Modeling high-frequency limit order book dynamics with support vector machines

Alec Kercheval and Yuan Zhang

Department of Mathematics
Florida State University

October 24, 2013
Outline

1. Support Vector Machines
   - Geometry and Computation via Support Vectors
   - Feature Space, Outliers, and Multiclass

2. Limit Order Book Dynamics and Model Architecture
   - Order Book Data
   - Model Architecture

3. Experimental Results
   - Computation Time
   - Trading Test
People

- Yuan Zhang, PhD student, FSU
- SVM: Burges, Cortes, Vapnik, Crammer, Singer, Joachims, Ng, Platt...
SVMs - Intuition

Figure 1: SVMs Intuition: Finding Optimal Separating Hyperplane
Definition (Geometric Margin)

Given a training example \((\vec{x}_i, y_i)\), \(y_i \in \{-1, +1\}\), the geometric margin of \((\vec{x}_i, y_i)\) with respect to \((\vec{w}, b)\) is

\[
\gamma_i = \frac{y_i}{\|\vec{w}\|} (\vec{w} \cdot \vec{x}_i + b)
\]

(1)

Given a training set \(T = \{ (\vec{x}_i, y_i); i = 1, \ldots, m \}\), the geometric margin of \(T\) with respect to \((\vec{w}, b)\) is

\[
\gamma = \min_{i=1,\ldots,m} \gamma_i
\]

(2)
Optimal Margin Classifier

For a training data set \( T = \{(\vec{x}_i, y_i); i = 1, \ldots, m\} \), let \( \hat{\gamma} = \gamma ||\vec{w}|| \) and solve:

\[
\max_{\hat{\gamma}, \vec{w}, b} \frac{\hat{\gamma}}{||\vec{w}||} = \gamma
\]

subject to:
\[
y_i(\vec{w} \cdot \vec{x}_i + b) \geq \hat{\gamma} = \gamma ||\vec{w}||, \quad i = 1, \ldots, m
\]

Apply scale constraint to (\( \vec{w}, b \)) to set \( \hat{\gamma} = 1 \), and notice that maximizing \( \hat{\gamma}/||\vec{w}|| = 1/||\vec{w}|| \) is equivalent to minimizing \( ||\vec{w}||^2 \):

\[
\min_{\vec{w}, b} \frac{1}{2} ||\vec{w}||^2
\]

subject to:
\[
y_i(\vec{w} \cdot \vec{x}_i + b) \geq 1, \quad i = 1, \ldots, m
\]
Lagrangian and Dual Problem

\[ L(\vec{w}, b, \vec{\alpha}) = \frac{1}{2} \| \vec{w} \|^2 - \sum_{i=1}^{m} \alpha_i[y_i(\vec{w} \cdot \vec{x}_i + b) - 1] \] (5)

\[ \max_{\vec{\alpha}} W(\vec{\alpha}) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j (\vec{x}_i \cdot \vec{x}_j) \]

s.t. \( \alpha_i \geq 0 \quad i = 1, \ldots, m \)

\[ \sum_{i=1}^{m} \alpha_i y_i = 0 \quad i = 1, \ldots, m \] (6)
Optimal Solutions

\[ \vec{w}^* = \sum_{i=1}^{m} \alpha_i^* y_i \vec{x}_i \]  \hfill (7)

\[ b^* = - \frac{\max_{i:y_i=-1} \vec{w}^* \cdot \vec{x}_i + \min_{i:y_i=+1} \vec{w}^* \cdot \vec{x}_i}{2} \]  \hfill (8)
Classification of new data

