Msc Supply Chain & Operation Management -Data Science

Exercises

EMLyon Business School - Fall 2026-2027 - Guillaume Metzler

- Fourteen short exercises, **two per algorithm**: k-nearest neighbours, classification trees, regression trees, support vector machines (linear and kernel), random forests, AdaBoost and gradient boosting.
- The seven parts are **independent**. Give **all intermediate computations** and round the final results to 3 decimals.

Part A – k-nearest neighbours

distances, voting, scaling, choice of k

Exercise A.1 Late shipments: k-NN classification

A carrier wants to predict whether a shipment will be **late**. Eight past shipments are described by the distance (in hundreds of km), the number of transfers, and whether they were late (1) or on time (0). A new shipment Q has a distance of 5 and 4 transfers (Figure 1).

shipment	distance	transfers	late
S1	1	4	0
S2	8	3	1
S3	1	3	1
S4	4	5	1
S5	6	2	0
S6	7	4	0
S7	9	1	1
S8	2	2	1

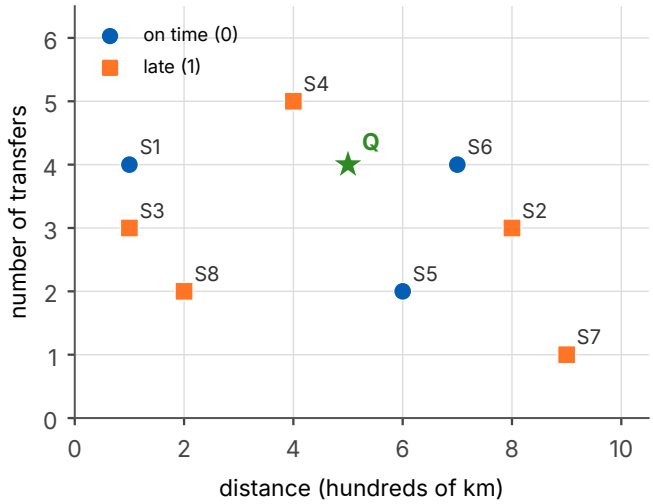


Figure 1 – The eight shipments and the new shipment Q.

(a) Compute the Euclidean distance from Q to each of the eight shipments and rank the neighbours.

(b) Give the prediction of Q for $k = 1$, $k = 3$ and $k = 5$ (majority vote). What do you observe? Why is an odd k convenient?

(c) The standard deviations of the two variables computed on the training set are $\sigma_{\text{dist}} = 3.2$ and $\sigma_{\text{transf}} = 1.3$ (the means cancel in the differences). Standardise and recompute the distance from Q to S2, S4 and S6 (the three nearest shipments after standardisation are among these). Give again the predictions for $k = 1$ and $k = 3$.

(d) Which of the two versions (raw or standardised) would you trust, and how would you choose k in practice?

Exercise A.2 Delivery cost: choosing k by leave-one-out

The cost y (k€) of six deliveries is recorded together with the load x (tonnes). We want to predict the cost of a new delivery by k-NN regression (prediction = mean of the y of the k nearest neighbours in x).

load x	1	2	4	8	10	11
cost y	4	7	8	15	14	19

(a) **Leave-one-out.** For $k = 1$ and $k = 2$, remove each delivery in turn, predict its cost with the other five and give the prediction error. Compute the leave-one-out mean squared error (LOO-MSE) for $k = 1$ and $k = 2$. (For $k = 3$ one finds a LOO-MSE of 14.444.)

(b) Which value of k do you keep? Why does the LOO-MSE not decrease when k becomes larger?

(c) Predict the cost of a delivery of 5.5 tonnes with the chosen k.

