Model Selection in large-scale problems

Lambert Schomaker

Discussion slides – October 2012
ModelSelection - discussion

General goal

› Solve large-scale classification and regression problems
› Large numbers of
  • Instances
  • Classes
  • Dimensions
› Small numbers of
  • Labeled data
  • Human resources
Applications (9fte)

› Image retrieval, face detection and recognition
› Handwritten word retrieval in historical documents
› Writer/hand identification forensic and paleographic
› Tissue classification (affymetrix)
› BCI/EEG analysis
› Spiking-neuron network models
› Star classification: brown dwarf hunting
› Chemical reaction tracking: Raman spectroscopy
› Robot behavior policy training (POMDPs)
› Camera-based text detection and recognition
Experience / opinion

› Computational intelligence or PhD sweat?
› Many methods from literature contain hidden pitfalls
› Did hundreds of thousands experiments with many features, SVMs, HMMs, Bayes on large data
› ➔ In large data, performances converge!
› Simple 1NN or nearest-centroid works great!
› If necessary: kmeans for class variants
› Reduce human interference
› What do you want to optimize, anyway?
<table>
<thead>
<tr>
<th>Discipline</th>
<th>#dimensions</th>
<th>#instances</th>
<th>#classes</th>
<th>Goal</th>
<th>Challenge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Astronomy</td>
<td>$10^2$-$10^3$ star features</td>
<td>$2 \times 10^6$</td>
<td>$10^2$</td>
<td>Star classification, brown dwarf classification</td>
<td>While the dimensionality is limited, the absolute values of measurements are important and precise</td>
</tr>
<tr>
<td>Genomics</td>
<td>$5 \times 10^4$</td>
<td>$4 \times 10^4$</td>
<td>$10^2$-$10^3$</td>
<td>Disease classification on tissue using RNA expression(Affymetrix) array data</td>
<td>While dimensionality is huge, measurements are noisy. The information is in their correlations: remove 'normal' cell state information, by, e.g., PCA</td>
</tr>
<tr>
<td>Catalysis</td>
<td>$3 \times 10^3$</td>
<td>$10^4$ in spectra $10^2$ time steps</td>
<td>$10$</td>
<td>Ligand effectivity in catalysis using Raman spectroscopy. Determine, in time, which cells in an array are delivering promising results in the ongoing reactions. Classification and regression, time series.</td>
<td>Currently, peak-detection heuristics are used to track chemical reactions. The redundancy in the spectral patterns is ignored, missing important information during expensive chemical processes</td>
</tr>
<tr>
<td>Pattern Recognition</td>
<td>$5 \times 10^3$</td>
<td>$3 \times 10^8$</td>
<td>$5 \times 10^4$</td>
<td>Retrieve and sort instances from dozens of large book collections in a live '24/7' machine learning setup</td>
<td>Bootstrap from one single label per class to reliable, large training sets. We want to translate our successes in continuous learning (Monk) to other disciplines</td>
</tr>
</tbody>
</table>
Separability versus Prototypicality in Handwritten Word Retrieval

Jean-Paul van Oosten

Prof. Lambert Schomaker
Classification versus Retrieval

Classification: \( i_{\text{recog}} = \arg\max_i P(C_i|X) \)

Retrieval: \( H = \text{sort} P(X_j|C) \)

Can we use SVMs for both classification and retrieval?
Word retrieval

Intuition: When using SVMs, if the distance from the decision boundary is large, the instance is probably a good candidate.
I.e., sort on signed distance to the decision boundary
Word retrieval using the discriminant value

Even though the accuracy of classifying “Zwolle” is 97%, the top of the hit list contains counter-intuitive examples.
ModelSelection - discussion

Precision $\approx$ Gain in the feedback-loop

- The user labels word instances by using hit lists.
- The higher the quality of the hit lists, the more labels the user can provide.
- Increments in harvested numbers of labels can be linked to better hit lists due to more instances, etc.
ModelSelection - discussion

Closer analysis of intuition

Regular testing procedures (test set and training set of similar size)

<table>
<thead>
<tr>
<th>Data set</th>
<th>Inst. / class</th>
<th>Accuracy</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test</td>
<td>120+</td>
<td>0.98</td>
<td>0.97</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>7-35</td>
<td>0.96</td>
<td>0.68</td>
<td>0.57</td>
</tr>
</tbody>
</table>
ModelSelection - discussion

Closer analysis of intuition

Test-set now contains realistic number of instances

<table>
<thead>
<tr>
<th>Data set</th>
<th>Inst. / class</th>
<th>Accuracy</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test</td>
<td>120+</td>
<td>0.98</td>
<td>0.97</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>7-35</td>
<td>0.96</td>
<td>0.68</td>
<td>0.57</td>
</tr>
<tr>
<td>+12k distractors</td>
<td>120+</td>
<td>0.99</td>
<td>0.97</td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td>7-35</td>
<td>0.97</td>
<td>0.68</td>
<td><strong>0.01</strong></td>
</tr>
</tbody>
</table>
Fundamental question

