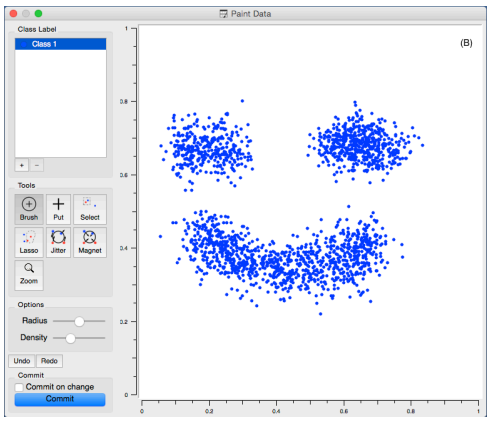
Final Exam for Introduction to Data Mining, November 2020

Part 1. Multiple-choice questions

Each question has one correct answer. Correct answers are worth +3 points, wrong answers deduct 1 point. 15 points from this part are required for a passing grade. Submit answers as: 1 b, 2 c, 3 b, etc.

1. We created some data (see the Paint widget on the right) to illustrate one of the techniques for data mining - see the picture on the right. Which technique did we have in mind?



   a) Clustering
   b) Naive Bayes
   c) Classification trees
   d) Classification with random forest

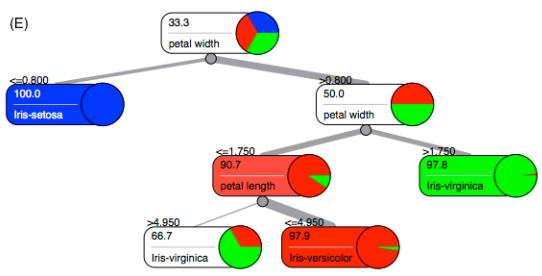
2. Somebody would like to make sense of perfumes, so he bought 20 different bottles and asked his wife who had a great sense of smell (and also liked to smell great) to compare each pair and assess their pairwise (dis)similarities on scale from 0 to 10. Several hours later, she's done. Now what?

   a) Hierarchical clustering
   b) Random forests
   c) Scatter plot
   d) K-means clustering

3. Why did not he use multidimensional scaling (MDS)?

   a) The data sample is too small for MDS.
   b) MDS cannot yield any useful information about relations between perfumes.
   c) MDS is not applicable because he does not have any information about individual perfumes (i.e. variables describing them), but just a matrix of their (dis)similarities.
   d) No idea. Cannot be any of the above. Maybe he has not attended our course and didn't know about MDS.

4. His wife also likes flowers. What does the picture on the right have to do with them?



   a) This is a classification tree for the famous Fisher's data on irises.
   b) This is a result from hierarchical clustering of Iris data.
   c) This is the nonsense he got when he tried using MDS on perfumes. (We told him he shouldn't use it, but he wouldn't listen.)
   d) This is a two-layer neural network for classification of Iris data.

5. The guy who spends too much on perfumes and flowers gets loans from a bank that uses one of the following systems to automatically decide whether a client is credit-worthy. Which?

   a) K-Means
   b) Scatter plot
   c) Random forest
   d) Silhouette plot

6. The sign above the bank says: "Enter in pairs". For every pair of customers, the bank gives the loan to one who is more likely to return it, according to some predictive model. Which of the following measures is most useful for assessing the performance the model in such context?

   a) Classification accuracy
   b) Precision
   c) Recall
   d) Area under ROC curve

7. The above bank also sends spam to selected groups of clients; different groups get different spam. Which method did they use to form the groups?