Part B – Classification trees

Gini impurity, gain, building a tree

Exercise B.1 Returned products: a tree with the Gini impurity

An online shop wants to predict whether a product will be **returned** (1) from its price (in euros) and the average rating of the customers (1 to 5 stars). Twelve products are observed (Figure 2).

product	price	rating	returned
P1	20	2	1
P2	25	1	1
P3	30	4	0
P4	35	2	1
P5	40	5	0
P6	45	3	0
P7	55	2	1
P8	60	4	0
P9	65	3	1
P10	70	5	0
P11	80	4	0
P12	90	3	1

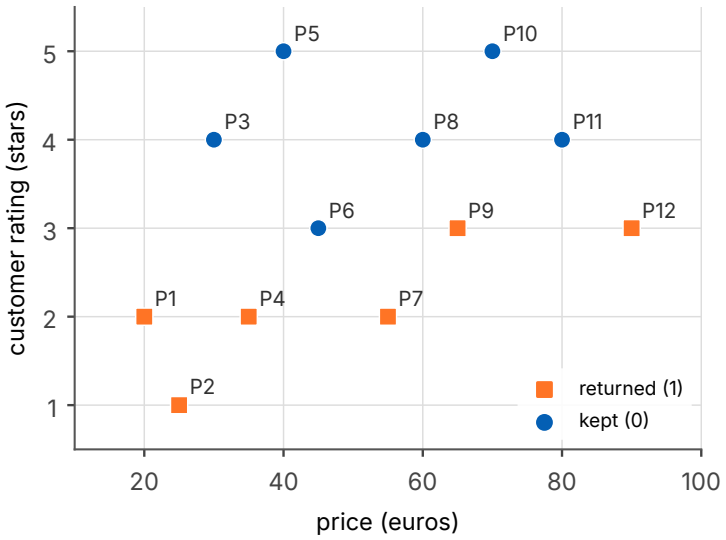


Figure 2 – The twelve products.

(a) Compute the Gini impurity of the root node.

(b) For each of the rules `rating` $\leq$ `3.5`, `rating` $\leq$ `2.5` and `price` $\leq$ `62.5`, give the number of products and of returns in each child, the Gini impurity of each child, the weighted Gini impurity and the gain. Which rule is chosen at the root?

(c) The left child (rating ≤ 3.5) is not pure. Find the best rule to split it (candidates: `rating` $\leq$ `2.5`, `price` $\leq$ `40`, `price` $\leq$ `50`) and draw the tree of depth 2 with the prediction of each leaf. What is its training accuracy?

(d) Predict the outcome of a product at 50 € with a rating of 2 and of a product at 75 € with a rating of 5.

Exercise B.2 Late orders: Gini impurity and categorical variables

A retailer records, for twelve orders, the sales channel (`Web` or `Store`), whether express delivery was chosen, and whether the order arrived late.

channel	delivery	orders	of which late
Web	Express	4	3
Web	Standard	4	1
Store	Express	2	0
Store	Standard	2	0

(a) Compute the Gini impurity of the root node.

(b) Compute the Gini gain of the split by `channel` and of the split by `delivery`. Which variable is chosen at the root?

(c) In the branch chosen as the best question, which question does the tree ask next? Compute its gain and draw the final tree with the prediction of each leaf. What is its training accuracy?

(d) A student adds the variable `order_id` (a different value for each order). What would be its Gini gain? Is it a good variable?

Part C – Regression trees

variance gain, leaves = means, no extrapolation, overfitting

Exercise C.1 Weekly demand: a regression tree on one variable

The weekly demand y (in pallets) of a product is observed during eight weeks (Figure 3).

