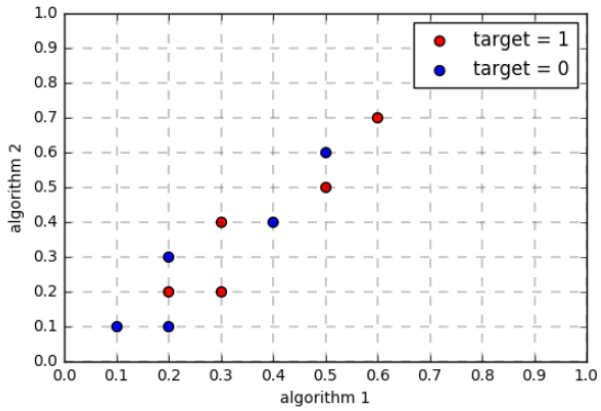
Метрики качества бинарного классификатора

1. Пусть даны выборка X, состоящая из 8 объектов, и классификатор $b(x)$, предсказывающий оценку принадлежности объекта к положительному классу. Предсказания $b(x)$ и реальные метки объектов приведены ниже:

$$\begin{aligned} b(x_1) &= 0.1, \quad y_1 = +1, \\ b(x_2) &= 0.8, \quad y_2 = +1, \\ b(x_3) &= 0.2, \quad y_3 = -1, \\ b(x_4) &= 0.25, \quad y_4 = -1, \\ b(x_5) &= 0.9, \quad y_5 = +1. \\ b(x_6) &= 0.3, \quad y_6 = +1, \\ b(x_7) &= 0.6, \quad y_7 = -1, \\ b(x_8) &= 0.95, \quad y_8 = +1. \end{aligned}$$

   Постройте ROC-кривую и вычислите $AUC - ROC$ для классификаторов $a(x, t)$, порожденных $b(x)$, на выборке X.

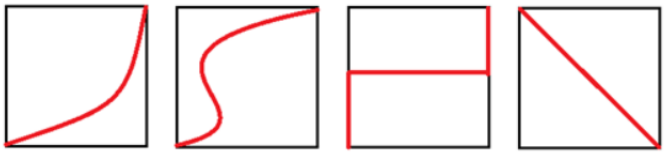
2. Постройте ROC-кривые для двух алгоритмов, предсказания которых изображены на рисунке. Посчитайте $AUC - ROC$ каждого из алгоритмов.



3. Ответьте на следующие вопросы:

   (a) В тестовой выборке 10 объектов, известно, что $AUC - ROC < 1$. Какое максимальное значение может быть у $AUC - ROC$?

   (b) Какие из этих кривых могут быть ROC-кривыми?



(c) Объекты нулевого класса получили оценки $0.1, 0.4, 0.5$, а первого – $0.2, 0.8$. Чему равен $AUC - ROC$?

(d) Как изменится значение $AUC - ROC$ после округления ответов на тестовой выборке до 2 знака после запятой ($0.7235 \approx 0.72$) ?

4. Пусть дана некоторая выборка X и классификатор $b(x)$, возвращающий в качестве оценки принадлежности объекта x к положительному классу 0 или 1 (а не вероятности).

   1) Постройте ROC-кривую для классификатора $b(x)$ на выборке X.

   2) Покажите, что AUC-ROC классификатора $b(x)$ может быть выражена через долю правильных ответов и полноту классификатора $a(x;t)$, получающегося при выборе некоторого порога $t \in (0;1)$ ($a(x) = [b(x) > 1]$). Помимо указанных величин в формулу могут входить N, N_+, N_-, число объектов, число положительных и отрицательных объектов в выборке X соответсвенно.

5. Алгоритм бинарной классификации выдает оценки вероятности принадлежности к положительному классу $b_i = \hat{\mathcal{P}}(y_i = +|x_i)$. Всего есть $N = 10000$ наблюдений. Если ранжировать их по возрастанию b_i, то окажется, что наблюдения с $y_i = 1$ и наблюдения с $y_i = 0$ образуют чередующиеся блоки различного размера:

$$\underbrace{- - - - - - - -}_{3N/8}\underbrace{+ + + + + + + +}_{3N/8}\underbrace{- - - - - - - -}_{N/8}\underbrace{+ + + + + + + +}_{N/8}$$

   Постройте ROC и PR кривые, определите площади под кривыми.