SVM: Algorithms of Choice for Challenging Data

Boriana Milenova, Joseph Yarmus, Marcos Campos
Data Mining Technologies
ORACLE Corp.
Overview

- SVM theoretical framework

- ORACLE data mining technology
  - SVM parameter estimation
  - SVM optimization strategy

- SVM on challenging data
SVM Model Defines a Hyperplane

- Linear models in feature space
- Hyperplane defined by a set of coefficients and a bias term

\[
\mathbf{w} \cdot \mathbf{x} + b = 0
\]
Maximum Margin Models

*Functional margin* = \( \min(y_i f(x_i)) \)

*Geometric margin* = \( \min(\frac{y_i f(x_i)}{||w||}) = \frac{1}{||w||} \)

\[ \min ||w|| \Rightarrow \max(\text{margin}) \]
SVM Optimization Problem

Minimize $||w||$ subject to $y_i f(x_i) \geq 1$

Lagrangian in primal space:

$$L_p(w) = \frac{1}{2} \langle w \cdot w \rangle - \sum \alpha_i \left[ y_i \left( \langle w \cdot x_i \rangle + b \right) - 1 \right]$$

subject to $\alpha_i \geq 0$

$$\frac{\partial L_p}{\partial w} = 0 \quad w = \sum \alpha_i y_i x_i$$

$$\frac{\partial L_p}{\partial b} = 0 \quad \sum \alpha_i y_i = 0$$
Duality

Lagrangian in dual space:

\[
L_D = \sum \alpha_i - \frac{1}{2} \sum \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle
\]

subject to \( \alpha_i \geq 0 \) \( \sum \alpha_i y_i = 0 \)

Dot products!
- dimension-insensitive optimization
- generalized dot products via non-linear map \( \phi \)

\[
K(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) \rangle
\]
Towards Higher Dimensionality via Kernels

1. Transform data via non-linear mapping \( \phi \) to an inner product feature space
2. Train a linear machine in the new feature space

Mercer’s kernels:

- symmetry
  \[ K(x_i, x_j) = K(x_j, x_i) \]
- positive semi-definite kernel matrix
- reproducing property
  \[ \langle K(x_i, \cdot) \cdot K(x_j, \cdot) \rangle = K(x_i, x_j) \]
Soft Margin: Non-Separable Data

$$L_p(w) = \frac{1}{2} \langle w \cdot w \rangle + C \sum \xi^k$$

subject to

$$y_i \left( \langle w \cdot x_i \rangle + b \right) \geq 1 - \xi_i$$

Capacity parameter $C$

trades off complexity and empirical risk
1-Norm Dual Problem

Lagrangian in dual space:

\[ L_D = \sum \alpha_i - \frac{1}{2} \sum \alpha_i \alpha_j y_i y_j K(x_i, x_j) \]

subject to \( 0 \leq \alpha_i \leq C \) \( \sum \alpha_i y_i = 0 \)

Quadratic problem

- linear and inequality constraints
SVM Regression

\[ L_p(w) = \frac{1}{2} \langle w \cdot w \rangle + C \sum (\xi_k + \hat{\xi}_k) \]

subject to

\[ \left( \langle w \cdot x_i \rangle + b \right) - y_i \leq \varepsilon + \xi_i \]

\[ y_i - \left( \langle w \cdot x_i \rangle + b \right) \leq \varepsilon + \hat{\xi}_i \]
SVM Fundamental Properties

- Convexity
  - single global minimum
- Regularization
  - trades off structural and empirical risk to avoid overfitting
- Sparse solution
  - usually only a fraction of training data become support vectors
- Not probabilistic

Solvable in polynomial time...
SVM in the Database

ORACLE Data Mining (ODM)
- commercial SVM implementation in the database
- product targets application developers and data mining practitioners
- focuses on ease of use and efficiency

Challenges:
- effective and inexpensive parameter tuning
- computationally efficient SVM model optimization
**SVM Out-Of-The-Box**

Inexperienced users can get dramatically poor results

**LIBSVM examples:**

<table>
<thead>
<tr>
<th></th>
<th>Out-of-the-box correct rate</th>
<th>After tuning correct rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Astroparticle Physics</td>
<td>0.67</td>
<td>0.97</td>
</tr>
<tr>
<td>Bioinformatics</td>
<td>0.57</td>
<td>0.79</td>
</tr>
<tr>
<td>Vehicle</td>
<td>0.02</td>
<td>0.88</td>
</tr>
</tbody>
</table>
SVM Parameter Tuning

- Grid search (+ cross-validation or generalization error estimates)
  - naive
  - guided (Keerthi & Lin, 2002)
- Parameter optimization
  - gradient descent (Chapelle et al., 2000)
- Heuristics
ODM On-the-Fly Estimates

☑ Standard deviation for Gaussian kernel
  - single kernel parameter
  - kernel has good numeric properties
    ☑ bounded, no overflow

☑ Capacity
  - key to good classification generalization

☑ Epsilon estimate for regression
  - key to good regression generalization
ODM Standard Deviation Estimate

Goal: Estimate distance between classes

3. Pick random pairs from opposite classes
4. Measure distances
5. Order descending
6. Exclude tail (90th percentile)
7. Select minimum distance
ODM Capacity Estimate

