Chair of Decision Sciences and Systems
TUM School of Computation, Information and Technology
Technical University of Munich

Business Analytics & Machine Learning
Homework sheet 7: Model Evaluation and Selection – Solution

Prof. Dr. Martin Bichler, Prof. Dr. Jalal Etesami
Julius Durmann, Markus Ewert, Johannes Knörr, Yutong Chao

Exercise H7.1 *Metrics*

You want to bet on soccer matches and try to predict match results. In order to improve your forecasts, you decide to use your knowledge on data mining and construct a decision tree. The table below compares the real outcome and your predicted outcome of 15 matches.

Calculate the accuracy, the true positive rate, the false positive rate and the true negative rate for your decision tree based on your predictions.

True Class	Predicted Class
1	1
0	1
1	1
1	1
1	0
0	0
0	1
1	0
0	0
0	0
0	0
1	1
1	1
1	0
0	0

Solution

Number of occurrences:

- True positives (TP): 5
- False negatives (FN): 3
- True negatives (TN): 5
- False positives (FP): 2

Measures:

True Class	Predicted Class	Event
1	1	TP
0	1	FP
1	1	TP
1	1	TP
1	0	FN
0	0	TN
0	1	FP
1	0	FN
0	0	TN
0	0	TN
0	0	TN
1	1	TP
1	1	TP
1	0	FN
0	0	TN

- True positive rate (recall) $= \frac{\text{TP}}{\text{TP+FN}} = \frac{5}{5+3} = \frac{5}{8} = 0.625$
- False positive rate (false alarm rate) $= \frac{\text{FP}}{\text{FP+TN}} = \frac{2}{2+5} = \frac{2}{7} \approx 0.286$
- True negative rate (specificity) $= \frac{\text{TN}}{\text{TN+FP}} = \frac{5}{5+2} = \frac{5}{7} \approx 0.714$
- Accuracy $= \frac{\text{TN+TP}}{\text{TP+FN+TN+FP}} = \frac{5+5}{5+3+5+2} = \frac{10}{15} \approx 0.667$

Exercise H7.2 *Gain curve, lift curve, ROC curve*

Use the given results of a classifier that outputs the probabilities of instances being positive to construct:

- a gain curve (10 % steps),
- a lift curve,
- an ROC curve.

Furthermore, mark a cutoff value of 0.87 in the plots.

Index	Probability	Class
1	0.991	+
2	0.977	+
3	0.973	+
4	0.945	+
5	0.918	+
6	0.915	-
7	0.906	+
8	0.889	-
9	0.873	+
10	0.871	+

11	0.869	-
12	0.866	-
13	0.862	+
14	0.852	-
15	0.837	+
16	0.831	-
17	0.829	-
18	0.811	-
19	0.787	-
20	0.779	-

Remember: A cutoff value of 0.87 means that we will classify an instance as positive if its probability is above 87 % and negative, otherwise.

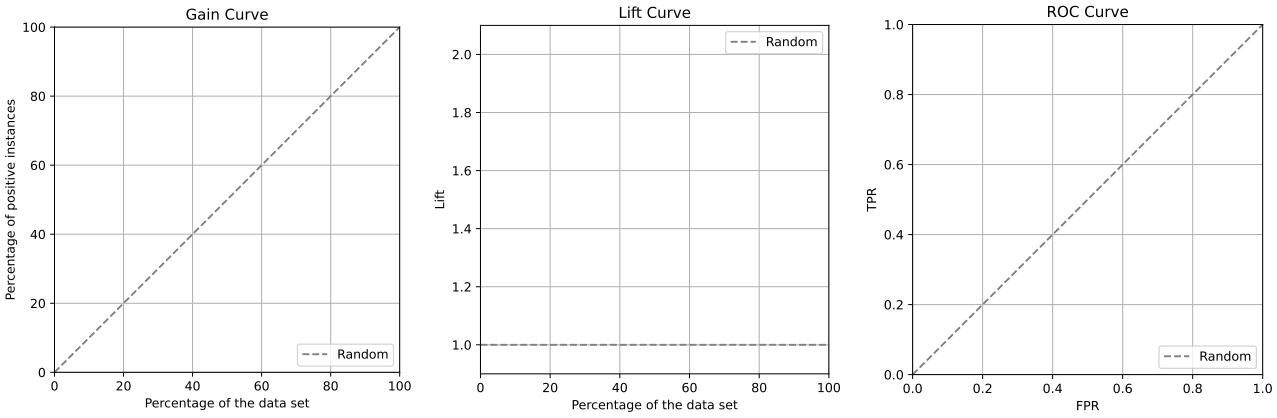

Figure 1 Exercise 8.2 template from left to right: Gain curve, lift curve, and ROC curve.

