2. Übung

Schriftliche Aufgaben

Vorname: Tom
Nachname: Tucek
Matrikelnummer: 1325775

1.3 Parameter- und Merkmalsauswahl

1.3.1: Ideale Konfiguration

Die ideale Konfiguration (= geringste Fehlerrate) scheint die Kombination der Merkmale 2, 3 und 4 zu sein, mit einem k von 3. Die Fehlerrate dabei beträgt 2% (= 1 falsch klassifizierter Datensatz.)

1.3.2: Begründung

Die Bilder der nächsten Seite zeigen die Verteilung der Merkmale bei jeweils 2 aktiven Merkmalen (in den Überschriften sichtbar). Die 2 Klassen sind unterschiedlich eingefärbt. Dadurch ist schnell sichtbar, dass die Kombination der Merkmale 3 und 4 die deutlichste Trennung erzeugt. (In 3D Scatter-Plots ist leider nicht viel mehr erkennbar.)

Außerdem ist auf dem Bild, welches die Kombination von 2 und 4 zeigt, sichtbar, dass die 2 Klassen beinahe durch eine gerade Linie getrennt werden können, allerdings gibt es ein paar Überschneidungen in der Mitte.

Die Kombination der Merkmale 2 und 4 hat bestenfalls eine Fehlerrate von 6%. Die Kombination der Merkmale 3 und 4 hat jedoch eine beste Fehlerrate von 4% bei z.B. k=3. Dies entspricht der besten Fehlerrate der Kombination aller 4 Merkmale. Allerdings kann eine bessere Fehlerrate erreicht werden, indem die 2 Kombinationen zusammengefügt werden, zur Kombination der Merkmale 2, 3 und 4.

Die Bilder auf der übernächsten Seite zeigen die Fehlerraten verschiedener Merkmalskombinationen. Hierbei ist zu erwähnen, dass alleine das Merkmal 4 bereits eine Fehlerrate von 6% bei k=3 oder k>=13 erreicht (Figure 1).

Die Kombination von Merkmal 3 und 4 erreicht als einzige 2er-Kombination eine Fehlerrate von 4% (Figure 2). Die Kombination der Merkmale 2, 3 und 4 erreicht schließlich die bestmögliche Fehlerrate von 2% bei k=3 (Figure 3), obwohl die Kombination aller 4 Merkmale bestenfalls eine Fehlerrate von 6% hat.

