Problems for submission 4 (10 points)

Due: 27.11.2024 at 20:00

- Please provide your solutions in a text file. The graphs in this file just help you to visualize the problem.
- Name of file: Surname1_Surname2_Surname3_Problems4.txt
- Add full names on top of the text file
- Please show all your work

Please submit your text file to the course website on ADAM under Student Submissions of Problem Solutions -> Problems 4

Introduction: ROC curves

Receiver Operating Characteristic (ROC) curves are widely employed in virtual screening to assess the performance of predictive models in distinguishing between active and inactive compounds. These curves graphically depict the trade-off between sensitivity and specificity across various classification thresholds. In the context of virtual screening, ROC curves provide a valuable tool for evaluating the ability of computational methods to prioritize potential drug candidates effectively. The area under the ROC curve (AUC) serves as a quantitative measure, with higher area under the curve (AUC) values indicative of superior model performance in discriminating true positives from false positives.

The data used for this exercise is artificially generated and functions solely as an illustrative toy example. In the forthcoming "Virtual Screening" exercise, you will generate authentic data and conduct a thorough analysis. This dataset comprises 30 compounds, each possessing the following properties:

- Compound: Unique compound number
- Class: The compounds are classified in 0 or 1 which could correspond to non-binding (0, False) and binding (1, True) ligands. The class represents the ground truth.
- Value: These constructed values represent calculated numbers. For example a docking score.

In virtual screening, our goal is therefore to deduce the class (binding or non-binding) from a predicted value, specifically the docking score.

compound	class	value	compound	class	value
1	0	1	16	1	5
2	0	2	17	1	6
3	0	2	18	1	6
4	0	3	19	1	6
5	0	3	20	1	7
6	0	3	21	1	7
7	0	4	22	1	7
8	0	4	23	1	7
9	0	4	24	1	7
10	0	4	25	1	7
11	0	5	26	1	8
12	0	5	27	1	8
13	0	6	28	1	9
14	1	4	29	1	9
15	1	5	30	1	10

Table 1: Artifically generated data. The table has been sorted according to class and value.

1. [3 pts] Utilize the data provided in Table 1 to create a histogram illustrating the distribution of values within the two classes. Employ a bin length of 1 for optimal representation. Additionally, consider employing distinct colors for each class (0 and 1) within the same histogram, or alternatively, generate separate histograms for Class 0 on the left and Class 1 on the right for clearer visualization. In the text file please provide a table with the following information: class, bin number, count. Please sort the table according to class and bin number.

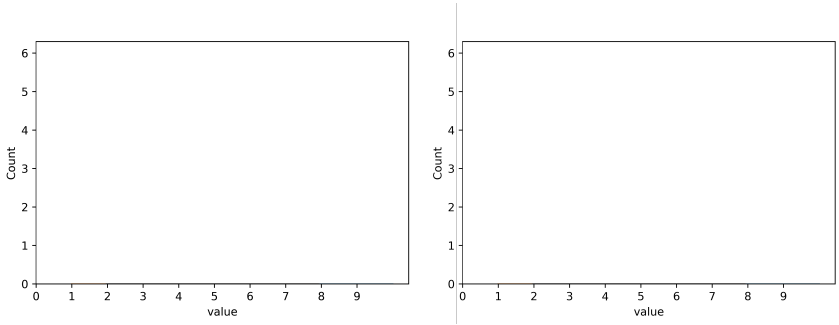

Figure 1: Histogram of class distribution. Left: Class 0, Right: Class 1

A confusion matrix summarizes the performance of a binary classifier by showing true and false positives, as well as true and false negatives:

Figure 2: Confusion matrix. Note: In Literature the Prediction and ground truth is sometimes flipped.

2. [3 pts] Employ the provided templates in figure 3 to construct confusion matrices for the data across various thresholds. To populate a confusion matrix, count the occurrences of true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) in figure 1 at the designated thresholds. The threshold depicts the value (for example docking score) which separates the two predicted classes. In the solution text file please provide the 4 tables with the corresponding threshold as table title. Use spaces to separate the cells.

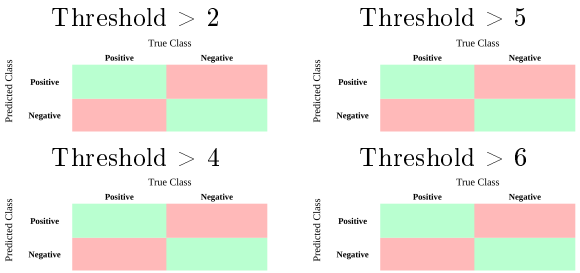

Figure 3: Calculated confusion matrices for four different thresholds.

3. [3 pts] ROC curves are constructed by plotting the true positive rate (TPR) against the false positive rate (FPR). Calculate the TPR and FPR for the three thresholds and plot the points in the plot below. A ROC curve starts at (0,0) and ends at (1,1). Try to interpolate the curve using your 4 calculated points. In the text file please provide the four points as (FPR, TPR).

$TPR = \frac{TP}{TP+FN}$ $FPR = \frac{FP}{FP+TN}$

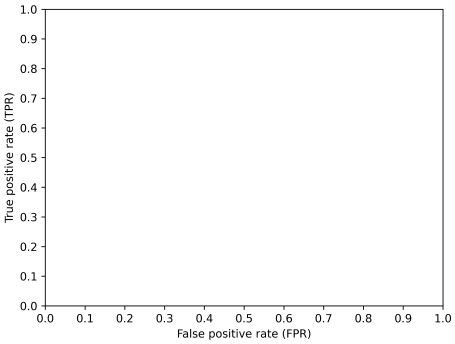

Figure 4: Formulas to calculate true positive and false positive rates and a template to draw the ROC curve.

4. [1 pts] Based on your previous interpolation what is your estimation of the area under the curve (AUC).
   (a) AUC $= 1$
   (b) AUC > 0.5
   (c) AUC < 0.5