Homework #1:

Exploratory analysis and simple Orange workflows

Find any data on the web, or retype it from any other data source, construct it on your own or find it in your lab. The data set could be small, but it should contain at least 20 samples and 4 features. It would be best if the data set is real, say, is relevant to your research or interest, comes from some publication, web page, or from some problem that you have worked on before. Use any data exploration techniques that we have learned about in our first lecture to find a visualization that shows anything interesting in this data. Interpret this visualization. Was your finding expected? Was it reported anywhere? Or is this something entirely new?

Submit your homework as a short report in PDF. The report should include a title of the homework ("Exploratory analysis and simple Orange workflows"), your name and email, and following sections:

Problem, a paragraph that describes the problem;

Material, where and how did you get the data, report on the size of the data (number of samples and features), and on the type of the features included;

Methods, up to two paragraphs on the methods you have used, preferably include a workflow and comment on it;

Results and Discussion, choose an interesting visualization or two and comment on them.

Report should not exceed two pages (this limit is strict!), use 11 pt Times. Submit your report as a PDF document; name this document as lastname-firstname-1.pdf where lastname is your lastname and first name your first name. Email the report to bzupan@gmail.com with subject "STAT-HW1" (omit the quotes).

Homework #2: Classification Accuracy

Consider two data sets from Gene Expression Omnibus (GEO, data sets with accession numbers GDS3713, GDS4182). GEO includes data sets on people or model organisms. For instance, GDS3713 data is on smoking and its effect on females. In rows are profiles of B lymphocytes that are characterized with a set of numbers. Numbers represent how much each gene (genes are attributes in the data) in the lymphocytes is expressed. For these homework you treat the data as is, no need to understand them. We just considered them because they contain many attributes and few samples, and the problem of such "fat" data is interesting for data mining. If you are courious, you may still look up the two data sets on GEO (say, for GDS3713, go to http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS3713) to obtain more information about the data sets.

Please download the data from http://bit.ly/1UkLQYV.

Both data sets contain a binary (two-valued) class, and the task is to predict a class value based on gene expression profile. In brief, the data contains tissue samples (in rows) and for each tissue sample they have recorded expression of genes. Each tissue is classified into one of two classes (say, disease or healthy). The question is if one can classify the tissue based on its gene expression profile. Questions like these are important for systems biology and clinical decision making, as profiling tissues through gene expression can help us to improve diagnostic and prognostic tools.

For the homework, construct a workflow where you use cross-validation to estimate the classification accuracy of the classification trees. Report on this accuracy for each of the two data sets. Tell us for which of them the machine learning method of classification trees performed better. Please explain how have you reached this conclusion.

Submit your homework as a short report in PDF. The report should include a title of the homework, your name and email, and following sections:

Problem, your short description of the problem (up to two sentences),

Material, one or two sentences on the data, also report on the size of the data (number of samples and features), and on the type of the features included;

Methods, a paragraph on methods you have used, preferably include a screenshot of the workflow;

Results and Discussion, one paragraph on results. Make sure you end with a sentence with your conclusion (for which data set the classification trees were a better predictor).

Report should not exceed one page (this limit is strict!), use 11 pt Times. Submit your homework as a PDF document to bzupan@gmail.com with subject "STAT-HW2" (please use just this string without qoutes in the subject, do not augment it with anything or change it).

A side note: the two data sets are as those coming from GEO, but for the size of the data sets we have reduced the number of features (genes) in the data.

A hint: Janez claims he's a psychic. Blaz introduced him to a student he met in front of Faculty of Medicine in Ljubljana, and Janez says he can tell, with over 95% accuracy, whether she studies medicine or machine engineering. Blaz doesn't seem impressed. Why?

