is an arbitrary origin and $\{x, x\} = \frac{1}{2}d(x, O) - d(x, O')$
- Examples:
  - $d^2(x, x') = x^2 + y^2 - dx, dy$
  - $d^2(x, x') = x^2 + y^2 - dx, dy$
- Empirical observations:
  - Similar behaviour as standard kernels

**Positive Definiteness**

- Equivalent conditions:
  - $d$ is Hilbertian metric
  - $k^{\beta}$ is pd for all $\beta \in [0, 2]$
  - $k^{\beta}$ is pd for all $\gamma \geq 0$
  - $k^{\beta}$ is pd for all $\gamma \geq 0$, $p \in \mathbb{N}$
- Counterexample by violating inequality
- Empirical observations:
  - Weak positive spin on real world data

**Tangent Distance Kernels**

- Assumption: differentiable transformations
- Tangent Distance: local linear approximation
- Distance of tangent spaces instead of manifolds
- Examples:
  - $d^2(x, x') = x^2 + y^2 - dx, dy$
  - $d^2(x, x') = x^2 + y^2 - dx, dy$
- Invariance in 2D
  - Scaling:
    - $d^2(x, x')$ varying $x$
  - Rotation:
    - $d^2(x, x')$ varying $x$
- More than linear invariances can be captured

**Application Raman Spectra**

- Project OMB in BMBF framework "Biospectroscopy"
- Detection of clean room contamination by Raman microscopy
- Raman spectroscopy

- Use of the "Raman effect", inelastic scattering of light
- Distribution of energy differences $\rightarrow$ Raman spectrum
- Non-destructive method, fingerprint of chemical bindings

- Pattern variations due to:
  - Measurement duration, background radiation
  - Thickness of sample, heterogeneity, growth time, nutrition conditions, temperature, photo bleaching

**Application Raman Spectra**

- Details on dataset (man1):
  - 2546 spectra, 1833 features
  - 20 classes
  - Uniqueness distributed

- Easy to classify
- Difficult to classify

---
Application Raman Spectra

Details on tangents:
- intensity scale: scale tangent 
- baseline shift: 

\[ P_r(x) = \prod_{m=0}^{l-3} \frac{1}{\prod_{m=0}^{l-3}} \frac{1}{(x-j)^2} \]

m = 0, 1, 2, 3

1 = 0, 1, 2, 3

Application Raman Spectra

Recognition results:
- LIBSVM L1, one-against-SVM, \( k_{ij} \)
- Simultaneous scaling and Lagrange tangents
- KdV grid-search for \((C, \gamma)\)
- L0 and class-wise average L0 minimization

| L0 norm \( ||v||_0 \) | L0 norm/\( ||v||_0 \) norm/\( ||v||_0 \) | L0 norm/\( ||v||_0 \) norm/\( ||v||_0 \) |
|-----------------|-----------------|-----------------|
| 0.000000 | 0.000000 | 0.000000 |
| 0.000999 | 0.000999 | 0.000999 |
| 0.001998 | 0.001998 | 0.001998 |
| 0.002997 | 0.002997 | 0.002997 |
| 0.003995 | 0.003995 | 0.003995 |
| 0.004994 | 0.004994 | 0.004994 |

Transformation Integration Kernels

Motivation:
- Maintain additivity by avoiding min/max-operations
- Extend Near-Integration framework w.r.t. kernels
- Adjustability of degree of invariance

Definition of TI-Kernels (w.r.t.):

\[ k_{ij}(x, x') = \int_0^1 \Phi(\tau(x)) d\tau \Phi(\tau(x')) d\tau = \int_0^1 k_{ij}(x, \tau(x')) d\tau \]

Transformation Integration Kernels

Invariance in 2D

- Invariance adjustability: highly nonlinear sinus shifts
- Effect in SVM: rotation
- Similar: Total invariance of linear, polynomial kernels