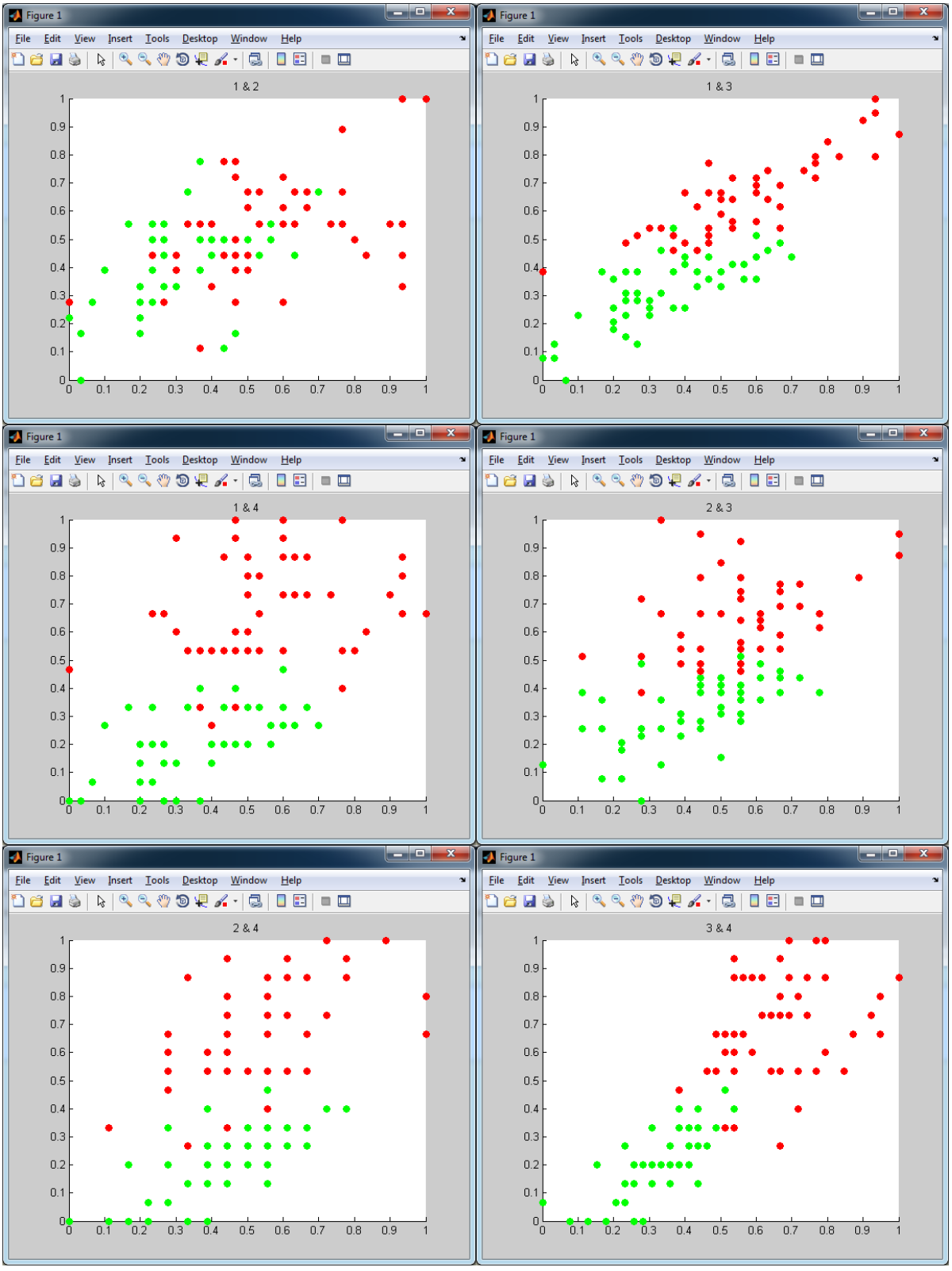


Figure 1
File Edit View Insert Tools Desktop Window Help
1 & 2
Figure 1
File Edit View Insert Tools Desktop Window Help
1 & 3
Figure 1
File Edit View Insert Tools Desktop Window Help
1 & 4
Figure 1
File Edit View Insert Tools Desktop Window Help
2 & 3
Figure 1
File Edit View Insert Tools Desktop Window Help
2 & 4
Figure 1
File Edit View Insert Tools Desktop Window Help
3 & 4

