Maschinelles Lernen
Übungsblatt 7

Dr. Steffen Rendle, Christoph Freudenthaler
Wirtschaftsinformatik und Maschinelles Lernen (ISMLL)
Universität Hildesheim

15. Dezember 2010
Abgabe bis 7. Januar 2011

KNN-Methode

Folgender Datensatz beschreibt die durchschnittliche Zufriedenheit von jeweils ca. 35 Büromitarbeitern aus 10 verschiedenen Abteilungen eines Finanzdienstleisters bezüglich den Arbeitsbedingungen. Es wurden 5 verschiedene Aspekte von Zufriedenheit abgefragt. Die Werte entsprechen dem Anteil der Mitarbeiter, die zur jeweiligen Frage mit *ja* geantwortet haben:

Gesamtzufriedenheit	**Beschwerdeumgang**	**Lernmöglichkeiten**	**Leistungsbelohnung**	**Kritikfähigkeit**
43	51	39	61	92
63	64	52	63	73
71	70	69	76	86
61	63	47	54	84
81	78	66	71	83
43	55	44	54	49
58	65	58	66	68
71	75	55	70	66
72	82	67	71	83
67	61	47	62	80

a) Betrachten sie nur die Spalten 1–3. Erzeugen Sie ein Streudiagramm mit den Datenpunkten der Spalten 2 und 3. Berechnen Sie mittels 1-NN und der Euklidischen Distanz die Vorhersage der *Gesamtzufriedenheit* auf Basis der erklärenden Variablen *Beschwerdeumgang* und *Lernmöglichkeiten* für eine weitere Abteilung, deren Anteile zur Zufriedenheit mit dem Beschwerdeumgang und zur Zufriedenheit mit den Lernmöglichkeiten bekannt sind. Der Anteil zufriedener Mitarbeiter mit dem Umgang mit Beschwerden der Mitarbeiter ist $53\,\%$ und der Anteil der von Mitarbeitern, die mit den Lernbedingungen zufrieden sind, liegt bei $58\,\%$. Tragen Sie den neuen Punkt in ihr Streudiagramm ein und markieren Sie den nächsten Nachbarn.

b) Stellen Sie nun zur Beschleunigung des 1-NN Verfahrens einen Suchbaum (basierend auf der Euklidischen Distanz) auf, der den 2-dimensionalen Raum der erklärenden Variablen *Beschwerdeumgang* und *Lernmöglichkeiten* in 4 disjunkte Abschnitte trennt. Zeichnen Sie die Grenzen der 4 Abschnitte ein und erweitern Sie das Streudiagramm um die Mittelwerte der 4 Abschnitte. Sagen sie anhand dieses Suchbaumes die Gesamtzufriedenheit für die 11. Abteilung voraus und stellen Sie den Vorhersageprozess grafisch dar. Was fällt ihnen auf? Liefert der Suchbaum (immer) das gleiche Ergebnis wie das langsamere 1-NN Verfahren ohne Suchbaum?

Hinweis: Zerlegen Sie den 2-dimensionalen Raum so, dass Sie jede Achse mit einer Trennlinie halbieren und in jeder Achsenhälfte $50\,\%$ der Datenpunkte (= 5 Datenpunkte) liegen.

c) Man kann KNN-Verfahren als distanzgewichtete Mittelwerte definieren. Geben Sie die Distanzfunktion (=Kernel) für das 1NN-Verfahren an. Welche Alternativen gibt es zum Kernel des KNN-Verfahrens?

Entscheidungsbäume II

a) Erstellen sie für die Daten aus Aufgabe 1 einen Regressionsbaum mit den erklärenden Variablen *Beschwerdeumgang* und *Lernmöglichkeiten*. Die Zielvariable ist erneut die Gesamtzufriedenheit. Verwenden Sie als Qualitätskriterium den quadratischen Fehler und erlauben Sie nur Blattknoten mit einer Mindestgröße von 2 Elementen. Erkennen Sie einen Unterschied zwischen dem gelernten Entscheidungsbaum und dem Suchbaum aus Aufgabe 1? Wenn ja, welchen?

b) Die folgenden Trainingsdaten seien gegeben:

Day	Outlook	Temp.	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Die Zielvariable *PlayTennis* mit den Werten *yes* und *no* muss für verschiedene Samstage in Abhängigkeit der Attribute für den jeweiligen Vormittag vorhergesagt werden.
Erstellen Sie einen binären Entscheidungsbaum anhand des in der Vorlesung vorgestellten Verfahrens („greedy strategy“).
Geben Sie für jeden Knoten die betrachteten möglichen Splits an. Verwenden Sie den entropy gain als Qualitätskriterium für den Split.

c) Bonus für Besucher der Vo Künstliche Intelligenz Vergleichen Sie die Suchstrategie der Entscheidungsbäume (suche aus allen erklärenden Variablen jene mit dem besten Split gem. Qualitätskriterium) mit der greedy-best-first Suche, die Sie aus der Vorlesung Künstliche Intelligenz kennen. Worin liegt der einzige Unterschied? Vergleichen Sie auch die 1NN-Strategie aus dem letzten Übungsblatt mit der greedy-best-first Suche. Erkennen Sie einen Unterschied? Wenn ja: welchen? Wenn nein: warum nicht?

Entscheidungsbäume

Geben Sie für jede der folgenden Booleschen Funktionen einen Entscheidungsbaum an:

1. $A \wedge B$
2. $A \vee B$
3. $A \oplus (B \vee C)$
4. $(A \vee B) \wedge (C \vee D)$

Dabei soll in jedem Entscheidungsknoten nur eine Variable abgefragt werden.
Hinweis: Das Symbol $\oplus$ steht für die XOR-Funktion.