Homework #3:

Classifiers and their Decision Boundaries

Consider the following classifiers:

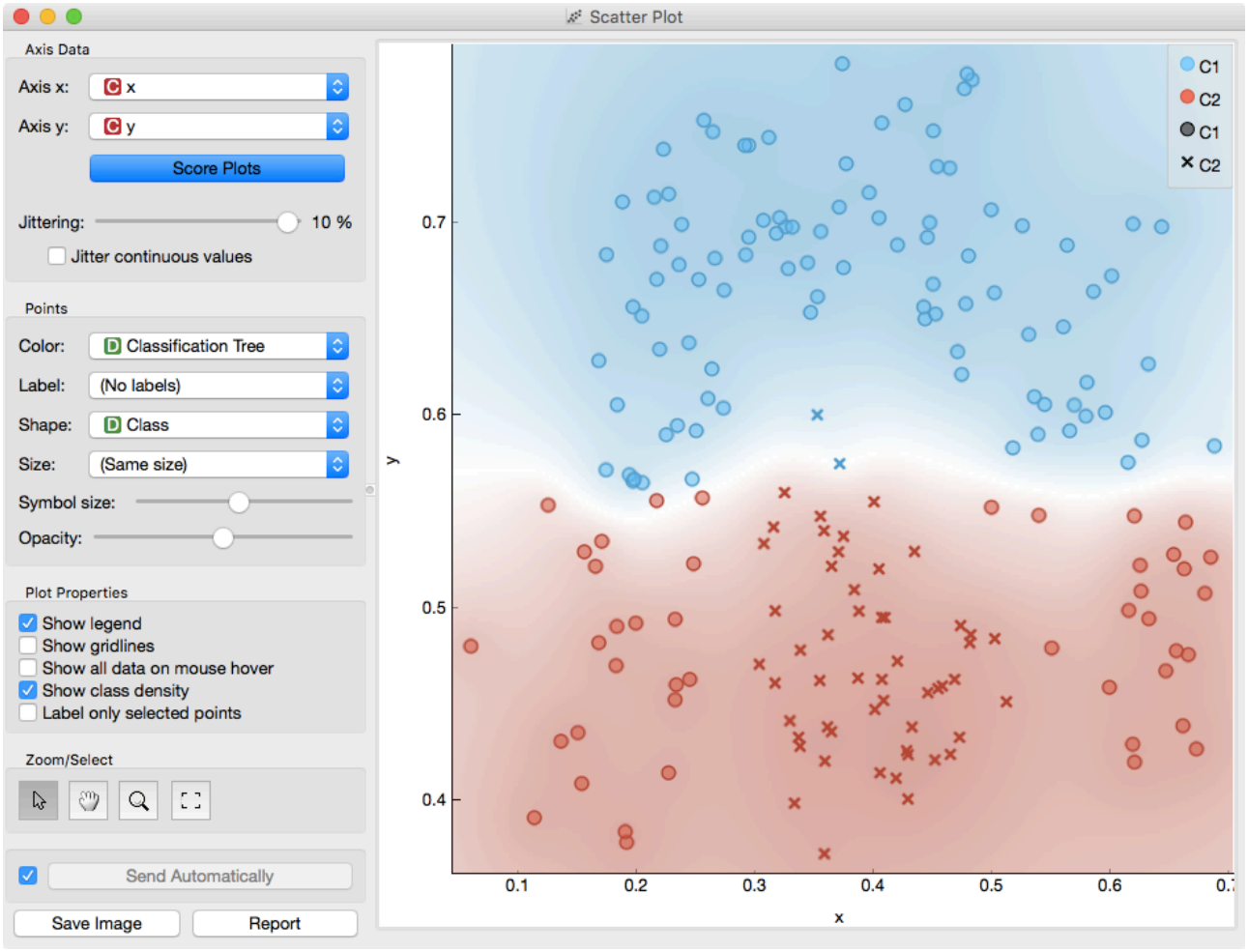
1. classification tree of the depth 1 (so-called a "stump", set the parameter "Limit the depth to" to 1)
2. classification tree of the depth 3
3. logistic regression
4. SVM with RBF (radial basis function) kernel and g=1
5. random forest with 50 trees
6. nearest neighbors classifier with number of neighbors set to 5

For each of the data sets above paint:

A. a data set where the classifier finds the "right" decision boundary

B. a data set where the classifier failed to find the "right" decision boundary

Demonstrate A and B using scatter plots. A minimal schema contains the Paint Data, Predictions and Scatter Plot, plus a learner (say, Classification Tree, receiving the data and passing a classifier to the Predictions). In the scatter plot, you can color the dots by the predicted class and set the shape to represent the true class value. For instance, for classification tree, the Scatter Plot widget could look something like the following:



(In the homework, just show us the graphs, not the entire widget).

