LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR INFORMATIK

Similarity Estimation using Bayes Ensembles

Similarity Estimation:

- Applications for similarity or dissimilarity terms:
  - Retrieval Queries (Ranking, Range Queries)
  - Clustering
  - Model Training
- Similarity estimates reflect actual data similarity, i.e.: $s(x_1, x_2) > s(x_1, x_3) \Rightarrow x_1$ more similar to x_2 than to x_3
- Commonly used: *Distance Measures*
- For $x_1, x_2 \in \mathbb{R}^d$: L_p-norms

$$L_p(x_1, x_2) = \left(\sum_{i=1}^{d} |x_{1,i} - x_{2,i}|^p \right)^{1/p}$$

- Pitfalls:
  - Target similarity is ignored
  - Irrelevant features / Correlated features
  - Large influence of single dimensions
  - Arbitrarily large distances

Our Solution:

- Split the problem into similar (SIM) and dissimilar (DIS) object pairs
- Control the influence of a dimension i as probability in [0,1] using a *Bayes Estimate* (BE):

$$\mathrm{BE}_i(x_1, x_2) = \frac{p_{\mathrm{DIS}} \cdot P((x_{1,i} - x_{2,i}) \mid \mathrm{DIS})}{p_{\mathrm{DIS}} \cdot P((x_{1,i} - x_{2,i}) \mid \mathrm{DIS}) + p_{\mathrm{SIM}} \cdot P((x_{1,i} - x_{2,i}) \mid \mathrm{SIM})}$$

- Global distance = *Bayes Ensemble Distance* (BED):

$$\mathrm{BED}(x_1, x_2) = \frac{1}{d} \cdot \sum_{i=1}^{d} \mathrm{BE}_i(x_1, x_2)$$

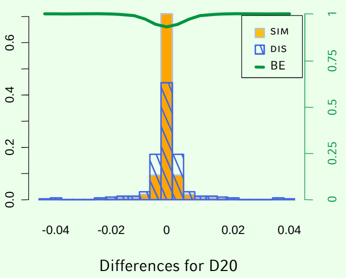
- More stable against outlier dimensions than a classical Naïve Bayes Classifier:

$$\mathrm{NB}(x_1, x_2) = \frac{1}{\mathrm{scale}} \cdot \prod_{i=1}^{d} \mathrm{BE}_i(x_1, x_2)$$

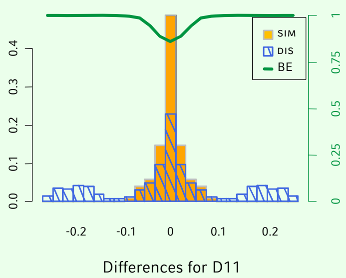
Difference Distributions:

- Similar (SIM) and dissimilar (DIS) object pairs can form differentiable difference distributions
- Example distributions for 32-dimensional color Histogram Data of 34 class image dataset

Suboptimal Dimension



Good Dimension



Feature Quality Assessment:

- Ensemble method: Introduce weights
- Relevance terms for variance difference of SIM and DIS

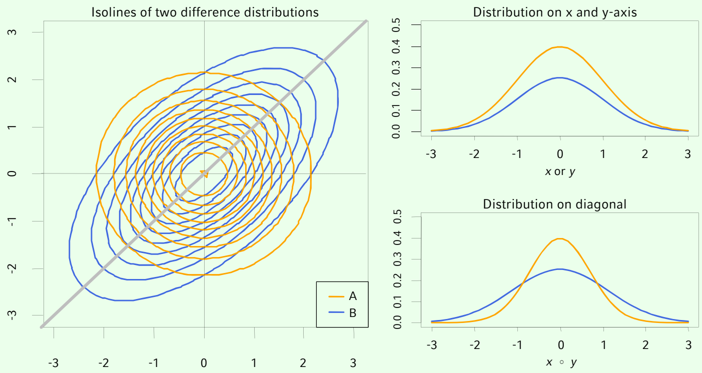
$$q_i = \sigma^2_{\mathrm{DIS}} - \sigma^2_{\mathrm{SIM}} = \operatorname*{avg}_{x_d \in \mathrm{DIS}}(x^2_{d,i}) - \operatorname*{avg}_{x_s \in \mathrm{SIM}}(x^2_{s,i})$$

- Include into ensemble:

$$\mathrm{BED}(x_1, x_2) = \left(\sum_{i=1}^{d} q_i \right)^{-1} \cdot \sum_{i=1}^{d} q_i \cdot \mathrm{BE}_i(x_1, x_2)$$

Feature Space Improvement:

- Dimensionality reduction
- Exploitation of correlated features



- Distance covariances of SIM and DIS

$$\Sigma_{\mathrm{SIM}} = \sum_{x_s \in \mathrm{SIM}} x_s^T \cdot x_s, \quad \Sigma_{\mathrm{DIS}} = \sum_{x_d \in \mathrm{DIS}} x_d^T \cdot x_d$$

- Target transformation $W = (w_1, \ldots, w_{d^*})$ to dimension d^*
- Maximize the variance difference:

$$\begin{aligned} \max \quad & w_i^T \cdot (\Sigma_{\mathrm{DIS}} - \Sigma_{\mathrm{SIM}}) \cdot w_i \\ \text{s.t.} \quad & w_i \perp w_j \ \forall i, j \in \{1, \ldots, d^*\} \end{aligned}$$

- Equivalent to solving EVD: $\lambda w = (\Sigma_{\mathrm{DIS}} - \Sigma_{\mathrm{SIM}}) \cdot w$

Algorithm:

Input: X with $x_i \in \mathbb{R}^d$, SIM, DIS, target dimension d^*

(1) Derive Σ_{SIM} and Σ_{DIS}

(2) Compute feature transformation $W \in \mathbb{R}^{d, d^*}$

(3) Get weights q_i ($i \in 1 \ldots d^*$) for new feature space $W^T X$ using the features' variance differences

Output: Bayes Ensemble Distance

$$\mathrm{BED}(x_1, x_2) = \left(\sum_{i=1}^{d^*} q_i \right)^{-1} \cdot \sum_{i=1}^{d^*} q_i \cdot \mathrm{BE}_i(W^T x_1, W^T x_2)$$

Conclusions on BEDs:

- Balanced, adaptive distance measure
- Easily interpretable
- Applicable to various datasets (discrete class labels, pair-wise similarity labels, regression target functions)

Database Group

Tobias Emrich, Franz Graf, Hans-Peter Kriegel, Matthias Schubert, Marisa Thoma

SSDBM