Übungsblatt 1, Business Analytics, SoSe 2011, 02.05.2011
Dr. Tomáš Horváth, Osman Akcatepe

1. Figur 1 zeigt die Klassifizierungsdaten mit zwei Klassen: Kreis-Quadrat. Die zwei Instanzen mit gepunkteten Linien, die 1 und 2 beschriftet worden sind, sind noch nicht klassifiziert worden. Welche Klassenetikette würden zu ihnen durch k-nächste-Nachbarn für k zugeteilt werden k= 1, k = 3 und k = 5?

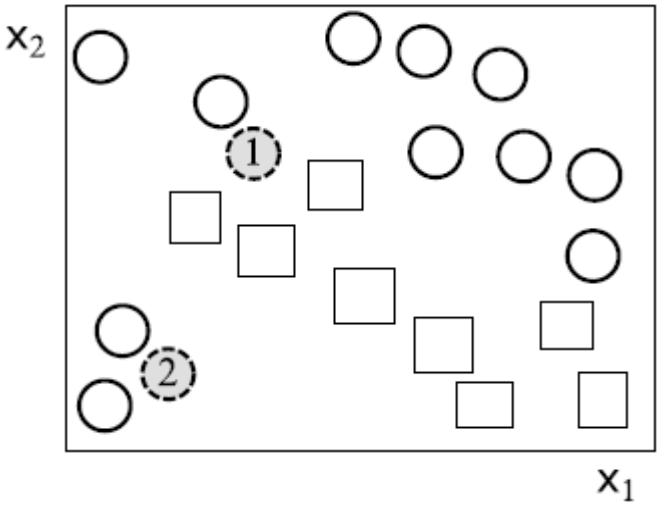

Figure 1: Ein Klassifikationsdataset

2. In einem verschiedenen Dataset mit gegebener Tabelle 1 kann das Feature x_1 drei mögliche Werte annehmen: Quadrat, Kreis und Sechseck. Das Feature x_2 kann irgendeinen Ganzzahlwert annehmen, und Sie können annehmen, dass es logisch ist, den Unterschied zwischen zwei von seinen Werten anzuschauen. Was wäre ein passender Weg, diese Features für die k-nächste-Nachbarn zu vertreten, (angenommen, dass es euklidische Distanz zwischen Instanzen gibt) ?

X1 ToyShape	X2 NumberOfSells	Y OfGoodQualityOrBadQuality
Quadrat	1	Gut
Quadrat	24	Schlecht
Kreis	0	Gut
Sechseck	0	Gut

Table 1: Ein künstliches Dataset

Übungsblatt 1, Business Analytics, SoSe 2011, 02.05.2011
Dr. Tomáš Horváth, Osman Akcatepe

3. Vielleicht wäre es besser, NumberOfSells durch das Dutzend zu messen. Nehmen Sie so an, dass wir x_2 von Tabelle 1 in Einheiten von zwölf messen würden. Zum Beispiel 1 => 1/12 und 24 => 2. Wie würde das k-nächste-Nachbarn beeinflusst?

4. Für jeden des Folgenden, das Probleme lernt, bitte anzuzeigen, ob es eine Voraussage (Prediction), Regression oder Klassifizierungsproblem ist.

(a) Eine Suchmaschine versucht, ob eine Website zu bestimmen, von Sport basiert auf der Anzahl von Zeiten, die Website enthält die Folgenden Wörter/Phrasen: „Sport", „Fußball", „Tennis", „Hockey", „Wahlen", "Menschenrecht' und „Party".
(b) Ein Landwirt hat verschiedene Beträge des Düngemittels auf verschiedenen Teilen von seinem Land benutzt. Er hat die durchschnittliche Höhe seiner Maises für jeden Teil aufgezeichnet. Jetzt will er lernen, wie die durchschnittliche Höhe seiner Maises vom Betrag des Düngemittels abhängt, den er benutzt.
(c) Jeden Frühling zählt ein Biologe die Anzahl der Nachkommenschaft in der gleichen Löwenbevölkerung. Basiert auf ihren Zahlen der letzten Jahre will er die Anzahl der Nachkommenschaft im nächsten Jahr schätzen.

5. Nehmen Sie an, dass eine Probe des vier Zahl [1,2,13,15] von einer Population gezogen ist. Um vorauszusagen, eine fünfte Zahl, die noch nicht von jener gleichen Population gezogen wird, kann einer „zentrale Tendenz" machen: Welcher Schätzer wird voraussagen am besten: der Median oder der Mittelwert oder das Midrange?

Viel Erfolg!

* Das Midrange ist definiert durch M = (min X + max X) / 2