Compute the sign of

$$\vec{w}^* \cdot \vec{x} + b^* = \left( \sum_{i=1}^{m} \alpha_i^* y_i \vec{x}_i \right) \cdot \vec{x} + b^*$$

$$= \sum_{i=1}^{m} \alpha_i^* y_i (\vec{x}_i \cdot \vec{x}) + b^*$$

By “complementary slackness", we have:

$$\alpha_i^* (y_i (\vec{w}^* \cdot \vec{x}_i + b^*) - 1) = 0, \quad \forall i = 1, \ldots, m$$

so

$$\alpha_i^* > 0 \text{ if and only if } \vec{x}_i \text{ lies on the margin (a Support Vector)}$$
Feature Mappings

\[ \vec{w} \cdot \phi(\vec{x}_i) + b = 0 \]  \hspace{1cm} (11)

\[ \phi: \mathbb{R}^n \mapsto \mathcal{H} \]

\[
\min_{\vec{w}, b} \frac{1}{2} \| \vec{w} \|^2 \\
\text{s.t. } y_i(\vec{w} \cdot \phi(\vec{x}_i) + b) \geq 1, \quad i = 1, \ldots, m
\]  \hspace{1cm} (12)

\[
\max_{\vec{\alpha}} W(\vec{\alpha}) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \phi(\vec{x}_i) \cdot \phi(\vec{x}_j) \\
\text{s.t. } \alpha_i \geq 0 \quad i = 1, \ldots, m \\
\sum_{i=1}^{m} \alpha_i y_i = 0 \quad i = 1, \ldots, m
\]  \hspace{1cm} (13)
Mercer’s Theorem

Note the inner products $\phi(\vec{x}_i) \cdot \phi(\vec{x}_j)$. When is this equal to a simpler $\kappa(\vec{x}_i, \vec{x}_j)$?

Example:

$$\kappa(\vec{x}, \vec{z}) = (\vec{x} \cdot \vec{z} + 1)^2 = \phi(\vec{x}) \cdot \phi(\vec{z})$$

where

$$\phi(\vec{x}) = [x_1x_1, x_1x_2, \ldots, x_nx_n, \sqrt{2}x_1, \ldots, \sqrt{2}x_n, 1].$$

**Theorem (Mercer’s Theorem)**

A function $\kappa : \mathbb{R}^n \times \mathbb{R}^n \mapsto \mathbb{R}$ is symmetric positive semi-definite if and only if there exists a function $\phi : \mathbb{R}^n \mapsto H$ such that $\kappa(\vec{x}, \vec{z}) = \phi(\vec{x}) \cdot \phi(\vec{z})$. 
Kernel Technique in Optimization Problem

New data classified according to:

\[
\hat{w}^* \cdot \phi(\mathbf{x}) + b^* = \left( \sum_{i=1}^{m} \alpha_i^* y_i \phi(\mathbf{x}_i) \right) \cdot \phi(\mathbf{x}) + b^*
\]

\[
= \sum_{i=1}^{m} \alpha_i^* y_i \kappa(\mathbf{x}_i, \mathbf{x}) + b^*
\]
Figure 2: example: linearly inseparable input data mapping into higher feature space
Training Data with Outliers

Figure 3: hyperplane swings with outliers
Support Vector Machines
Limit Order Book Dynamics and Model Architecture
Experimental Results

Soft-margin SVMs

\[ \begin{align*}
\min_{\vec{w}, b, \xi} & \quad \frac{1}{2} \| \vec{w} \|^{2} + C \sum_{i=1}^{m} \xi_i \\
\text{s.t.} & \quad y_i (\vec{w} \cdot \vec{x}_i + b) \geq 1 - \xi_i, \quad i = 1, \ldots, m \\
& \quad \xi_i \geq 0, \quad i = 1, \ldots, m.
\end{align*} \] (15)

And the dual optimization problem with kernel:

\[ \begin{align*}
\max_{\vec{\alpha}} & \quad W(\vec{\alpha}) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \kappa(\vec{x}_i, \vec{x}_j) \\
\text{s.t.} & \quad 0 \leq \alpha_i \leq C \quad i = 1, \ldots, m \\
& \quad \sum_{i=1}^{m} \alpha_i y_i = 0 \quad i = 1, \ldots, m
\end{align*} \] (16)
Multi-class SVMs — Binary Reduction

Solving a multi-class learning problem: use a set of binary classification tasks and build a binary classifier independently for each label $l_1, \ldots, l_k$ with one-against-all or one-against-one methods.

