Multi-Repräsentierte Ähnlichkeitsschätzer

Bisher:

- Distanzmetrik/Skalarprodukt = (Un-)ähnlichkeitsmaß
- Ähnlichkeit ist linear und kann beliebig groß werden (Kernel) oder Unähnlichkeit ist linear und kann unendlich ansteigen.
- Kombinationen von Gewichtungsfunktionen brauchen Trainingsbeispiele

Aber:

- Ähnlichkeit und Unähnlichkeit sind beschränkt:
  - ab einer gewissen Ähnlichkeit werden Objekte als gleich wahrgenommen
  - man unterscheidet ab einer gewissen Unähnlichkeit nicht weiter
- Trainingsbeispiele, die die Ähnlichkeit 2er Objekte beschreiben sind schwer zu erzeugen. (Selbst Menschen labeln häufig inkonsitent!)
- Ein und derselbe Distanzwert kann bei unterschiedlicher Objektähnlichkeit beobachtet werden.

Multi-Repräsentierte Ähnlichkeitsschätzer

Lösungsansatz:

Beschreibe Ähnlichkeit, als Wahrscheinlichkeit, dass ein Benutzer beide Objekte als ähnlich betrachtet.

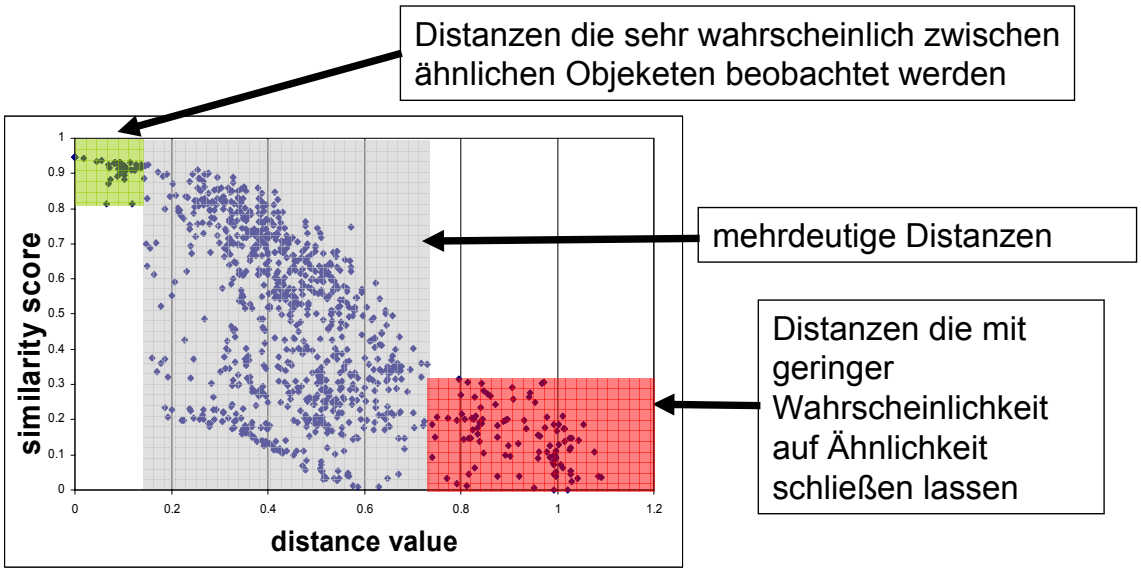
$\Rightarrow$ Abstand in einer Repräsentation wird zu einem Feature das statistisch mit der Ähnlichkeitsaussage korreliert ist.

$\Rightarrow$ Ähnlichkeit wird also als bedingte Wahrscheinlichkeit angesehen.

$\Rightarrow$ Die Unähnlichkeitswahrscheinlichkeit kann dann als Ranking-Kriterium für Ähnlichkeitsanfragen, kNN-Klassifikatoren und distanzbasierte Algorithmen verwendet werden.

Multi-Repräsentierte Ähnlichkeitsschätzer

Beobachtung: Betrachte Distanzen und Ähnlichkeitsaussagen.



$\Rightarrow$ Distanzen sind nur für große und kleine Werte mit Ähnlichkeit korreliert.

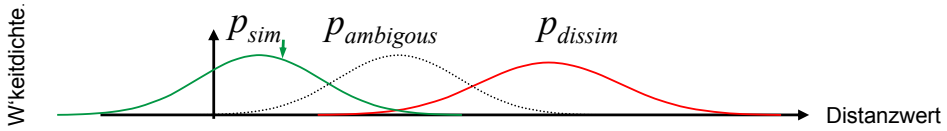
$\Rightarrow$ mittlere Distanzen können alles bedeuten.