Acceleration

- Single Kernel Evaluation: Integral Reduction (IR)
  - If \( T \) is invertible and compatible with \( k \)
    \[ \int_0^1 k(x, \tau(x')) d\tau = \int_0^1 k(x, \tau(t')) d\tau \]
  - => square-root complexity reduction
- SVM-Training, SV-extraction (SV)
  - Perform initial ordinary SVM-training
  - Extract SVs
  - Train invariant SVM on the SVs
  - => linear complexity reduction
Application USPS-Digits

- Details on dataset:
  - standard benchmark dataset
  - 7291 training, 2007 test patterns of handwritten digits given by 16x16 gray-value bitmaps
  - samples:
    - easy to classify
    - difficult to classify

![Classification Examples](image)

- SVM Recognition Results:
  - increasing rotation integration range
  - increasing xy translation integration range
  - both TI-kernels superior over base kernel
  - xy-integration better than rotations
  - TI-kernels yield state-of-the-art results as VSV

![Recognition Results](image)

Application USPS-Digits

- Complexity comparison to VSV:
  - shared non-optimized parameters,
    - xy-translations, ±-2-pixels, 3.8 shifts per sample

<table>
<thead>
<tr>
<th>Method</th>
<th>test error [%]</th>
<th>test time [sec]</th>
<th>train time [sec]</th>
<th>average [sec]</th>
</tr>
</thead>
<tbody>
<tr>
<td>TI-SVM</td>
<td>3.6</td>
<td>1171</td>
<td>226</td>
<td>444</td>
</tr>
<tr>
<td>TI-SVM, IB</td>
<td>3.6</td>
<td>130</td>
<td>16</td>
<td>412</td>
</tr>
<tr>
<td>TI-SVM, SW+IB</td>
<td>3.6</td>
<td>146</td>
<td>190</td>
<td>440</td>
</tr>
<tr>
<td>VSV-SVM</td>
<td>3.8</td>
<td>146</td>
<td>190</td>
<td>442</td>
</tr>
</tbody>
</table>

- TI-kernels produce small models
- no recognition degradation from acceleration
- Acceleration techniques largely successful
- Accelerated TI-SVM consistently faster than VSV

Pseudo-Euclidean Spaces (PE)

- Real finite dimensional vector spaces \( \mathbb{R}^{d,0} = \mathbb{R}^d \times \mathbb{R}^d \) of signature \((p,q)\)
- symmetric (indefinite) inner product
  \( (x,z)_p = x^T z^T - x^T z = z^T M z \quad M = \text{diag}(1_p, -1_q) \)
- squared norm
  \( \|x\|_p^2 = (x,x)_p = x^T M x \)
- squared distance
  \( d(x,z)_p = (x-z,x-z)_p \)
- orthogonality
  \( x^T M z = x^T M^T z = 0 \)
- hyperplanes
  \( \forall c: \langle x, w \rangle + b = 0 \)

pE Feature Space Embedding

- Data dependent pE-embedding:
  - Given data \( \{x_i\}_{i=1}^n \rightarrow \) existence of pE space \( \mathbb{R}^{d,0} \)
  - + sym. kernel \( \kappa \)
  - + embedding \( \Phi : \mathcal{X} \rightarrow \mathbb{R}^{d,0} \)

- Representation of kernel
  \( k(x, z) = \langle \Phi(x), \Phi(z) \rangle \)

- Construction by Eigendecomposition
  \( \Phi(x) : = \sqrt{\lambda} \sum_{i=1}^n \phi_i(x) \)
  \( A = \text{diag} \{ \lambda_i \} \)
  \( \kappa = \sum_{i,j} \phi_i \phi_j \)

Indefinite Kernels in SVM

![Indefinite Kernels](image)
Geometrical Interpretation of Kernel Operations

- Norm of feature vector $k(x, x) = |\Phi(x)|_2$
- Kernel induced distance $d^2(x, x) = k(x, x) - 2k(x, x) + k(x, x)$
- Linear combinations $\sum \alpha_k \Phi(x, x) = \sum \alpha_k \Phi(x, x) \cdot \mu$
- Centering of Kernel Matrix $K = \frac{1}{n} K + 1$ $\Rightarrow$ $\sum \Phi_i(x) = 0$
- Projections, variance analysis, Eigendecompositions, Optimization problems... general kernel methods?