Submit your homework as a short report in PDF. The report should include a title of the homework, your name and email, and following sections:

Problem, in your words, describe what was the task of this homework (one short paragraph),

Material, list the classifiers that are being tested;

Methods, provide the schema that was used for testing the classifiers and describe it in one or two sentences;

Results and Discussion, report on results. Mainly, just provide screen shot of success/failure (A and B) for each of the classifiers. With each graph, you may report also on AUC scores you get using your data and 10-fold cross-validation. For now, it is sufficient to know that AUC scores close to 1.0 are excellent, and scores around 0,5 are very poor. It may happen that for some of the classifiers you would not be able to paint a data set matching A and B. If this is the case, please provide your intuition why. Conclude with a short paragraph summarizing your findings.

Report should not exceed three pages (this limit is strict!), use 11 pt Times. Submit your homework as a PDF document to bzupan@gmail.com with subject “STAT-HW3”.

Happy painting!

Homework #4: ROC Curve

Warm-up exercise

You won a lottery prize, you just don't yet know which. The probability that you'll get $5 is 0.8 and the probability that you'll get $10 is 0.2. What is the expected value of the prize?

Solution: if you get $5 with a probability of 0.8 and $10 with a probability of 0.2, the expected prize is the weighted average of the prizes, 0.8 x $5 + 0.2 x $10 = $6.

Now for the homework:

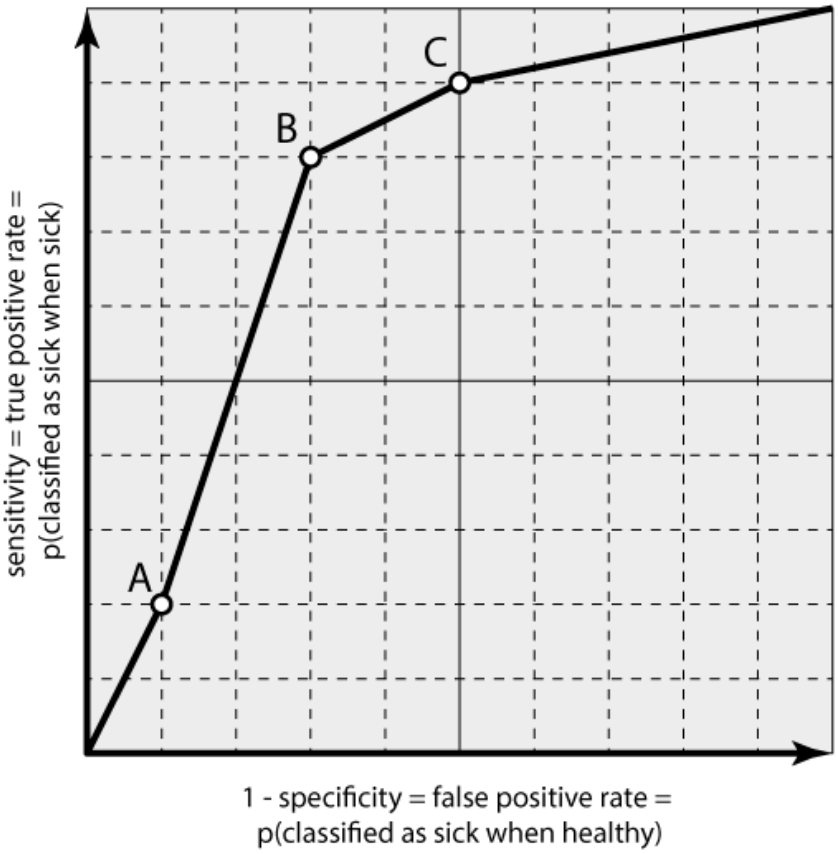
Sara treats hamsters for Chomsky disease. About one half of hamsters she sees have this disease (luckily, the disease is not serious; it only makes hamsters run backwards on the running wheel).