week x	1	2	3	4	5	6	7	8
demand y	20	24	22	40	44	42	60	58

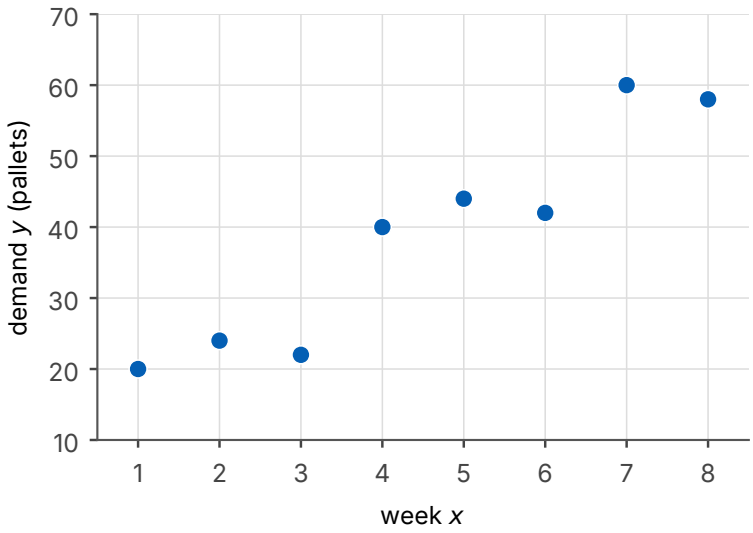


Figure 3 – Weekly demand.

(a) Compute the mean and the variance of y in the root node (variance with denominator n).

(b) Compute the variance gain of the three rules $x \leq 2.5$, $x \leq 3.5$ and $x \leq 6.5$. Which rule is chosen? Give the prediction of the tree of depth 1 and its training MSE.

(c) Refine the tree: in the right child ($x > 3.5$), the best rule is $x \leq 6.5$. Give the three leaves of the tree of depth 2, their predictions and the training MSE.

(d) Predict the demand of weeks 5, 10 and 20 with the tree of depth 2. What do you think of these predictions for weeks 10 and 20? Which kind of model would be more adequate and why?

Exercise C.2 Promotions: two variables and overfitting

The units sold (in hundreds) of a product are recorded for eight weeks with different discounts (%) and advertising budgets (k€). Three further weeks are kept aside as a **test set**.

Training set

week	1	2	3	4	5	6	7	8
discount (%)	0	0	5	5	10	10	15	15
adv. (k€)	1	3	1	3	1	3	1	3
units	20	28	24	30	38	46	52	54

Test set

week	T1	T2	T3
discount (%)	12	3	8
advertising	2	3	1
units	47	25	35

(a) Compute the variance gain of the rules `discount` $\leq$ `7.5`, `discount` $\leq$ `12.5` and `advertising` $\leq$ 2 at the root. Which one is chosen?

(b) Grow the tree to depth 2: in each child, compare the best rule on each variable. Give the four predictions and the training MSE.

(c) Predict the three test weeks with the tree of depth 1 and with the tree of depth 2 and compute the test MSE of each. A tree grown until each leaf contains a single week has a training MSE of 0 and predicts 38, 30 and 38 for the three test weeks: compute its test MSE.

(d) Conclude on the choice of the depth of a regression tree.

Part D – Support vector machines

margin, support vectors, dual form, kernel trick

Exercise D.1 Supplier screening: a linear SVM

Six suppliers are described by two scores x_1 and x_2; the class is $y = +1$ (accepted) or $y = -1$ (rejected) (Figure 4). A linear SVM with a very large C (hard margin) is trained; the dual solution found by the algorithm is given in the table.

point	x_1	x_2	y	α
A	4	2	$+1$	0.125
B	2	4	$+1$	0.125
C	5	4	$+1$	0
D	1	1	-1	0.25
E	0	1	-1	0
F	1	0	-1	0

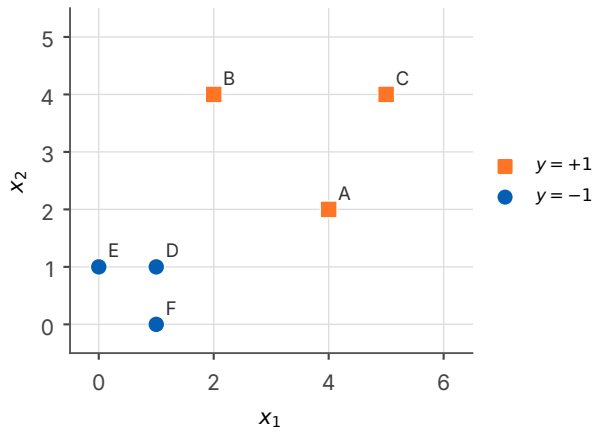


Figure 4 – The six suppliers.

(a) Check that $\sum_i \alpha_i y_i = 0$. Which points are the support vectors?