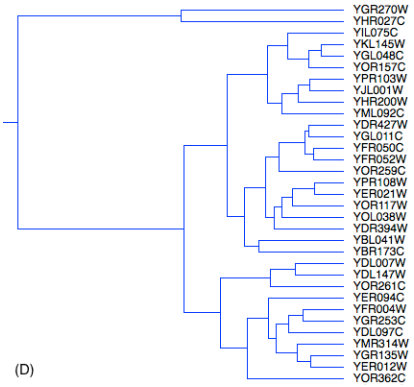
   a) K-Means
   b) Scatter plot
   c) Random forest
   d) Silhouette plot

8. Talking about the silhouette plot: the length of the line in the silhouette plot reflects:

   a) how central an instance is to the cluster
   b) how similar the data instance is to other clusters
   c) the predictive power of each data instance
   d) the variance of the data instance

9. Does the figure on the right show a silhouette plot?

   a) Yes, but with labels on the right
   b) No, this is an MDS in vertical layout
   c) No, this is a dendrogram of hierarchical clustering
   d) No, this is a nomogram for logistic regression.



10. So, which of the following statements is generally true for logistic regression with different levels of regularization? (Two of them are *often* true, but one is *almost always* true.)

    a) strong regularization improves performance on training data
    b) strong regularization improves performance on testing data
    c) weak regularization improves performance on training data
    d) weak regularization improves performance on testing data

Part 2: Open-answer questions

Answer each question with a sentence or two, but not more than three.

1. Larger classification trees take into account more features than smaller ones, and hence provide for more informed decisions. Current tree induction algorithms that strive to find small trees thus don't make sense. Object!

2. Suppose that we have a problem in which predictors are expensive to measure. We want to select a smaller subset of variables that will still yield high predictive accuracy. Will we use naive Bayesian classifier or logistic regression? Why? When using the chosen method, which approach will we use to regulate the number of variables?

3. We have a prediction problem in which the outcome is a consequence of a large number of individually insignificant factors. Which method should perform better: classification trees or logistic regression? Why?

4. Consider a problem with two independent variables, x and u, and a dependent variable y. Linear regression would find a model of form $y = b_0 + b_1x + b_2u$. How can you use linear regression to fit a non-linear plane of form $y = b_0 + b_1x + b_2u + b_3x^2 + u\ b_4u^2 + b_5ux$.

5. After answering the previous question, tell us whether linear models can have non-linear boundaries. If no, explain why the trick from the previous question doesn't work. If yes, tell us which model is known for using this trick.

6. If Facebook wanted to cluster its users, would it use k-means or hierarchical clustering? Explain why. (Let us assume they use standard methods without any tweaks.)

7. Assume you have some data in which instances are split into groups. Explain the meaning of the silhouette coefficient for a point and for a group.

Part 3. Hands-on Data Analysis

Although we'd like to keep the exam tool-agnostic, this part requires Orange because it is too difficult to formulate concrete questions that can be answered in a short time and are not biased towards a certain tool.

Answer each of the below questions in a paragraph of reasonable length. Provide evidence. Preferably, attach an Orange schema and the report file.

In 2013, friends of students of the Statistics class at FSEV UK participated in a study about their preferences and habits. Results are collected in the file reponses.csv, which is provided with this exam.

1. Remove variables "height" and "weight" and run k-means clustering. How many clusters does the data contain?

2. What do the clusters actually represent? With which variable are they strongly related?

3. Would the clustering be similar if this variable is removed? Demonstrate. What does this show?

4. To further investigate the reality of this clustering, show an MDS in whose computation you ignore height, weight and the above variable, but in the MDS, color the points according to the above variable and set the size to represent weight. (Move these variables to meta attributes so they are not used in computation of MDS but still available for visualization.)

5. Use a suitable machine learning / predictive modeling algorithm to discover which variables are most strongly correlated with gender. Justify the selection of the algorithm (you can also explain why you rejected other alternatives). Explain what you found.

6. Find interesting pairs of variables using Sieve. Browse through the list and point at some obvious pairs (variables that indicate a similar thing), expected relations and perhaps some stereotypical ones.

7. Cluster variables using hierarchical clustering; use Pearson correlation as a measure of distance. Set a reasonable cut-off point in clustering. Comment the findings.