- One-against-all method: construct $k$ binary one-against-all SVM models; Assign the label with the largest value

$$
\sum_{i=1}^{m} \alpha_i^* y_i \kappa(x_i, \tilde{x}) + b^*
$$
Multi-class SVMs — Model Building Examples

Figure 4: one-against-all & one-against-one models training
### Message book & Order Book (AAPL)

**Message book**

<table>
<thead>
<tr>
<th>Time(sec)</th>
<th>Price($)</th>
<th>Volume</th>
<th>Event Type</th>
<th>Direction</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k - 1$</td>
<td>34203.011926972</td>
<td>585.68</td>
<td>18</td>
<td>execution</td>
</tr>
<tr>
<td>$k$</td>
<td>34203.011926973</td>
<td>585.69</td>
<td>16</td>
<td>execution</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$k + 4$</td>
<td>34203.011988208</td>
<td>585.74</td>
<td>18</td>
<td>cancellation</td>
</tr>
<tr>
<td>$k + 5$</td>
<td>34203.011990228</td>
<td>585.75</td>
<td>4</td>
<td>cancellation</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$k + 8$</td>
<td>34203.012050158</td>
<td>585.70</td>
<td>66</td>
<td>execution</td>
</tr>
<tr>
<td>$k + 9$</td>
<td>34203.012287906</td>
<td>585.45</td>
<td>18</td>
<td>submission</td>
</tr>
<tr>
<td>$k + 10$</td>
<td>34203.089491920</td>
<td>586.68</td>
<td>18</td>
<td>submission</td>
</tr>
</tbody>
</table>

**Order book**

<table>
<thead>
<tr>
<th>Ask$^1$</th>
<th>Bid$^1$</th>
<th>Ask$^2$</th>
<th>Bid$^2$</th>
<th>Ask$^3$</th>
<th>Bid$^3$</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k - 1$</td>
<td>585.69</td>
<td>16</td>
<td>585.44</td>
<td>167</td>
<td>585.71</td>
<td>118</td>
</tr>
<tr>
<td>$k$</td>
<td>585.71</td>
<td>118</td>
<td>585.44</td>
<td>167</td>
<td>585.72</td>
<td>2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$k + 4$</td>
<td>585.71</td>
<td>118</td>
<td>585.70</td>
<td>66</td>
<td>585.72</td>
<td>2</td>
</tr>
<tr>
<td>$k + 5$</td>
<td>585.71</td>
<td>118</td>
<td>585.70</td>
<td>66</td>
<td>585.72</td>
<td>2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$k + 8$</td>
<td>585.71</td>
<td>100</td>
<td>585.44</td>
<td>167</td>
<td>585.80</td>
<td>100</td>
</tr>
<tr>
<td>$k + 9$</td>
<td>585.71</td>
<td>100</td>
<td>585.45</td>
<td>18</td>
<td>585.80</td>
<td>100</td>
</tr>
<tr>
<td>$k + 10$</td>
<td>585.68</td>
<td>18</td>
<td>585.45</td>
<td>18</td>
<td>585.71</td>
<td>100</td>
</tr>
</tbody>
</table>
Dynamic Metrics

- spread crossing metric:

<table>
<thead>
<tr>
<th>Scenarios</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P^{bid}<em>{t+\Delta t} &gt; P^{ask}</em>{t}$</td>
<td>upward spread crossing</td>
</tr>
<tr>
<td>$P^{ask}<em>{t+\Delta t} &lt; P^{bid}</em>{t}$</td>
<td>downward spread crossing</td>
</tr>
<tr>
<td>$P^{ask}<em>{t+\Delta t} \geq P^{bid}</em>{t}$ &amp; $P^{bid}<em>{t+\Delta t} \leq P^{ask}</em>{t}$</td>
<td>stationary, no crossing</td>
</tr>
</tbody>
</table>