(b) Compute $\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i$, then the intercept b from a support vector of the positive class ($y_i(\mathbf{w}^\top \mathbf{x}_i + b) = 1$).

(c) Compute the value of the decision function $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$ for the six points. What do you notice for the support vectors? Compute the width of the margin $2/\|\mathbf{w}\|$.

(d) Classify the new suppliers $N_1 = (3, 2)$ and $N_2 = (2, 1)$. Are they inside the margin?

(e) What happens to the separating hyperplane if point C is removed from the training set? And point A? Explain.

(f) What would change if we used a smaller C (soft margin)?

Exercise D.2 Kernels: when a line is not enough

Part 1: feature map. A carrier wants to separate good deliveries ($y = +1$) from problematic ones ($y = -1$) according to a single score x (Figure 5).

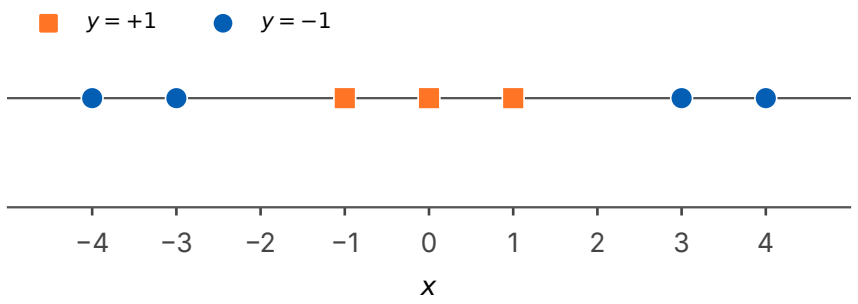


Figure 5 – One-dimensional data.

(a) Explain why no linear classifier on x (a threshold) can separate the two classes.

(b) Consider the map $\phi(x) = (x,\, x^2)$. Give the image of the seven points and show that the classes become linearly separable. Which horizontal line $x^2 = \text{const}$ has the largest margin?

(c) In this feature space, the maximum-margin classifier is $f(x) = w_1 x + w_2 x^2 + b$ with $w_1 = 0$, $w_2 = -\frac{1}{4}$, $b = \frac{5}{4}$. Check that the points $x = \pm 1$ and $x = \pm 3$ satisfy $y\, f(x) = 1$ and compute the margin width $2/\|\mathbf{w}\|$. Classify $x = 2$ and $x = -2.5$.

Part 2: Gaussian (RBF) kernel. Another SVM with the kernel $K(x, x') = \exp\big(-\gamma(x - x')^2\big)$, $\gamma = \frac{1}{2}$, is trained on four points: $x = -2$ and $x = 2$ (class -1), $x = -1$ and $x = 1$ (class $+1$). The fitted decision function is $f(x) = \sum_i \alpha_i y_i K(x_i, x) + b$ with

$$\alpha_1 y_1 = -2.22, \quad \alpha_2 y_2 = +2.22, \quad \alpha_3 y_3 = +2.22, \quad \alpha_4 y_4 = -2.22, \qquad b = -0.15$$

for the points $x_1 = -2$, $x_2 = -1$, $x_3 = 1$, $x_4 = 2$ respectively.

(d) Compute $f(0)$ and $f(3)$ and give the predicted classes. Useful values: $e^{-0.5} = 0.607$, $e^{-2} = 0.135$, $e^{-4.5} = 0.011$, $e^{-8} = 0.0003$.

(e) What is the value of $f(x)$ for a point very far from the training data ($x = 10$)? What does it mean?

(f) What happens when γ becomes very large? And very small? How are C and γ chosen?

