Data Mining und Maschinelles Lernen

TECHNISCHE UNIVERSITÄT DARMSTADT

7. Übungsblatt

Aufgabe 1 Nearest Neighbour

Gegeben sei folgende Beispielmenge:

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	26	High		No
D2	Sunny	28	High	Strong	No
D3	Overcast	29	High	Weak	Yes
D4	Rain	23	High	Weak	Yes
D5	Rain		Normal	Weak	Yes
D6	Rain	12	Normal	Strong	No
D7	Overcast	8		Strong	Yes
D8	Sunny	25	High	Weak	No
D9	Sunny	18	Normal	Weak	Yes
D10	Rain	20	Normal	Weak	Yes
D11	Sunny	20	Normal	Strong	
D12	Overcast	21	High	Strong	Yes
D13		26	Normal	Weak	Yes
D14	Rain	24	High	Strong	No
D15	Sunny	23	Normal	Weak	No
D16	Sunny	21	Normal	Weak	Yes

a) Um fehlende Werte zu behandeln, kann man diese einfach auffüllen, indem man die am naheliegendsten Nachbarn zu diesem Beispiel verwendet. Benutzen Sie hierfür 3-NN zum Ausfüllen dieser Werte. Normieren Sie das numerische Attribut, wie im Skript beschrieben (Foliensatz Instance-based Learning, Folie "*Distance Functions for Numeric Attributes*"), nehmen Sie für nominale Attribute die 0/1-Distanz (Foliensatz Instance-based Learning, Folie "*Distance Functions for Symbolic Attributes*") und benutzen Sie als Distanzfunktion für 3-NN die Manhattan-Distanz.

   Beziehen Sie für das Auffüllen von Werten die Klassifikation der Beispiele mit ein oder nicht? Warum? Überlegen Sie sich auch, wie sie beim Auffüllen mit fehlenden Attributen in den Nachbarn umgehen. Verwenden Sie die so ausgefüllten Werte auch für die nächsten Aufgaben.

   Welche Distanzfunktion ergibt sich für das numerische Attribut?

b) Benutzen Sie für die Berechnung von k-NN die gleichen Eckdaten wie in der vorherigen Aufgabe (Normierung für numerische Attribute, 0/1-Distanz für nominale Attribute und die Manhatten Distanz). Klassifizieren Sie so mit 1-NN das folgende Beispiel.

   - D17: Outlook=Sunny, Temperature=23, Humidity=High, Wind=Strong

c) Testen Sie nun verschiedene k. Für welches k ändert sich die Klassifikation gegenüber $k = 1$?

d) Berechnen Sie den Klassifikationswert obiger Instanz mittels abstandsgewichtetem NN (Inverse Distance Weighting). Überlegen Sie sich hierzu, wie Sie diese Methode auf nominale Attribute anwenden können.

e) Gehen Sie nun vom originalen, unveränderten Datensatz von Aufgabe 1a) aus und benutzen Sie für die Berechnung von k-NN wieder für numerische Attribute die normierte Distanzfunktion und für nominale Attribute diesmal die *Value Difference Metric (VDM)* (Foliensatz Instance-based Learning, Folie "*Distance Functions for Symbolic Attributes*"). Nehmen Sie für die Berechnung der *VDM* einen Wert von $k = 1$ an und normieren Sie die Distanzen mit der

Anzahl der Klassen. Überlegen Sie sich dabei auch, was mit fehlenden Attributwerten geschieht. Klassifizieren Sie so das Beispiel aus Aufgabe b), verwenden Sie dabei 1-NN und die euklidische Distanz (Foliensatz *Instance-based Learning*, Folie "*Distance Functions*"). Ändert sich die Klassifikation im Vergleich zur Aufgabe b)?