Table: spread crossing outcomes
## Data Attributes

<table>
<thead>
<tr>
<th>Basic Set</th>
<th>Description $i = \text{level index}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$v_1 = { P_i^{\text{ask}}, V_i^{\text{ask}}, P_i^{\text{bid}}, V_i^{\text{bid}} }_{i=1}^n,$</td>
<td>price and volume ($n$ levels, $n=10$)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time-insensitive Set</th>
<th>Description $i = \text{level index}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$v_2 = { (P_i^{\text{ask}} - P_i^{\text{bid}}), (P_i^{\text{ask}} + P_i^{\text{bid}})/2 }_{i=1}^n,$</td>
<td>bid-ask spreads and mid-prices</td>
</tr>
<tr>
<td>$v_3 = { P_n^{\text{ask}} - P_1^{\text{ask}}, P_1^{\text{bid}} - P_n^{\text{bid}} },$</td>
<td>max-min price differences</td>
</tr>
<tr>
<td>$v_4 = {</td>
<td>P_{i+1}^{\text{ask}} - P_i^{\text{ask}}</td>
</tr>
<tr>
<td>$v_5 = { 1/n \sum_{i=1}^n P_i^{\text{ask}}, 1/n \sum_{i=1}^n P_i^{\text{bid}}, 1/n \sum_{i=1}^n V_i^{\text{ask}}, 1/n \sum_{i=1}^n V_i^{\text{bid}} },$</td>
<td>mean prices and volumes</td>
</tr>
<tr>
<td>$v_6 = { \sum_{i=1}^n (P_i^{\text{ask}} - P_i^{\text{bid}}), \sum_{i=1}^n (V_i^{\text{ask}} - V_i^{\text{bid}}) },$</td>
<td>accumulated differences</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time-sensitive Set</th>
<th>Description $i = \text{level index}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$v_7 = { dP_i^{\text{ask}}/dt, dP_i^{\text{bid}}/dt, dV_i^{\text{ask}}/dt, dV_i^{\text{bid}}/dt }_{i=1}^n,$</td>
<td>price and volume derivatives</td>
</tr>
<tr>
<td>$v_8 = { \lambda_{la}^{\Delta t}, \lambda_{lb}^{\Delta t}, \lambda_{ma}^{\Delta t}, \lambda_{mb}^{\Delta t}, \lambda_{ca}^{\Delta t}, \lambda_{cb}^{\Delta t} }$</td>
<td>average intensity of each type</td>
</tr>
<tr>
<td>$v_9 = { 1 { \lambda_{la}^{\Delta t} &gt; \lambda_{la}^{\Delta t} }, 1 { \lambda_{lb}^{\Delta t} &gt; \lambda_{lb}^{\Delta t} }, 1 { \lambda_{ma}^{\Delta t} &gt; \lambda_{ma}^{\Delta t} }, 1 { \lambda_{mb}^{\Delta t} &gt; \lambda_{mb}^{\Delta t} } }$,</td>
<td>relative intensity indicators</td>
</tr>
<tr>
<td>$v_{10} = { d\lambda_{ma}/dt, d\lambda_{lb}/dt, d\lambda_{mb}/dt, d\lambda_{la}/dt }$,</td>
<td>accelerations (market/limit)</td>
</tr>
</tbody>
</table>
Attribute refinement via Information Gain

We form an “economical” set of attributes by selecting a subset ranked by Information Gain = entropy reduction.

The information gain of \( Y \) contributed by \( X \):

\[
IG(X) = H(Y) - H(Y|X)
\]  (17)
Entropy and Conditional Entropy

Given the training set with labels in $Y$, the entropy of $Y$ is:

$$H(Y) = - \sum_{y \in Y} p(y) \log_2(p(y))$$  \hspace{1cm} (18)$$

The conditional entropy of $Y$ after observing the feature $X$ is:

$$H(Y|X) = - \sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) \log_2(p(y|x))$$  \hspace{1cm} (19)$$
Performance Measurements

For each class:

- **Precision(P):** the ratio between the number of correctly labelled samples to the total number of samples labelled in the class;

- **Recall(R):** the ratio of correctly labelled samples to the total number of samples whose true labels belong to the class.

