MA4022/ MA7022   Data Mining and Neural Networks
Computational Task 1, 2020
Due till 14.02.2020
100 marks available

Psychological predisposition to heroin use.

Your work should answer the question: Does the psychological predisposition to drug consumption exist?

Nowadays, after many years of research and development, psychologists have largely agreed that the personality traits of the modern Five Factor Model (FFM) constitutes the most comprehensive and adaptable system for understanding human individual differences. The FFM comprises Neuroticism (N), Extraversion (E), Openness to Experience (O), Agreeableness (A), and Conscientiousness (C). The five traits can be summarized thus:

N *Neuroticism* is a long-term tendency to experience negative emotions such as nervousness, tension, anxiety and depression (associated adjectives: anxious, self-pitying, tense, touchy, unstable, and worrying);
E *Extraversion* manifested in characters who are outgoing, warm, active, assertive, talkative, and cheerful; these persons are often in search of stimulation (associated adjectives: active, assertive, energetic, enthusiastic, outgoing, and talkative);
O *Openness to experience* is associated with a general appreciation for art, unusual ideas, and imaginative, creative, unconventional, and wide interests (associated adjectives: artistic, curious, imaginative, insightful, original, and wide interest);
A *Agreeableness* is a dimension of interpersonal relations, characterized by altruism, trust, modesty, kindness, compassion and cooperativeness (associated adjectives: appreciative, forgiving, generous, kind, sympathetic, and trusting);
C *Conscientiousness* is a tendency to be organized and dependable, strong-willed, persistent, reliable, and efficient (associated adjectives: efficient, organised, reliable, responsible, and thorough).

Two additional characteristics of personality are proven to be important for analysis of substance use, Impulsivity (Imp) and Sensation-Seeking (SS).
Imp *Impulsivity* is defined as a tendency to act without adequate forethought;
SS *Sensation-Seeking* is defined by the search for experiences and feelings, that are varied, novel, complex and intense, and by the readiness to take risks for the sake of such experiences.

Seven psychological traits were used to characterise the participants: N, E, O, A, C, Imp, and SS.

Task 0. Preparation of dataset for analysis.

The dataset is online http://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29

There are more attributes than you need. Prepare the table. For every participant, leave the following information: 7 psychological traits and *heroin* user/non-user (*in the last year*).

The user/non-user classification will be the main task.

Task 1. Descriptive statistics: For both classes (users and non-users) find the mean values of the 7 attributes and their standard deviations. Evaluate the 95% confidence intervals *for mean values.*
(Take the definitions from any elementary textbook in statistics. A very simple online tutorial about 95% confidence interval is here:
http://www.itl.nist.gov/div898/handbook/eda/section3/eda352.htm
A very simple textbook, The Little Handbook of Statistical Practice, is here:
http://www.jerrydallal.com/LHSP/LHSP.htm).
Create graphical illustration ("psychological profiles" of heroin users and non-users with confidence intervals). **(20 marks)**

Task 2. Report, which differences between these means for users and non-users are significant. For significance evaluation use p-values. **(10 marks)**

Task 3. Try to create predictors user/non-user by one attribute (7 such predictors). For this purpose, create histograms for each attribute and each class and select the best threshold for each attribute x for the decision rule: if x>a then one class (users or non-users) and if x<a then another class (non-users or users) (the optimal cut). Find the classification error for each attribute. Which attribute gives the best prediction? Arrange the attributes in their prediction ability. **(15 marks)**

Task 4. Test 1NN and 3NN classification rules. Present the classification errors. Which rule is better? **(20 marks)**

Task 5. Find in the literature description and explanation of Fisher's linear discriminant. Read, understand and write a comprehensive description of the algorithm with main formulas and explanation (not more than 1 page!) **(10 marks)**

Task 6. Apply Fisher's linear discriminant to the prepared data set. Analyse the quality of classification. Compare to 1NN and 3NN methods. **(15 marks)**

Extra 10 marks for clear and well-written report.