Multi-Repräsentierte Ähnlichkeitsschätzer

Idee:

- Modelliere die Unsicherheit der Aussage als Wahrscheinlichkeitsdichte.
- zur Bestimmung der Ähnlichkeit werden zunächst 2 Verteilungen über die Distanzen von ähnlichen und unähnlichen Objekten betrachtet.
- die Verteilung der Unsicherheit kann dann als kombinierte Wahrscheinlichkeit betrachtet werden, dass beide Verteilungen den gleichen Distanzwert liefern.
- Die Unsicherheit in einem Interval von Distanzen wird dann wie folgt berechnet:

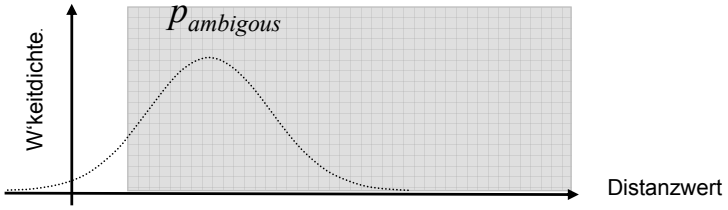
$$P_{ambiguous}(a,b) = \frac{\int_{a}^{b} p_s(x)p_d(x)dx}{\int_{-\infty}^{\infty} p_s(x)p_d(x)dx}$$



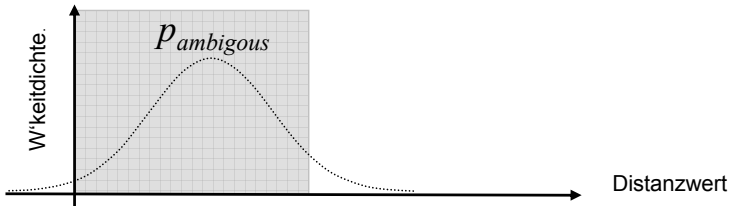
Multi-Repräsentierte Ähnlichkeitsschätzer

Ein Distanz hat eine sichere Aussage, wenn sie nicht zweideutig ist:

- eine sichere Aussage bei der die Distanz kleiner ist als die Mehrzahl der unsicheren Aussagen deutet auf Ähnlichkeit hin.



- eine sichere Aussage bei der die Distanz größer ist als die Mehrzahl der unsicheren Aussagen deutet auf Unähnlichkeit hin.



Multi-Repräsentierte Ähnlichkeitsschätzer

Die Wahrscheinlichkeit, dass 2 Objekte bzgl. einer Repräsentation ähnlich sind kann man also folgendermaßen bestimmen:

$$L^i_{sim}(o_1,o_2) = P^i_{definite}\left(d_i(o_1,o_2) < \delta\right) = P^i_{ambiguous}\left(d_i(o_1,o_2) \geq \delta\right)$$

Kombiniert man diese Wahrscheinlichkeiten für alle Repräsentationen ergibt sich folgendens Ähnlichkeitsmaß:

$$P_{SIM}(o_1,o_2) = \prod_{i=1}^{R} L^i_{sim}(o_1,o_2) \cong \sum_{i=1}^{R} \ln\left(L^i_{sim}(o_1,o_2)\right)$$

Wird eine Distanz benötigt wird das Komplement verwenden:

$$P_{DISSIM}(o_1,o_2) = \prod_{i=1}^{R} L^i_{dissim}(o_1,o_2) \cong \sum_{i=1}^{R} \ln\left(L^i_{dissim}(o_1,o_2)\right)$$

Multi-Repräsentierte Ähnlichkeitsschätzer

Wie bestimmt man die Dichtefunktionen für $p_{sim}(x)$ und $p_{dissim}(x)$?

Training mit Beispielen:

- benutze ein ausreichendes Sample von Objektpaaren
- Label können als graduelle Ähnlichkeit oder binär vorliegen (Bei Einteilung in Klassen sind alle Objekte einer Klasse ähnlich.)
- bestimme eine Normalverteilung bei der Ähnlichkeits- bzw. Unähnlichkeitswert eines Objektvergleichs das Gewicht in der Berechnung darstellt.

$$\mu_r^{sim} = \frac{\sum_{o_1,o_2 \in S} sim(o_1,o_2) \cdot d_r(o_1,o_2)}{\sum_{o_1,o_2 \in S} sim(o_1,o_2)}$$

$$Var_r^{sim} = \frac{\sum_{o_1,o_2 \in S} sim(o_1,o_2) \cdot \left(d_r(o_1,o_2) - \mu_r^{sim}\right)^2}{\sum_{o_1,o_2 \in S} sim(o_1,o_2)}$$

