UE Statistische Mustererkennung
WS 2021
Angaben zur 2ten Aufgabengruppe

1 Aufgabe UE-II.1

Generieren Sie je 1000 Stichproben (*samples*) mit Umfang 5/30/100/500 für die Normalverteilung $N(\mu, \sigma^2) = N(4, 3)$. Hierzu steht unter MATLAB die Funktion *normrnd* und unter Python die Funktion *numpy.random.normal* zu Verfügung. Beachten Sie jedoch, dass diese Funktionen die Standardabweichung σ und nicht die Varianz σ^2 als Argument erwarten.

Dies ist eine einfache Monte-Carlo-Simulation, in welcher wir die wahre Verteilung von Schätzern durch die empirische Verteilung von 1000 zufällig generierten Schätzwerten annähern.

a) Stellen Sie die empirische Verteilung folgender Schätzer:

   I) des Stichprobenmittels,

   II) der Stichprobenvarianz und

   III) des Stichprobenmedians

   für jeden der 4 Stichprobenumfänge in Histogrammform dar (Funktion *hist* unter MATLAB bzw. *numpy.histogram* unter Python).

   Plotten Sie auch die wahre DF des Stichprobenmittels und der Stichprobenvarianz (achten Sie auf die Skalierung).

b) Berechnen Sie für jeden der Schätzer dessen Mittel und Standardabweichung aus den je 1000 Realisierungen. Vergleichen Sie Ihre Ergebnisse mit den theoretischen Resultaten bzg. der Verteilung des Stichprobenmittels (I) und der Stichprobenvarianz (II); für den Medianschätzer (III) vergleichen sie dessen Mittel mit dem wahren Median (die theoretische Varianz des Stichprobenmedians wurde in der VO nicht besprochen und braucht daher nicht bestimmt zu werden). Wie verhält sich die Verteilung des Stichprobenmittels zu jener des Stichprobenmedians?

c) Berechnen Sie für jedes Stichprobenmittel (I) ein symmetrisches 95%-Schätzintervall, wobei Sie die Populationsvarianz $\sigma^2 = 3$ als bekannt voraussetzen können. Wie groß ist – für jeden der 4 Stichprobenumfänge – der Anteil der Stichproben, deren Schätzintervall das wahre Mittel 4 überdeckt?

   Wie ändert sich das Ergebnis, wenn die Populationsvarianz aus der Stichprobe geschätzt werden muss, jedoch weiterhin die Normalverteilungsquantile zur Berechnung der Intervallgrenzen herangezogen werden?

d) *non-parametric bootstrap*
Generieren Sie jeweils eine Stichprobe für jeden Umfang 5/30/100/500. Erzeugen Sie aus jeder dieser Stichproben – sagen wir vom Umfang N – durch Auswahl mit Zurücklegen *(sampling with replacement)* 1000 Stichproben vom Umfang N, sogenannte *bootstrap samples*. Der Unterschied zum ursprünglichen Ansatz liegt also darin, dass die 1000 bootstrap samples nicht unabhängig aus der – als bekannt vorausgesetzten – zugrundeliegenden Verteilung generiert werden, sondern durch *resampling* aus einer Stamm-Stichprobe. Gehen Sie weiter wie unter a) und b) vor. Welche Eigenschaften bzw. Parameter der zugrundeliegenden Verteilung lassen sich aus den *bootstrap samples* gut ableiten, welche nicht?

Hinweis: für das resampling stehen unter MATLAB: *datasample*, unter Python: *numpy.random.choice* zur Verfügung.