Part E – Random forests

bootstrap, out-of-bag error, averaging, variance

Exercise E.1 Out-of-bag estimate of a small forest

A random forest of 4 trees is trained on 8 orders O1–O8 (label $y = 1$: cancelled). The bootstrap samples drawn for each tree (observation indices) and the predictions of each tree for the eight orders are given below.

	bootstrap sample (8 draws)
Tree 1	O1, O1, O3, O4, O4, O6, O7, O8
Tree 2	O2, O2, O3, O5, O5, O6, O8, O8
Tree 3	O1, O2, O4, O5, O7, O7, O7, O8
Tree 4	O3, O3, O4, O5, O6, O6, O7, O1

	O1	O2	O3	O4	O5	O6	O7	O8
true label y	1	0	1	1	0	0	1	0
Tree 1	1	0	1	1	0	0	1	0
Tree 2	1	0	1	0	0	0	1	0
Tree 3	1	0	1	1	0	1	1	0
Tree 4	1	0	1	1	1	0	1	1

(a) For each tree, list the out-of-bag observations.

(b) Compute the OOB prediction of each order (majority vote of the trees for which it is out-of-bag) and the OOB error rate of the forest.

(c) Compute the training error of each tree on the whole training set. Compare with the OOB error and comment.

(d) The probability that an observation is never drawn in a bootstrap sample of size 8 is $(1 - 1/8)^8$. Compute it. If we had only one tree, how many orders would on average be out-of-bag?

Exercise E.2 Averaging trees: errors and variance

A random forest of 5 regression trees predicts the delivery time (days) of three shipments. The predictions of each tree and the observed values are:

	observed	Tree 1	Tree 2	Tree 3	Tree 4	Tree 5
Shipment 1	50	44	58	47	55	46
Shipment 2	30	36	25	33	27	34
Shipment 3	42	40	47	38	45	36

(a) Compute the forest prediction for each shipment, and its squared error.

(b) Compute the mean squared error of each of the five trees on the three shipments, and their average. Compare with the MSE of the forest.

(c) For shipment 1, verify that: (average squared error of the trees) = (squared error of the forest) + (variance of the five predictions). What does this tell about averaging?

(d) The predictions of the trees have a variance $\sigma^2 = 36$ and a pairwise correlation $\rho = 0.25$. Compute the variance of the average of T trees for $T = 9$, for $T \to \infty$, and the smallest T for which it is within 10 % of its limit.

(e) The data have 16 explanatory variables. How many variables are drawn at each split with the usual $\sqrt{d}$ rule? Which role does this parameter play?

Part F – AdaBoost

weighted error, vote weight α, weight update, weighted vote

Exercise F.1 AdaBoost with three given stumps

Ten orders have labels $y_i \in \{-1, +1\}$ ($+1$: late). Three weak learners (stumps) are produced one after the other by AdaBoost; their predictions are given. Round 1 starts with equal weights.

	O1	O2	O3	O4	O5	O6	O7	O8	O9	O10
label y_i	$+1$	$+1$	-1	-1	$+1$	-1	$+1$	-1	$+1$	-1
$h_1(\mathbf{x}_i)$	$+1$	-1	-1	-1	$+1$	-1	-1	-1	$+1$	-1
$h_2(\mathbf{x}_i)$	$+1$	-1	-1	$+1$	$+1$	-1	$+1$	-1	$+1$	-1
$h_3(\mathbf{x}_i)$	$+1$	$+1$	-1	$+1$	$+1$	-1	$+1$	-1	-1	$+1$

(a) Round 1: which orders does h_1 misclassify? Compute ε_1 and α_1.

(b) Update the weights and normalise them. Which weight do the misclassified orders now carry in total?

(c) Round 2: compute ε_2 and α_2. Why is ε_2 larger than ε_1 although h_2 makes the same number of errors as h_1?

(d) After the second update, the weights are $\frac{1}{22}$ for O1, O3, O5, O6, O8, O9, O10, 0.4 for O2, 0.1 for O4 and $\frac{2}{11}$ for O7. Compute ε_3 and α_3.

(e) Compute the final score $\sum_{t=1}^{3} \alpha_t h_t(\mathbf{x}_i)$ of each order, the prediction $H(\mathbf{x}_i) = \operatorname{sign}(\cdot)$ and the training accuracy. Comment.