Solution

It is given a probabilistic model where the output is the estimated probability that a given instance belongs to the positive class (2nd column). The actual classes are shown in the 3rd column.

The evaluation metrics in use operate in the supervised binary classification setting, so one must apply a threshold in order to obtain a classifier from the probabilistic model given. That is, at a cutoff value of 0.87 all instances with a probability higher than 0.87 would be classified as positive instances. Only the probabilities and the resulting ordering are a product of our model, but the threshold is something that we apply after the fact. It is (at this stage of the modeling process) independent of the model and may be set to different values depending on the objective.

Gain-curve: The gain-curve answers the question of how many instances should be classified as positive to achieve a certain proportion of true positive instances. In order to construct the gain-curve, we sort all instances by their score (in this case: the predicted probability of their class being positive). Starting from the highest score, different portions of the instances are classified as positive. As a result, more and more instances may be classified wrong. Starting from the two highest ranked instances (corresponding to $x = 10\,\%$ of the data), and assuming them to be classified as positive, we ask what the true positive rate TPR = TP/(TP+FN) is: $y = 2/(10 + 0) = 20\,\%$. We iterate in this manner over the whole sorted data set.

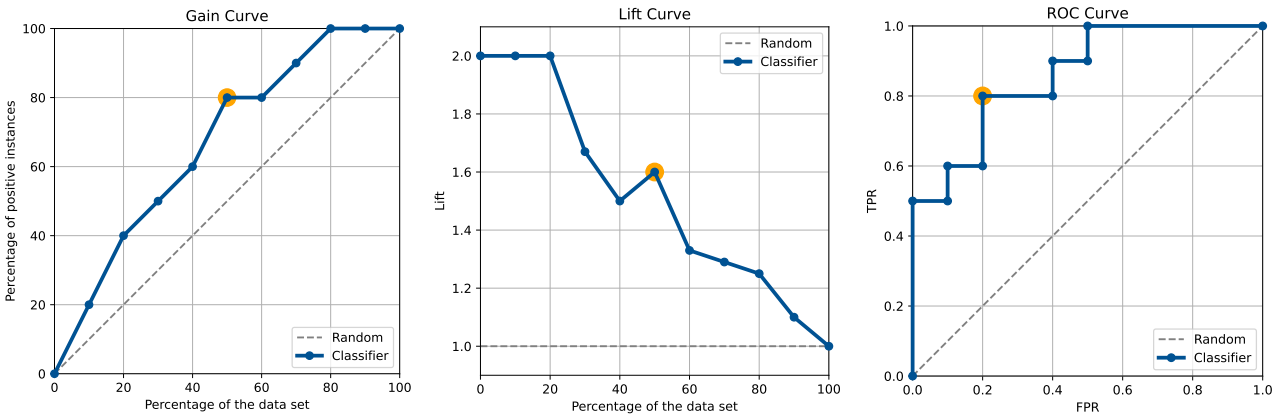

Figure 2 Exercise 8.2 solution from left to right: Gain curve, lift curve, and ROC curve.

Lift-curve: To get the corresponding lift-curve, we divide the y-value by the x-value for every point of the gain-curve.

ROC-curve: The ROC-curve is similar to the gain-curve, but the x-axis shows the false positive rate (FPR) instead of the percentage of instances classified as positive. It visualizes the relationship between the true positive rate (hit rate / recall) and the false positive rate at different thresholds (TPR = TP/(TP+FN) on the vertical axis and FPR = FP/(FP+TN) on the horizontal axis. Notice that the denominators are fixed given the dataset since TP+FN = # actual positives and FP+TN = # actual negatives).

