

*Prof. Dr. Vladimir Shikhman*  
*Professur für Wirtschaftsmathematik*  
*Technische Universität Chemnitz*

*Übungsleiter: David Müller*  
*david.mueller@mathematik.tu-chemnitz.de*

**Mathematische Grundlagen von Big Data Analytics (SS 2018)**  
**Übung 8: Classification II**

1) Sie arbeiten in der Controlling Abteilung eines Unternehmens. Die Marketingabteilung möchte den Absatz des produzierten Produktes erhöhen und dafür gezielt mit der Produktqualität werben. Hierfür muss selbstverständlich eine konstante Qualität sichergestellt werden. Sie werden von Ihrem Chef mit der Überprüfung vertraut, wobei Sie natürlich nicht jedes Exemplar begutachten können, sondern lediglich basierend auf einer Stichprobe Ihre Entscheidung treffen müssen. Aus der langjährigen Erfahrung bezüglich des Produktes sind die Wahrscheinlichkeiten für ein für die Zustände "defekt" und "intakt" bekannt.  $\Theta$  bezeichne die Zufallsvariable, welchen den wahren Produktzustand beschreibt:

$$Pr(\Theta = def) = \pi, \quad Pr(\Theta = int) = 1 - \pi.$$

Zwecks Ihrer Kontrolle lassen Sie ein Testgerät installieren und führen  $n$  unabhängige Messungen durch. Dabei erkennt das Gerät ein defektes Produkt mit der Wahrscheinlichkeit  $p \in [0, 1]$  und die Funktionsfähigkeit eines intakten Produktes mit  $q \in [0, 1]$ . Die Zufallsvariable  $X$  steht für die Anzahl der Defektmeldungen  $0 \leq k \leq n$ .

- a) Geben Sie die Formeln für die bedingten Wahrscheinlichkeiten von genau  $k$  Fehlermeldungen an.
- b) Formulieren Sie eine Entscheidungsregel für das Ausmustern eines Produktes.
- c) Berechnen Sie basierend auf b) eine Formel in Abhängigkeit von  $n$ , mit der Sie entscheiden ab welcher Anzahl von Defektmeldungen Sie das Produkt aussortieren. (Hinweis: Benutzen Sie den Satz von Bayes)
- d) Angenommen Ihnen steht eine sehr große Anzahl von Daten zur Verfügung (Big Data). Bestimmen Sie den für die Aussonderung benötigten Prozentsatz.

2) Nach dem Wechsel Ihres Arbeitgebers stehen Sie nun erneut vor der Aufgabe die Qualitätsüberprüfung. Ihnen wird mitgeteilt, dass im Mittel 5 % der Produkte defekt sind. Das verfügbare Testgerät erkennt ein defektes Produkt zu 90 % und ein intaktes zu 70 %.

- a) Wieviele Defektmeldungen müssen aufkommen, damit das Produkt bei 10 Testmessungen ausgesondert wird?

b) Für eine sehr große Anzahl von Daten: Ab wann wird ein Produkt ausgemustert?

3) Zeigen Sie, dass man mithilfe benötigter Annahmen über den Satz von Bayes im Klassifikationsproblem zur linearen Diskriminanzanalyse kommt. Verweisen Sie hierbei explizit auf die Annahmen.