Tasks to MMAD

Test 1 - Models

Task 1

A. Generate data from scalar regression model

$$y_t = bu_t + ay_{t-1} + k + e_t$$

where $b = 0.6$, $a = 0.4$, $k = 2$, u_t is normal variable with expectation -1 and variance 1 and e_t is normal noise with expectation 0 and variance 0.2.

B. Using the generated data perform estimation of parameters of the following models.

1. Model with the same structure as was used for simulation.
   Here, the estimated parameters should have the same values as those used for simulation.
2. Change the structure of the estimated model (higher or lower model order).
   Now, the estimates are approximations and the resulting estimates will differ from those in simulation. The goal of estimation are not the same parameters but good prediction of the output.
3. In each step of estimation remember the current estimates. After the estimation is finished plot the evolution of the estimates.
   Notice, how the parameters estimates are gradually refined.

Task 2

A. Generate data from categorical model

$$f\left(y_t|u_t\right) = p_{y_t|u_t}$$

where $y_t \in \{1, 2\}$, $u_t \in \{1, 2\}$, $f\left(u_t\right) = [0.4,\ 0.6]$ and

$p_{y\|u}$	$u = 1$	$u = 2$
$y = 1$	0.1	0.2
$y = 2$	0.9	0.8

B. Use the generated data for estimation of a categorical model with the same structure as for simulation.

Hints:

- For the regression model you can follow the Laboratory Chapter 1, Section 1.1 and reference to Scalar regression model: theory and program.
- For the categorical model Chapter 1, Section 1.4: theory and program.
- Normal generator

$$y(t) = \text{mean} + \text{sqrt}(\text{variance}) * rand(1, 1, 'n')$$

- Categorical generator

$$y(t) = \text{sum}(\text{cumsum}(th(y_t|:)) < rand(1, 1, 'u')) + 1$$

where th is the matrix given in the following table

u	1	2
$y = 1$	$th_{1\|1}$	$th_{1\|2}$
$y = 2$	$th_{2\|1}$	$th_{2\|2}$

Important

For details of Scilab language consult the Scilab help (icon ?). It is very useful.

Test 2 - Initialization of model estimation

Task 1

Consider the flipping coin task (1-head, 2-tail). For the dataset $\{1, 1, 2, 1, 2\}$ show (on the paper) the evaluation of the estimate of the parameter p - probability of 1 (head). Use the initial parameter:

- $p = 0.5$ as strong information,
- $\hat{p} = 0.5$ as week information,
- $\hat{p} = 0.1$ as strong information,
- $\hat{p} = 0.1$ as week information.

Hint: Look at the Laboratory Chapter 3, Initialization of estimation.

Task 2

Generate data from the Poisson model with the parameter $\lambda = 4$ (use the inbuilt function `grand(m,n,"poi",lam)`- see Scilab help).

Perform estimation of the model with

1. very vague but true initial condition,
2. strong and true initial condition,
3. very vague and false initial condition,
4. strong and false initial condition.

Compare the graphs of the evolution of the parameters.

Hint: λ is estimated as expectation. The statistics are S - sum and κ - count. The point estimate is $\hat{\lambda} = S/\kappa$.

Task 3

Simulate data from normal regression model

$$y_t = c_1 x_{1;t} + c_2 x_{2;t} + k + e_t$$

where $c_1 = 2$, $c_2 = -1$, $k = 1$ and $r = .1$ (variance of the noise e). x_t are normal variables with expectations -1 and 3 and variances 1.

Perform estimation with

- true initial parameters and weak/strong information,
- false initial parameters and weak/strong information.

Hint: For estimation consult the Laboratory Chapter 1 Models and their estimation, Section Scalar regression model. Here, the extended regression vector is

$$\Psi_t = [y_t, x_{1;t}, x_{2;t}, 1]' .$$

Test 3 - Prediction

Task 1

Simulate data from the normal regression model

$$y_t = ay_{t-1} + bu_t + k + e_t$$

where $a = 0.7$, $b = 1$, $k = 3$ and noise $sd = .1$. The control variable set as:
1) constant, 2) sinusoidal signal, 3) noise.

At the same time loop (i.e. on-line) perform estimation of parameters and zero-step prediction of the output y_t.

Hints:

Consult Laboratory, section 2.2 - Prediction for regression model.

The sine-signal generator is `sin(2*%pi*t/nd)`

Task 2

For the same model as in the Task 1, with input as a sine-signal plus random noise, perform multistep prediction for a general number np of steps ahead.

Hint: Consult the Laboratory, Section 2.3 - K-step prediction for regression model with unknown parameters. Have a look also at the description of the program!

Test 4 - Classification

Task 1

Simulate data from a multimodal system with two normal components f_1 and f_2 randomly switching with the pointer c_t: $f(c_t) = [0.4, 0.6]$ where $f_1(y_t) = N_{y_t}(\mu = 2, r = 0.5)$ and $f_2(y_t) = N_{y_t}(\mu = 5, r = 1)$. Draw histogram of the dataset. Store the data in the `data.csv` file.

Hint: You must introduce the pointer variable c_t with the given categorical distribution and generate its value. According to it you generate normal y_t from either f_1 or f_2.

For storing you must join together data: `data=[c y];` and use the command `csvWrite(data,'data.csv',';')`.

For extracting the data, use `data=csvRead('data.csv',';');` and `c=data(:,1); y=data(:,2);`

Task 2

1. For the data from `data.csv` perform classification into two clusters described by the components f_1 and f_2 with known parameters (i.e. estimate the values of the pointer $c_{p;t}$ and compare it with the pointer c_t from the simulation).

2. For the data from `data.csv` perform estimation of the components f_1 and f_2 (which now are considered unknown) using the values of the pointer c_t (which is now known). It is classification (learning) with a teacher.

3. Divide the data from data.csv into larger part yL (learning data) and the rest yT (testing data). For yL perform learning with a teacher. Then use the final estimates of component parameters and perform classification for the data yT.

4. For the data from `data.csv` perform mixture estimation. For initialization use estimation with a teacher for the first 10 data records. Then estimate both the pointer and the component parameters. Remember the pointer estimates (classes of classification) and compare them with the simulated pointer. Determine accuracy `Acc=(number of correct classifications)/(number of data).`

Test 5 - Knime (clustering and classification)

Task 1

In the directory "dataCars" find the dataset "cars.csv" (the description can be found in the file "car.names").

1. In KNIME perform Naive Bayes learning for the first 75% data records.
2. For the rest of data perform classification of the target variable "ACCEPT".
3. Evaluate accuracy of classification.

Task 2

In dataset "iris.csv" use the first two variables and perform k-means clustering with five components.

1. In KNIME show scatterplot with detected components distinguished by color and shape.
2. Using icon "csvWriter" copy the data with the detected clusters to disk.
3. In Scilab show histograms of the clustered variables in individual detected clusters.

Task 3

For data data03.csv from the directory DATA estimate the decision tree for the target variable T using variables A, B, C as the explanatory ones.

1. Show the decision tree.
2. Write the decision tree rules.
3. Determine the accuracy of th classification.
4. Correct the values of the target variable to obtain accuracy equal to one.

Task 4

For data data04.csv from the directory DATA perform hierarchical clustering.

1. Show the dendrogram .
2. Create clusters for the Distance Threshold equal to 0.9 (icon Hierarchical Cluster Assigner).

3. Compare the results from the dendrogram with the clustered data in the table (the same icon).

4. Write the points in the clusters obtained.