She can make two kinds of mistakes:

- If she fails to detect the disease when it's present, the associated cost (lawsuits etc.) is $1000.
- If she treats a hamster that is actually healthy, the cost (lawsuits etc.) is $600.

(Don't worry about her, in both cases she charges enough to survive.)

Her choice whether to administer the cure will be based on the classifier that predicts the probability of the disease from the observed symptoms. The classifier she uses is not perfect, as shown in the ROC curve.



1. What is the false positive rate of each point marked on the ROC curve?
2. What is the true positive rate of each point?
3. How do you compute false negative rate from one of these two?
4. Each of the two kinds of Sara's mistakes is associated with one of the above three probabilities. For each kind, which is the associated probability?
5. What is the expected cost of mistakes for each point on the ROC curve?

Advanced problem (up to 20 bonus points)

Sara has accidentally put a sick and a healthy subject (that is, hamster) in the same cage. Now she doesn't know which is which. She is going to diagnose both hamsters and administer the cure to the one which she believes is more likely to be sick. What is the probability that she'll pick the ***wrong*** one?

Submit your homework as a short report in PDF where you answer to above questions. The report should include a title of the homework, your name and email. It should be one page long (this limit is strict!), use 11 pt Times. Submit your homework as a PDF document to bzupan@gmail.com with subject “STAT-HW4”.

Homework #5: k-Means Clustering

Download data set GDS4168 from http://bit.ly/1ivkCxR (the documentation on this data set is at at http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS4168).

1. Use classification tree and logistic regression to construct a model for this data. Use Test & Score to estimate their accuracy.

2. Run k-means clustering. How well do the clusters correspond to classes? In the lectures we used Sieve to check this. This data set has very few instances, so Sieve could be misleading (try it!) We recommend using the Distributions or Box plot widget instead. In the k-Means, add the cluster as a feature, not a class.

3. Did both approaches work well? Could you propose the reason why - or why not?

Submit your homework as a short report in PDF where you answer to above questions. The report should include a title of the homework, your name and email. It should be one page long (this limit is strict!), use 11 pt Times. Submit your homework as a PDF document to bzupan@gmail.com with subject “STAT-HW5”.

Homework #6: Image Analytics

Gather a collection of 30 to 100 images, preferably related to biology, medicine, natural sciences, or even better, your research or interest. If you can not find those, just find any image set of your liking on the web or even in your photo album. Make sure they are in jpg or png format, and not too big to avoid overburdening the embedding server. Put images in a folder (and sub-folders, to indicate classes), and load them using Import Images widget. Check them out in Image Viewer to make sure they are loaded OK.

Now apply the skills you have learned in this class to get some insight into your image set.

1. Cluster images using either hierarchical clustering or k-means, and comment on the quality and meaningfulness of clusters.
2. Sort your images into groups (classes) by placing them in appropriate sub-folders. Can image classes be predicted from feature-characterized images? Report on cross-validated or leave-one-out accuracy. Perhaps you can also comment on types of mistakes that your selected learner makes.
3. Project images into two-dimensional space (use either PCA or MDS). Tell us if the projection makes any sense. For illustration, you can include the graph with projection and labels of the plots, and comment on the groups you can spot from this visualization.
4. Do anything else that you can think of and makes sense.

Submit your homework as a short report in PDF where you tell us about the data and the analysis results. The report should include a title of the homework, your name and email. It should be at most three pages long (this limit is strict!), use 11 pt Times. Submit your homework as a PDF document to bzupan@gmail.com with subject “STAT-HW6”.