Let us go through the procedure for different cutoff values:

1. If we set the cutoff value to 0.991, all predicted probabilities would fall below the threshold and be classified as negatives. The corresponding operating point on the ROC curve is: TPR = 0 /(0+10) = 0 and FPR = 0/(0+10) = 0.

2. If we move the threshold further down, e.g., to 0.98 (just so we “capture” the first instance), we obtain: TPR = 1/(1+9) = 0.1 and FPR = 0/(0+10) = 0. Hence, we move one step upwards from (0, 0) to (0, 0.1) on the curve.

3. Observe that the first five instances are all actual positives, so the true positive rate will keep increasing and the false positive rate will remain at zero as we keep lowering the threshold until we fall below 0.915 (instance 6 with output 0.915 is an actual negative). Thus, the result is a vertical line from (0, 0) up to (0, 0.5).

4. At a cutoff value between 0.906 and 0.915 the model would classify the first six instances as positives, out of which only five are actual positive. So we get TPR = 5/(5+5) = 0.5 and FPR = 1/(1+9) = 0.1 which is one step to the right from (0, 0.5) to (0.1, 0.5).

5. We repeatedly apply this procedure until the cutoff value is lower than the lowest predicted probability, which results in positive predictions for all 20 instances and translates to TPR = 10/(10+0) = 1 and FPR = 10/(10+0) = 1.

The results for this use case are depicted in Figure 2.

Exercise H7.3 *Cross validation*

Design a stratified 5-fold cross-validation for the below-mentioned table. Each data point consists of four attributes (A1 - A4) and a class label (Class).

Nr.	A1	A2	A3	A4	Class
1	4.9	3.1	1.5	0.1	1
2	5.0	3.2	1.2	0.2	0
3	5.5	3.5	1.3	0.2	0
4	4.9	3.1	1.5	0.1	1
5	4.4	3.0	1.3	0.2	1
6	5.1	3.4	1.5	0.2	0
7	5.0	3.5	1.3	0.3	1
8	4.5	2.3	1.3	0.3	1
9	4.4	3.2	1.3	0.2	0
10	5.0	3.5	1.6	0.6	0
11	5.1	3.8	1.9	0.4	0
12	4.8	3.0	1.4	0.3	1
13	5.1	3.8	1.6	0.2	0
14	4.6	3.2	1.4	0.2	1
15	5.3	3.7	1.5	0.2	0
16	5.0	3.3	1.4	0.2	0
17	7.0	3.2	4.7	1.4	1
18	6.4	3.2	4.5	1.5	0
19	6.9	3.1	4.9	1.5	1
20	5.5	2.3	4.0	1.3	1

Solution

The proportion of the two types of class labels is [10, 10] = 1:1 and the size of the data set is 20. In 5-fold cross-validation, the original sample is randomly partitioned into five equal size subsamples. Every subsample consists of 20/5 = 4 instances. In stratified k-fold cross-validation, the folds are selected so that each fold contains roughly the same proportions of the two types of class labels. Regarding a sample with four instances per subsample and a proportion of 1:1 of the two types of class labels every subsample has to contain two instances of the class 0 and two instances of the class 1.

Example partition (note that this is only one of many possible solutions):

- P1 = $\{1, 2, 3, 4\}$
- P2 = $\{5, 6, 7, 9\}$
- P3 = $\{8, 10, 11, 12\}$
- P4 = $\{13, 14, 15, 17\}$
- P5 = $\{16, 18, 19, 20\}$

Cross-validation:
Note: The attributes from above-mentioned table are not necessary for the solution.

Step	Training	Validation
1	P2 $\cup$ P3 $\cup$ P4 $\cup$ P5	P1
2	P1 $\cup$ P3 $\cup$ P4 $\cup$ P5	P2
3	P1 $\cup$ P2 $\cup$ P4 $\cup$ P5	P3
4	P1 $\cup$ P2 $\cup$ P3 $\cup$ P5	P4
5	P1 $\cup$ P2 $\cup$ P3 $\cup$ P4	P5