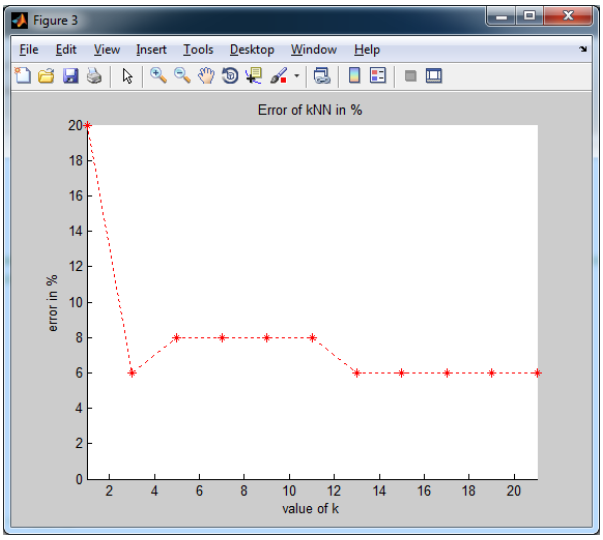


Figure 1: Fehlerrate von Merkmal 4 alleine

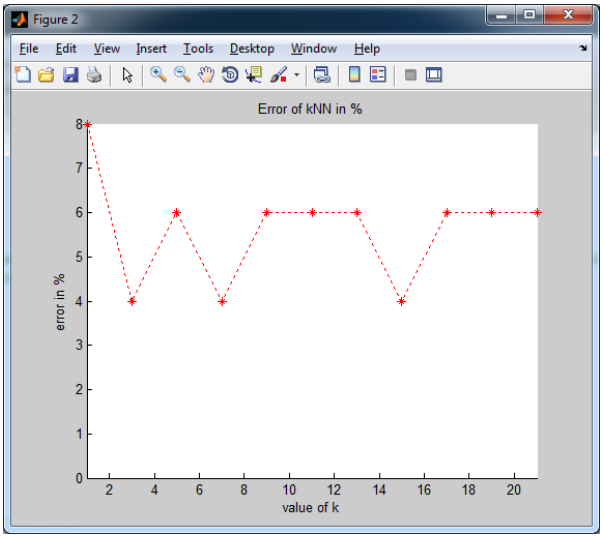


Figure 2: Fehlerrate von Kombination 3&4

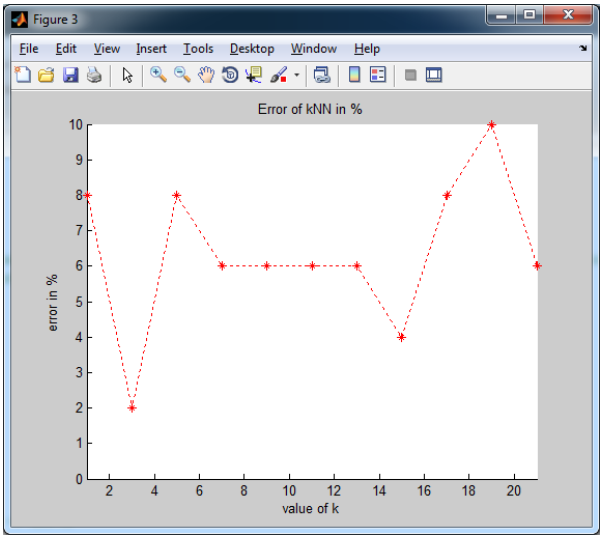


Figure 3: Fehlerrate von Kombination 2&3&4

1.3.3: Parameter k

Bei einem geraden Wert für k könnte es passieren, dass die k nächsten Nachbarn zur Hälfte zu einer Klasse und zur anderen Hälfte zur anderen Klasse gehören, was eine eindeutige Klassifizierung nicht möglich machen würde.

Beispielsweise könnten bei k=8 vier Nachbarn der Klasse 1, sowie vier Nachbarn der Klasse 2 gefunden werden. Folgend kann nicht alleine durch die jeweilige Anzahl der Nachbarn entschieden werden, wie klassifiziert wird und es müsste entweder mit einer dritten Möglichkeit (nicht klassifiziert) oder mit einem komplexeren Verfahren gearbeitet werden.

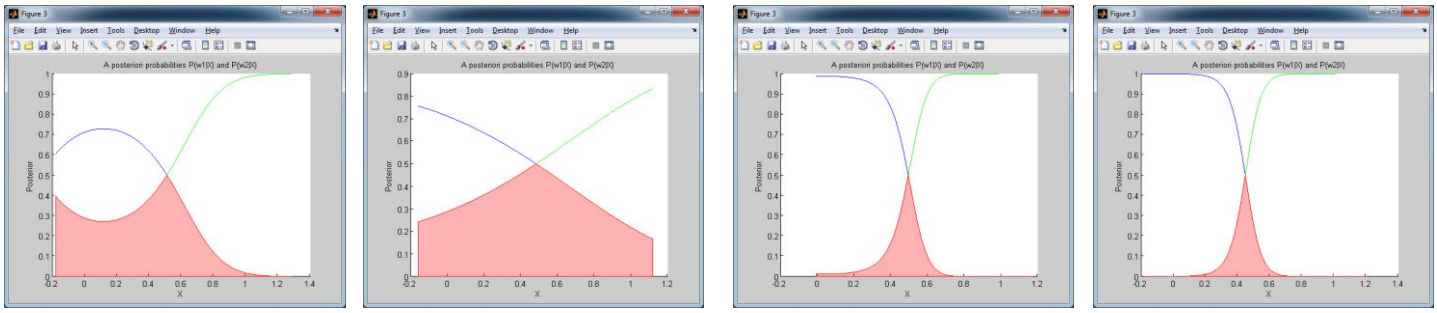
2.6 Merkmalsauswahl und Evaluierung

2.6.1 Merkmalsauswahl

Die Fehlerraten der Merkmale lauten wie folgt:

Merkmal 1	34%
Merkmal 2	38%
Merkmal 3	10%
Merkmal 4	08%

Demzufolge hat Merkmal 4 die geringste Fehlerrate bei der Verwendung des Bayes-Theorems. Anhand der folgenden 4 Abbildungen (Merkmal 1 bis 4 von links nach rechts) kann gesehen werden, dass nur Merkmal 3 und Merkmal 4 geeignete Posteriors haben, welche den Bereich der Fehlerwahrscheinlichkeit gering halten.



Die Posteriors von Merkmal 1 und Merkmal 2 liegen zu nahe beieinander, was zu mehr falschen Klassifikationen führt. Die Fläche der Fehlerrate von Merkmal 3 ist nur minimal größer als die von Merkmal 4.

2.6.2 kNN vs. Bayes-Theorem

Wie bereits in Figure 1 zu sehen war, hat Merkmal 4 mit dem kNN-Klassifikator bei k=3 eine Fehlerrate von 6%. Demzufolge schneidet der kNN-Klassifikator in diesem Fall besser ab.

Wenn man sich die Verteilung des 4. Merkmals ansieht (beispielsweise am 6. Bild auf Seite 2, nur vertikale Verteilung), so erkennt man, dass die 2 Klassen halbwegs getrennt sind, bis auf die Ausläufer, die praktisch über der Grenze und im Feld der anderen Klasse liegen. Da der Bayes-Theorem-Klassifikator in unserem Fall nur mit dieser Grenze arbeitet, werden diese Ausreißer stets falsch klassifiziert, während sie mit dem kNN-Klassifikator eventuell richtig klassifiziert werden können.