Prof. Dr. Katharina Morik,
Prof. Dr. Claus Weihs
Dr. Wouter Duivesteijn,
M.Sc. Sarah Schnackenberg,
B.Sc. Melanie Dagge

Dortmund, 09.07.15

Abgabe: bis Fr, 17.07.15, 10:00 Uhr an
schnackenberg@statistik.tu-dortmund.de
und wouter.duivesteijn@tu-dortmund.de

Übungen zur Vorlesung

Wissensentdeckung in Datenbanken

Sommersemester 2015

Blatt 13 (Statistik und Informatik)

Aufgabe 13.1 (5 Punkte)
Im EWS liegt der bekannte Schweizer-Banknoten-Datensatz `bank.txt`. Er enthält die Ergebnisse von Längenmessungen an 200 Schweizer 1000-Franc-Scheinen (100 echten und 100 gefälschten) in der Einheit mm. Welche Längen genau gemessen wurden, können Sie der Infodatei (`info.txt`) sowie dem Bild in „Figure 1.1“ entnehmen.

a) Führen Sie Hauptkomponentenanalysen für den gesamten Datensatz durch und zwar sowohl

   - auf der Basis von Kovarianzen als auch
   - auf der Basis von Korrelationen.

   Dies ist in R mithilfe der Funktionen `princomp` und `prcomp` möglich.

b) Wieviele Hauptkomponenten würden Sie wählen, um eine Dimensionsreduktion durchzuführen? Schauen Sie sich dazu den `screeplot` und den Prozentsatz der erklärten Gesamtvarianz (z. B. mit der Funktion `summary`) an.

   Für welche der beiden Hauptkomponentenanalysen aus Teil a) sollten die Loadings überhaupt interpretiert werden? Interpretieren Sie die Loadings der ersten Hauptkomponente.

c) Erstellen Sie für beide Hauptkomponentenanalysen aus Teil a) einen Biplot (in R mit der Funktion `biplot` möglich). Welche Scores-Struktur liegt vor? Vergleichen und interpretieren Sie die Biplots.

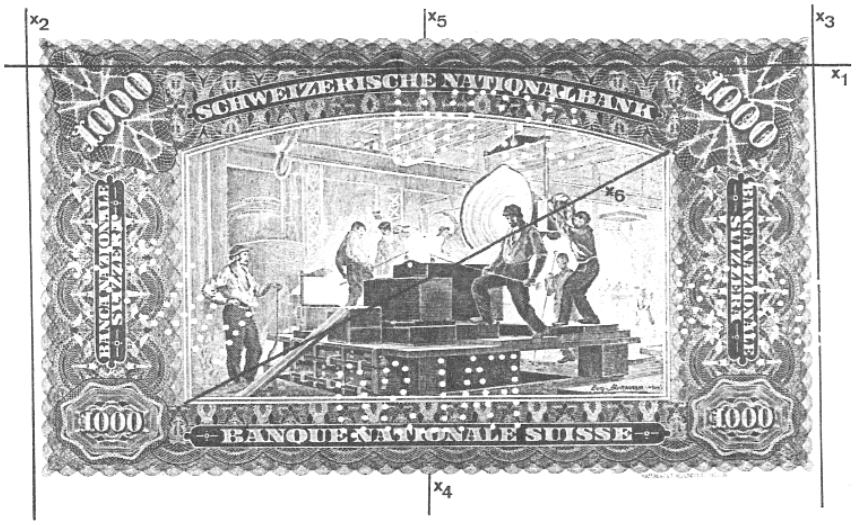

Figure 1.1 An old Swiss 1000-franc bill

Aufgabe 13.2 (2 Punkte)
Wählen Sie einen beliebigen Anwendungsfall für die real-zeitliche Verarbeitung von Datenströmen aus Handel, Logistik, Industrie, Infrastruktur oder Social Media aus.

Beschreiben Sie dafür jeweils eine Aggregationsaufgabe und eine Vorhersageaufgabe und gehen Sie jeweils auf folgende Punkte ein:

- Welche Vorteile bzw. Nachteile bringt die Real-zeitliche Betrachtung der beschriebenen Aufgabe?
- Welche Daten werden benötigt und könnte es Probleme bei der Zuverlässichkeit der Daten geben (Concept Drifts oder Sensorausfälle)?
- Welche Grenzen von Latenz und Durchsatz müssen eingehalten werden?
- Können die benötigten Ressourcen (Kommunikation, Speicher, CPU, Energie) für die Aufgabe einfach skaliert werden oder gibt es Einschränkungen bei der Skalierung der beschriebenen Aufgabe (Weltweiter Einsatz, nur ausgewählte Prozesse oder Kunden, Sampeln nur alle 10 Sekunden...)?

Aufgabe 13.3 (3 Punkte)
Die Zeichen ('a','b','c','.',' ') des folgenden Datenstroms sollen gezählt werden:
D= It is time to apply these design principles consciously from the start instead of rediscovering them each time.

a.) Zählen Sie die Zeichen einmal Vollständig und einmal mit LossyCounting (Wie in der Vorlesung beschrieben).

b.) Beschreiben Sie 3 Einsatzmöglichkeiten von Zählen in anderen Algorithmen bzw. Verfahren.