- **$F_\beta$-Measure:** the harmonic mean of precision and recall
  
  $$F_\beta = \frac{(1 + \beta^2)PR}{\beta^2P + R},$$

  and when $\beta = 1$, $F_1 = \frac{2PR}{P + R}$. 

## Performance with Attribute Refinement

<table>
<thead>
<tr>
<th>Ticker</th>
<th>Label</th>
<th>Original Set</th>
<th>Economical Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAPL</td>
<td>U (↑)</td>
<td>77.8 74.7 76.2</td>
<td>75.1 74.2 74.6</td>
</tr>
<tr>
<td></td>
<td>D (↓)</td>
<td>80.4 83.3 81.8</td>
<td>80.5 82.3 81.4</td>
</tr>
<tr>
<td></td>
<td>S (–)</td>
<td>99.1 98.8 99.0</td>
<td>99.3 98.6 99.0</td>
</tr>
<tr>
<td>GOOG</td>
<td>U (↑)</td>
<td>85.4 86.8 86.1</td>
<td>85.5 82.6 84.0</td>
</tr>
<tr>
<td></td>
<td>D (↓)</td>
<td>83.0 79.3 81.2</td>
<td>79.5 80.5 80.0</td>
</tr>
<tr>
<td></td>
<td>S (–)</td>
<td>98.6 99.5 99.0</td>
<td>98.6 99.5 99.0</td>
</tr>
</tbody>
</table>

**Table:** cross-validation performance comparisons after attribute refinement, $\Delta t = 5$
## Training and Testing Time Cost

<table>
<thead>
<tr>
<th>Ticker</th>
<th>Training time(s)</th>
<th>Prediction time(ms)</th>
<th>Support vectors</th>
<th>Training Set Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAPL</td>
<td>5.390</td>
<td>0.0311</td>
<td>103</td>
<td>2000</td>
</tr>
<tr>
<td>GOOG</td>
<td>3.640</td>
<td>0.0237</td>
<td>43</td>
<td>800</td>
</tr>
<tr>
<td>Mean</td>
<td>4.515</td>
<td>0.0274</td>
<td>73</td>
<td>1400</td>
</tr>
</tbody>
</table>

**Table: Model time cost: full feature sets, $\Delta t = 5$**
Trading strategy Test: prediction and bid-ask spread

Figure 5: spread crossing prediction of AAPL and bid-ask spread, $\Delta t = 5$
Figure 6: spread crossing prediction of AAPL with profit-loss curve, $\Delta t = 5$
Trading Strategy Test: statistics

<table>
<thead>
<tr>
<th>Duration (s)</th>
<th>Mean (Tick)</th>
<th>Std. Dev. (Tick)</th>
<th>Avg. # of Trades</th>
</tr>
</thead>
<tbody>
<tr>
<td>300</td>
<td>1.5390</td>
<td>1.4352</td>
<td>13.0</td>
</tr>
<tr>
<td>600</td>
<td>1.7781</td>
<td>1.1113</td>
<td>22.0</td>
</tr>
<tr>
<td>900</td>
<td>1.9781</td>
<td>1.5113</td>
<td>31.0</td>
</tr>
<tr>
<td>1800</td>
<td>2.8658</td>
<td>1.8491</td>
<td>68.0</td>
</tr>
<tr>
<td>3600</td>
<td>3.5862</td>
<td>1.1083</td>
<td>126.0</td>
</tr>
</tbody>
</table>

Table: Window size for 30 sliding window tests. Spread crossing, $\Delta t = 5$ (events). Strategy: long $100.00$ on upward crossing signal and short back 5 events later; short $100.00$ on downward signal. After the window duration ends, the accumulated profit/loss reported. 1 tick = $0.01$. 
The End

Thank you!