Few-shot Learning with Novel Metrics

Patrik Gergely, Ana Sofia Uzsoy, Stuart Burrell

Summary

Motivation
Traditional methods must be re-trained to learn new classes, which is computationally expensive.

The Task
Classify unseen classes with a pre-trained model from a handful of examples.

A Solution
Prototypical Networks, Snell et al. (2017)

Future Directions
Learnable Metrics: model learns the best distance metric for the task
Doc2Vec Pre-training: use pretrained models to improve NLP performance
NLP-Focused architectures: use recurrent or LSTM layers to improve NLP performance

Results

Table 1: Percentage accuracy of different metrics trained on three different datasets. Testing was done 5-way with either 1-shot or 5-shot.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Omniglot</th>
<th>Mini-imagenet</th>
<th>Reuters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Snell's Euclidean</td>
<td>98.8</td>
<td>97.9</td>
<td>49.42</td>
</tr>
<tr>
<td>Squared Euclidean</td>
<td>93.32</td>
<td>96.31</td>
<td>48.35</td>
</tr>
<tr>
<td>KL Divergence</td>
<td>67.46</td>
<td>78.13</td>
<td>38.86</td>
</tr>
<tr>
<td>Generalized 1-div.</td>
<td>74.08</td>
<td>87.45</td>
<td>28.83</td>
</tr>
<tr>
<td>Cosine Similarity</td>
<td>72.69</td>
<td>83.30</td>
<td>38.07</td>
</tr>
<tr>
<td>Cosine with Softmax</td>
<td>82.24</td>
<td>88.45</td>
<td>39.88</td>
</tr>
</tbody>
</table>

Hypothesis: Euclidean distance outperforms cosine similarity as it is a Bregman divergence. 
Finding: KL divergence and Generalized 1-divergence are two Bregman divergences which are outperformed by cosine similarity.

Mathematical Background

Pseudo-metrics

- Squared Euclidean: $\sum_{i=1}^{d} (x_i - y_i)^2$
- Generalised I-divergence: $\sum_{i=1}^{d} x_i \log(x_i) - \sum_{i=1}^{d} x_i (x_i - y_i)$
- KL Divergence: $\sum_{i=1}^{d} x_i \log(x_i/y_i)$
- Cosine Similarity: $\frac{x \cdot y}{||x|| ||y||}$
- Squared Mahalanobis: $\frac{1}{2} (x - y)^T Q^{-1} (x - y)$

Bregman Divergences

The Bregman divergence for generating function $\phi$ is given by

$$d_\phi(x, y) = \phi(x) - \phi(y) - \langle y, \nabla \phi(y) \rangle.$$

Architecture

Default values from Snell et al. (2017) are 5-shot 60-way train, 20-way test with 15 query points for each.

References