Ludwig-Maximilians-Universität München
Institut für Informatik
Dr. Arthur Zimek
Erich Schubert

Knowledge Discovery in Databases

SS 2013

Übungsblatt 10: Kernel und Regression

Aufgabe 10-1 Kernel-Funktionen

Wie in der Vorlesung erklärt, zeichnet sich eine Kernel-Funktion ("Kernel") durch positive (Semi-)Definitheit aus. Eine Matrix A ist positiv definit, falls ihre Eigenwerte nichtnegativ sind, oder alternativ formuliert, falls für all $x \in \mathbb{R}^d$ gilt: $x^\top \cdot A \cdot x \geq 0$

Zeigen Sie, dass folgende Funktionen Kernels sind, falls x und $\hat{x}$ Vektoren im $\mathbb{R}^d$ sind:

(a) $k_1(x, \hat{x}) = 1$

(b) $k_2(x, \hat{x}) = 3 \cdot x^\top \cdot \hat{x}$

(c) $k_3(x, \hat{x}) = 3 \cdot x^\top \cdot \hat{x} + 5$

Aufgabe 10-2 Lineare Regression

Das Gehalt einer Person hängt von der Anzahl der Jahre ab, in denen die Person ihren Beruf ausgeübt hat. Um diesen Zusammenhang genauer zu untersuchen, kann man ein lineares Regressionsmodel lernen. Als Trainingsmenge stehen uns die Jahre an Berufserfahrung und die Gehälter folgender Personen zur Verfügung.

Erfahrung in Jahren	Gehalt in (1000$)
3	30
8	57
9	64
13	72
3	36
6	43
11	59
21	90
1	20
16	83

(a) Berechnen Sie eine Regressionsgerade, die dazu dienen soll, das voraussichtliche Gehalt auf Basis der Berufserfahrung abzuschätzen. Bestimmen Sie hierzu die Gerade, die den quadratischen Fehler minimiert.

(b) Bestimmen Sie den quadratischen Fehler der berechneten Gerade, um abzuschätzen, wie gut die Regressiongerade den Zusammenhang erklärt.

(c) Berechnen Sie mit Hilfe Ihrer Regressionsgerade das vorraussichtliche Gehalt für Personen mit den folgenden Jahren an Berufserfahrung:
Person A: 20
Person B: 8
Person C: 11

Aufgabe 10-3 Evaluierung von Clusterings

Gegeben seien zwei Clusterings $\{A_1, \ldots, A_5\}$ und $\{B_1, \ldots, B_4\}$,
und folgende Matrix mit den Größen der Schnittmengen $|A_i \cap B_j|$:

	A_1	A_2	A_3	A_4	A_5
B_1	2	1	11	0	13
B_2	0	10	0	3	1
B_3	15	0	4	0	0
B_4	0	0	2	8	1

- Bewerten Sie das Clustering, indem Sie einmal nur die Zeilenmaxima (als "Precision") und einmal nur die Spaltenmaxima (als "Recall") betrachten. Kombinieren Sie diese Werte mit dem F-Measure.
- Berechnen Sie in jeder Zeile eine "Precision" und jeder Spalte einen "Recall", indem sie den größten Wert als "true positive" annehmen. Berechnen Sie das F-Measure aus der mittleren Precision der Zeilen und dem mittleren Recall der Spalten.
- Berechnen Sie für jede Zelle ein F-Measure (indem sie die "Precision" bzgl. der Zeilensumme und den "Recall" bezüglich der Spaltensumme berechnen). Suchen Sie für jede Zeile und Spalte das Maximum, und berechnen Sie daraus je einen mittleren F-Score für alle Zeilen und alle Spalten, sowie den Mittelwert aus diesen beiden.
- Pair Counting: um nicht immer genau einen Wert aus einer Zeile oder Spalte in Betracht zu ziehen, werden alle Objektpaare betrachtet die in beiden Clusterings zusammen in einem Cluster sind. Ein Objekt bilde dabei kein Paar mit sich selbst (also keine Paare der Form (x, x)), d.h. Paare existieren in einem Clustering A:

  $$(x, y) \in P(A) \Leftrightarrow \exists_{A_i \in A} x \in A_i \wedge y \in A_i \wedge x \neq y$$

  Berechnen Sie Precision und Recall der Paare und das F-Measure daraus. Berechnen Sie auch den Rand Index, Adjusted Rand Index und Jaccard Index mit den Formeln aus der Vorlesung.