Exercise F.2 AdaBoost with stumps on one variable

Eight shipments are described by one score $x \in \{1, \dots, 8\}$; the label is $y = +1$ (late) or -1. The weak learners are the stumps $h_{c,s}(x) = s$ if $x \le c$ and $-s$ otherwise, with $c \in \{1.5, 2.5, \dots, 7.5\}$ and $s \in \{+1, -1\}$.

x	1	2	3	4	5	6	7	8
y	$+1$	$+1$	$+1$	-1	-1	$+1$	-1	-1

(a) **Round 1** (weights $\frac{1}{8}$). Compute the weighted error of the stumps $h_{c,+1}$ for $c = 2.5, 3.5, 4.5$ and 6.5. The stump with the smallest error among all 14 stumps is the best one: which is it? Compute its α_1 and the new weights.

(b) **Round 2.** The best stump of the second round is $h_{6.5,+1}$. Compute its weighted error ε_2, α_2 and the new weights.

(c) **Round 3.** The best stump of the third round is $h_{5.5,-1}$. Compute ε_3 and α_3.

(d) Compute the final score at each $x \in \{1, \dots, 8\}$ and the training accuracy of $H = \operatorname{sign}\left(\sum_t \alpha_t h_t\right)$. What class is predicted for $x = 9$?

(e) Each stump is monotone in x, but the label pattern is not: how does AdaBoost manage to fit it?

Part G – Gradient boosting

residuals, learning rate, number of trees, early stopping

Exercise G.1 Gradient boosting by hand

The weekly demand y (in hundreds of units) of a product is observed during five weeks $x = 1, \dots, 5$: $y = 12, 14, 26, 28, 40$. We use gradient boosting with the squared loss, regression stumps and a learning rate $\eta = 0.5$.

(a) Give F_0 and its training MSE; compute the residuals.

(b) **Round 1.** Among the rules $x \leq 2.5$, $x \leq 3.5$ and $x \leq 4.5$, which gives the largest variance gain on the residuals? Give the values of the two leaves of the stump h_1 (mean residual of each side), the new model $F_1 = F_0 + \eta h_1$ and its training MSE.

(c) **Round 2.** New residuals; among the same three rules, which is the best? Give h_2, F_2 and its training MSE.

(d) Predict the demand of week 6 with F_2. What is the limitation?

(e) What would change with $\eta = 1$ and with $\eta = 0.1$? Why is a small η combined with many trees often preferred?

Exercise G.2 An ensemble of stumps and tuning the number of trees

Part 1. A gradient boosting model with $F_0 = 40$ and $\eta = 0.1$ has been trained for a demand series with three stumps: $h_1(x) = -8$ if $x \leq 5$ and $+10$ otherwise; $h_2(x) = -5$ if $x \leq 3$ and $+3$ otherwise; $h_3(x) = -2$ if $x \leq 7$ and $+6$ otherwise. Three new weeks are observed: week 2 (demand 36), week 6 (demand 45) and week 9 (demand 44).

(a) Compute $F_3(x) = F_0 + \eta\,(h_1 + h_2 + h_3)$ for the three weeks, the residuals and the MSE. What would the fourth tree try to fit?

Part 2. The same kind of model (trees of depth 2) is trained on 60 weeks of demand and evaluated on 300 other weeks (validation set), for three learning rates. The RMSE is (the training RMSE is given for $\eta = 0.5$ only):

number of trees T	1	5	10	25	50	100	200
$\eta = 0.5$: training	6.50	3.97	3.34	1.59	0.66	0.15	0.01
$\eta = 0.5$: validation	7.65	5.86	5.71	6.34	6.62	6.85	6.91
$\eta = 0.1$: validation	9.81	8.03	6.90	5.88	5.54	5.81	6.31
$\eta = 0.02$: validation	10.32	9.83	9.30	8.10	6.96	6.07	5.57

(b) For each learning rate, give among the listed values the number of trees with the smallest validation RMSE, and this RMSE.

(c) Describe what happens with $\eta = 0.5$: compare the training and the validation RMSE when T grows. What is this phenomenon?

(d) What does the row $\eta = 0.02$ suggest? Which configuration would you choose and how would you refine the choice of T (early stopping)?

(e) The validation RMSE of the chosen configuration is 5.5. Can we announce that the model will have this error on future data? What should be done?