2 Aufgabe UE-II.2

Gegeben seien zwei Klassen mit mit *priors* $P(\omega_1) = 0.3$ und $P(\omega_2) = 0.7$, wobei sich ein interessierendes Merkmal X für jede der beiden Klassen gemäß $X_i \sim N(\mu_i, \sigma_i^2)$ mit $(\mu_1 = -7, \sigma_1^2 = 30)$ sowie $(\mu_2 = 4, \sigma_2^2 = 13)$ verteile.

a) Schreiben Sie Funktionen zur Berechnung der $p(x|\omega_i)$, der Randverteilung $p(x)$ (evidence) des Merkmals sowie der *posteriors*.

b) Klassieren Sie anhand der folgenden Merkmalsausprägungen unter Verwendung der *Bayes decision rule*: $-15, -10, -5, 0, 5, 10$

c) Ermitteln Sie grafisch (durch Plotten der *posteriors*) die Entscheidungsgrenze.

d) Plotten Sie die *evidence*, also die Rand-DF von X.

3 Aufgabe UE-II.3

Das Merkmal X gebe die in einem Antikörpertest gebundene Menge an Antikörpern an. Ein größerer Wert weist also auf eine Infektion hin.

Die Verteilung von X für gesunde H_0 und infizierte H_1 Probanden sei bekannt und durch $X|H_0 \sim N(4,1)$ bzw. $X|H_1 \sim N(5,1)$ gegeben.

Führen Sie für die beiden Prävalenzpaare $P(H_0) = 0.9, P(H_1) = 0.1$ sowie $P(H_0) = 0.99, P(H_1) = 0.01$ folgende Berechnungen bzw. Visualisierungen durch:

a) Plotten Sie den positiven Vorhersagewert als Funktion der Entscheidungsgrenze.

b) Bestimmen Sie die optimale Entscheidungsgrenze (graphisch) und die Bayes-Fehlerrate.

c) Plotten Sie die ROC-Kurve.

Hinweise:

- $P(+|H_1)$ ergibt sich durch Integration der likelihood von H_1 über die Entscheidungsregion von H_1

$$P(+|H_1) = \int_{x^*}^{\infty} p(x|H_1)dx, \tag{1}$$

  wobei x^* die Entscheidungsgrenze bezeichnet.

- Zur Erinnerung: die Dichtefunktion wird im Englischen als *pdf*, die Verteilungsfunktion als *cdf* bezeichnet. Integrale der DF der Normalverteilung lassen sich in Python z.B. mittels *scipy.stats.norm.cdf*, in MATLAB mit *normcdf* berechnen. Das Integral über den rechten Schwanzbereich $[x, +\infty[$ erhalten Sie mit $1 - cdf(x)$.

- Der Zusammenhang zwischen dem positiven Vorhersagewert und den gegebenen Größen läßt sich vermittels des Bayes-Theorems herstellen:

$$P(H_1|+) = \frac{P(+|H_1)P(H_1)}{P(+|H_1)P(H_1) + P(+|H_0)P(H_0)} \tag{2}$$

4 Aufgabe UE-II.4

Die **Bernoulli-Verteilung** $B(1,\theta)$ mit Parameter $0 \leq \theta \leq 1$ beschreibt einen Zufallsversuch, der nur zwei mögliche Ausfälle haben kann (z.B. Münzwurf). Für eine Bernoulli-verteilte Zufallsvariable $X \sim B(1,\theta)$ gilt:

$$P(X=1) = \theta, P(X=0) = 1-\theta, \tag{3}$$

mit

$$\begin{aligned} \mathcal{E}[X] &= \theta & (4)\\ Var[X] &= \theta(1-\theta). & (5) \end{aligned}$$

Die Summe $Z = \sum_{i=1}^{N} X_i \sim Bi(N,\theta)$ von N iid Bernoulli-Variablen X_i ist **binomial-verteilt**:

$$P(Z=n) = \left(\begin{array}{c} N \\ n \end{array} \right) \theta^n (1-\theta)^{(N-n)} \tag{6}$$

a) Leiten Sie – **unter Verwendung der aus der VO bekannten Eigenschaften** des Erwartungswertes und der Varianz der Summe von iid Zufallsvariablen – den Erwartungswert und die Varianz der Binomialverteilung her.

b) Zeigen Sie dass der relative Anteil Z/N der "guten" Ausfälle ein erwartungstreuer, asymptotisch konsistenter Schätzer des Parameters θ ist.