Goal: Allocate sufficient capacity to separate typical examples

2. Pick m random examples per class
3. Compute $y_i$ assuming $\alpha = C$
   \[ y_i = \sum_{j=1}^{2m} C y_j K(x_j, x_i) \]
5. Exclude noise (incorrect sign)
6. Scale C, $y_i = \pm 1$ (non bounded sv)
   \[ C = y_i / \sum_{j=1}^{2m} y_j K(x_j, x_i) \]
8. Order descending
9. Exclude tail ($90^{\text{th}}$ percentile)
10. Select minimum value
Some Comparison Numbers

LIBSVM examples:

<table>
<thead>
<tr>
<th></th>
<th>Out-of-the-box</th>
<th>On-the-fly estimates</th>
<th>Grid search + xval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Astroparticle Physics</td>
<td>0.67</td>
<td>0.97</td>
<td>0.97</td>
</tr>
<tr>
<td>Bioinformatics</td>
<td>0.57</td>
<td>0.84</td>
<td>0.85</td>
</tr>
<tr>
<td>Vehicle</td>
<td>0.02</td>
<td>0.71</td>
<td>0.88</td>
</tr>
</tbody>
</table>
Goal: estimate target noise by fitting a preliminary model

3. Pick \( m \) random examples
4. Train SVM model with \( \varepsilon \rightarrow 0 \)
5. Compute residuals on remaining data
6. Scale \( \varepsilon_t = \left( \varepsilon_{t-1} + \sigma_n \right)/2 \)
7. Retrain
## Comparison Numbers Regression

<table>
<thead>
<tr>
<th></th>
<th>On-the-fly estimates RMSE</th>
<th>Grid search RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boston housing</td>
<td>6.57</td>
<td>6.26</td>
</tr>
<tr>
<td>Computer activity</td>
<td>0.35</td>
<td>0.33</td>
</tr>
<tr>
<td>Pumadyn</td>
<td>0.02</td>
<td>0.02</td>
</tr>
</tbody>
</table>
Optimization Approaches

❖ QP solvers
  - MINOS, LOQO, quadprog (Matlab)

❖ Gradient descent methods
  - Sequentially update one $\alpha$ coefficient at a time

❖ Chunking and decomposition
  - Optimize small “working sets” towards global solution
  - Analytic solution possible (SMO - Platt, 1998)
Chunking strategy

/* WS working set */
select initial WS randomly;
while (violations)
{
    Solve QP on WS;
    Select new WS;
}
ODM Working Set Selection

- Avoid oscillations
  - overlap across chunks
  - retain non-bounded support vectors
- Choose among violators
  - add large violators
- Computational efficiency
  - avoid sorting
Who to Retain?

/* Examine previous working set */
if (non-bounded sv < 50%)
{
    retain all non-bounded sv;
    add other randomly selected up to 50%;
}
else
{
    randomly select non-bounded sv;
}
Who to Add?

create violator list;

/* Scan I - pick largest violators */
while (new examples < 50% AND WS Not Full)
{
    if (violation > avg_violation)
        add to WS;
}

/* Scan II - pick other violators */
while (new examples < 50% AND WS Not Full)
{
    add randomly selected violators to WS;
}
SVM in Feed-Forward Framework

\[ y_i = \sum_j \alpha_j y_j K(x_i, x_i) \]
DOF in Neural Nets / RBF
DOF in SVM
SVM vs. Neural Net / RBF

<table>
<thead>
<tr>
<th></th>
<th>SVM</th>
<th>NN / RBF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regularization</td>
<td>✓</td>
<td>—</td>
</tr>
<tr>
<td>Global minimum</td>
<td>✓</td>
<td>—</td>
</tr>
<tr>
<td>Compact model</td>
<td>—</td>
<td>✓</td>
</tr>
</tbody>
</table>
Text Mining

Domain characteristics:
- thousands of features
- hundreds of topics
- sparse data
SVM in Text Mining

Reuters corpus

~10K documents, ~10K terms, 115 classes

Accuracy: recall / precision breakeven point

<table>
<thead>
<tr>
<th>Method</th>
<th>Naive Bayes</th>
<th>Rocchio</th>
<th>C4.5</th>
<th>K-NN</th>
<th>SVM linear</th>
<th>SVM non-linear</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.72</td>
<td>0.80</td>
<td>0.79</td>
<td>0.82</td>
<td>0.84</td>
<td>0.86</td>
</tr>
</tbody>
</table>

Joachims, 1998
Biomining

microarray data

Domain characteristics:
- thousands of features
- very few data points
- dense data
SVM on Microarray Data

Multiple tumor types
144 samples, 16063 genes, 14 classes
Accuracy: correct rate

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive Bayes</td>
<td>0.43</td>
</tr>
<tr>
<td>Weighted voting</td>
<td>0.62</td>
</tr>
<tr>
<td>K-NN</td>
<td>0.68</td>
</tr>
<tr>
<td>SVM linear</td>
<td>0.78</td>
</tr>
</tbody>
</table>

Ramaswamy et al., 2001
Other domains

High dimensionality problems:
- image (color and texture histograms)
- satellite remote sensing
- speech

Linear kernels sufficient in most cases
- data separability
- single parameter tuning (capacity)
- small model size
Final Note

SVM classification and regression algorithms available in ORACLE 10G database

Two APIs

- JAVA (J2EE)
- PL/SQL
References