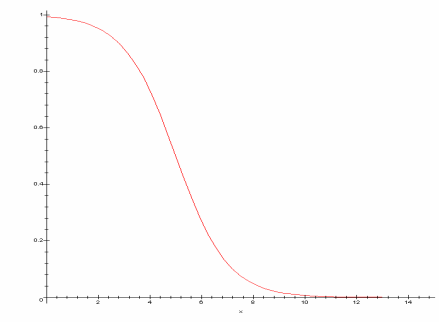
$$\mu_r^{dissim} = \frac{\sum_{o_1,o_2 \in S} dissim(o_1,o_2) \cdot d_r(o_1,o_2)}{\sum_{o_1,o_2 \in S} dissim(o_1,o_2)}$$

$$Var_r^{dissim} = \frac{\sum_{o_1,o_2 \in S} dissim(o_1,o_2) \cdot \left(d_r(o_1,o_2) - \mu_r^{dissim}\right)^2}{\sum_{o_1,o_2 \in S} dissim(o_1,o_2)}$$

Multi-Repräsentierte Ähnlichkeitsschätzer

- Nach der Bestimmung der Verteilung muss noch das Integral über den Dichtefunktionen berechnet werden.
- Da die Stammfunktion der Normalverteilung unbekannt ist, muß die kumulierte Dichtefunktion approximiert werden.
- Dies kann mit Hilfe einer Sigmoidfunktion erreicht werden:
  - a: regelt Verschiebung
  - b: regelt Steigung

$$sigmoid_{a,b}(x) = \frac{1}{1 + e^{a+b \cdot x}}$$



- Anpassen des Sigmoiden über numerische Methoden (Regression)

Multi-Repräsentierte Ähnlichkeitsschätzer

Training ohne Beispiellabel:

Idee: Objekte die in allen Repräsentationen sehr kleine bzw. sehr große Distanzen haben, sind wahrscheinlich auch semantisch ähnlich bzw. unähnlich.

Vorgehen: Co-Learning

- Bilde ein Qualitätskriterium, dass die Konsistenz der Wahrscheinlichkeitsaussagen über alle Repräsentationen hinweg mißt
- Verwende EM-Ansatz der dieses Qualitätsmaß minimiert

Ziel: Die Ähnlichkeitsaussagen für dasselbe Objektpaar sollen in jeder Repräsentation möglichst gleich sein.

Multi-Repräsentierte Ähnlichkeitsschätzer

EM-Algorithmus zum iterativen Anpassen der Ähnlichkeit:

E-Step: Qualitätsmaß für die Übereinstimmung:

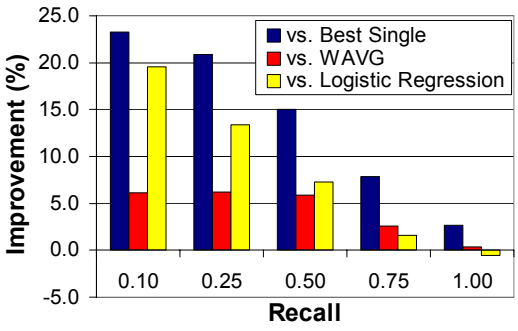
$$L_s^{AVG}(o_1,o_2) = \frac{\sum_{i=1}^{R} L_S^i(o_1,o_2)}{R}$$

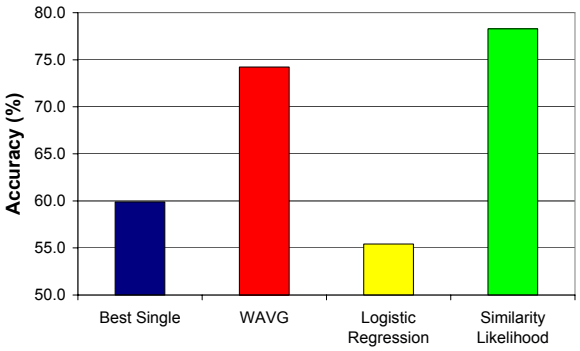
$$Var(S) = \sum_{o_1,o_2 \in S} \sum_{i=1}^{R} \left[L_S^i(o_1,o_2) - L_s^{AVG}(o_1,o_2) \right]^2$$

M-Step: Benutze denselben Ablauf wie bei der Methode mit Trainingsdaten. Anstatt der beobachteten Ähnlichkeitswerte wird allerdings $L_s^{AVG}(o_1,o_2)$ verwendet.

Multi-Repräsentierte Ähnlichkeitsschätzer

Ergebnisse: 500 Audiotitel in 6 Repräsentationen.





- Verbesserung der Precision für verschiedene Recall-Werte
- 1 NN Klassifikation
- WAVG entspricht dem normalisiertem Durchschnitt
- LogRegression ist die Anwendung logistischer Regression zur bestimmung von Gewichten (vgl. Folie 273)