Optimal Separation of Convex Hulls

- Formalization train: $\min_{x^*} \frac{1}{n} x^* T \Phi(x^*)$ classify: sign of $f(x) = \frac{1}{n} x^* T \Phi(x)$
- Separable case
- Non-separable case
- Training and test avoid explicit PE-embedding
- Reasonable solution: $w^T M w \geq 0$
- Optimum exists and obtained in span $\{ \Phi(x) \}$

Capacity Estimate

- VC-bound for PE hyperplanes in $\mathbb{R}^d$
  - data embedded s.th. $\varepsilon |\Phi(x)|_2 \leq |\Phi(x)|_2 \leq R^2$
  - $\mathcal{F} := \{ \text{sign}(w, x) \}$ set of hyperplanes, which are canonical wrt the data and satisfy
  $$\lambda \geq |w|_2 \leq \Lambda^2$$
  then holds $h(\mathcal{F}) \leq \frac{R^2}{2 \lambda}$
- SVM: minimize $|w|_2 = w^T M w$ while maintaining strict positivity

Indefinite SVM in PE Space

- SVM "primal": $\min_{x^*} \frac{1}{n} x^* T \Phi(x)$
  - $\geq 1$
  - $w^T M w \geq 0$ Classifier
  - Only bounded SVM can be misclassified
  - SVM is CH-classifier:
    - Feasible, stationary points + local optima transfer between SVM and CH if $w^T M w \geq 0$
    - E.g. Nonzero SVM-solution $\alpha \Rightarrow$ CH-solution $\alpha$
    - $\alpha = 2w \sum \alpha_i$
    - CH and SVM hyperplanes are parallel, even identical if coefficients are not bounded

Numerics

- Convergence to stationary point, libsvm
- Multiple solutions
- Uniqueness in extreme indefinite cases

Practical Criteria for Indefinite SVM

- Criterion for suitability: $\# \text{SV}$
  - No (few) bounded $\alpha_i \Rightarrow$ (few) training errors
- Criterion for unsuitability: $w^T M w \leq 0$
- after training: $\sum \alpha_i \Phi(x_i) \Phi(x_j)$
- before training: signature (PS)
- Criterion for suitability: Distance of Class Means
  - If DCM is positive, sufficiently low $C$ yields solution

Image References:
- [Image of Geometrical Interpretation of Kernel Operations]
- [Image of Optimal Separation of Convex Hulls]
- [Image of Capacity Estimate]
- [Image of Indefinite SVM in PE Space]
- [Image of Numerics]
- [Image of Practical Criteria for Indefinite SVM]
Summary

- Principle: Incorporation of transformation knowledge in kernels for improving kernel methods
- General framework I: IDS-kernels
  - Distance substitution kernels with invariant distances
  - Application: Tangent Distance Kernels on Spectra
- General framework II: TF-kernels
  - Integration over transformations
  - Application: Offline HWR (USPS digits)
  - Efficient for few invariances, but positive definiteness
- Indefinite SVM
  - Constructive geometric interpretation: pCH-separation
  - Convergent implementation required
  - Practical criteria for suitability

Perspectives

- Invariance
  - Combination with invariant representations
  - Learning of transformation directions
- Kernel design
  - Relating the p-ness of kernels yields much wider flexibility
  - Applications with proximity data: nonlinear analysis!
- Kernel methods
  - Acceptance/robustness against indefinite kernels?
  - Interpretation in p-spaces as basis for geometries/numerical/statistical analysis and new methods
- Non-convex optimization
  - Large scale implementations with convergence statements
  - Number and quality of solutions.

Thank You! 😊

Any questions?