› How is it possible that accuracy is not a good predictor of precision in a retrieval context?
› How is it possible that recall and precision, as computed according to standard machine learning practices, are not good predictors of precision in a retrieval context?
ModelSelection - discussion

Approach

› Analyse the precision drop in presumably well-performing classifiers
› Explore a number of methods to counteract the precision drop
› Present a convenient approach suitable for 24/7 machine learning
Distance to the boundary
Distance to the boundary
Separability versus Prototypicality

› Separability: The ability to categorise and separate class instances from non-class instances

› Prototypicality: The similarity of an instance to the canonical class prototype (e.g., measured as the distance to the centroid or prototype of the class).
Goal

› Maximising separation between classes
› Optimal ranking according to prototypicality
ModelSelection - discussion

Method

Two-stage ranking process, using nearest centroid:
1. For each instance, produce the most likely class C
   \[ i_{\text{recog}} = \arg\max_i P(C_i | X) \]
2. Rank each instance classified as C with a secondary feature or method
   \[ H = \text{sort} P(X_j | C) \]

This reduces the number of distractors in the second stage, and can be used to optimise both separability and prototypicality.
Experimental evaluation

› Compare ‘single-step, direct retrieval’ performance to the two-stage ranking process.
› Single-step, direct retrieval means ranking all instances with the distance to each class centroid.
› Two features selected from large set of feature methods
  1. Based on biologically inspired features
  2. 100 x 50 pixel scaled word image
ModelSelection - discussion

Performance measures

1. Top-1 recognition accuracy

2. Precision in top-n = \( \frac{N_{\text{correct}}}{\min(n, |H|)} \)

3. Recall for class C = \( \frac{N_{\text{obtained}}}{N_{\text{targets}}} \)

4. Levenshtein (edit) distance between query and top-word ASCII values
Data set

- Historical document collection from the Dutch Queen’s office. 1404 classes with 7 or more human labelled instances; 84,000 instances.
- Each class is categorised based on the number of labelled instances to investigate the effect in different stages of the labelling process:
  - 7-35
  - 35-60
  - 60-120
  - 120+
ModelSelection - discussion

Classification Accuracy

<table>
<thead>
<tr>
<th>Feature</th>
<th>7-35</th>
<th>35-60</th>
<th>60-120</th>
<th>120+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature 1</td>
<td>0.62</td>
<td>0.93</td>
<td>0.92</td>
<td>0.94</td>
</tr>
<tr>
<td>Feature 2</td>
<td>0.62</td>
<td>0.86</td>
<td>0.87</td>
<td>0.93</td>
</tr>
</tbody>
</table>

7 folds
84288 word instances

<table>
<thead>
<tr>
<th>Fold</th>
<th>Number of instances in train-set</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>72354</td>
</tr>
<tr>
<td>2</td>
<td>72417</td>
</tr>
<tr>
<td>3</td>
<td>72210</td>
</tr>
<tr>
<td>4</td>
<td>72177</td>
</tr>
<tr>
<td>5</td>
<td>72228</td>
</tr>
<tr>
<td>6</td>
<td>72089</td>
</tr>
<tr>
<td>7</td>
<td>72253</td>
</tr>
</tbody>
</table>
ModelSelection - discussion

Precision performance

![Precision performance graph]

- **Average of reranking methods**
- **Direct-retrieval with qp**
- **Direct-retrieval with img**

Number of labeled instances per class
ModelSelection - discussion

Average Levenshtein distance

The graph shows the average Levenshtein distance between labeled instances per class. The y-axis represents the Edit distance, and the x-axis represents the number of labeled instances per class. The graph includes three lines:

- avg. of reranking methods
- direct-retrieval with qp
- direct-retrieval with img

The data points indicate a decrease in Edit distance as the number of labeled instances increases.
Re-rank feature comparison

- Separating classes with feature 1, then re-ranking with feature 2 works better than other configurations.
- This configuration has a large impact in the 7-35 instances per class category, while choice of features has almost no impact in the 120+ category.
Bootstrapping

› During the bootstrapping phase, where there are very few labelled instances per class, the choice of feature can be very important.

› Features will be selected per script and collection to optimise both separability and prototypicality.
Monk

Jumps in number of word labels can be aligned with transition to a different size category
Conclusion

› A large-scale handwritten word retrieval system should optimise for both separability and prototypicality.

› By using the two-stage ranking method, two goals are met:

1. The number of distractors is reduced
2. Instances are ranked in terms of their prototypicality with respect to their class, leading to a more intuitive hit list
3. Take-home message: a method good at recognition does not need to be so in retrieval