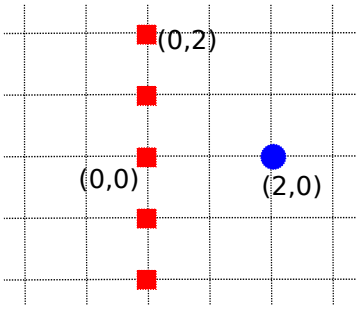
Теоретическое задание 1

решение принимается до 20.03.16

№ 1. Нарисуйте границу классов для алгоритма kNN в случае, когда размерность пространства признаков $d = 2$, при этом один класс представлен одной точкой, а другой множеством точек, лежащих на одной прямой (см. рис), используется метрика Минковского $p = 2$. В чем особенность этой границы?



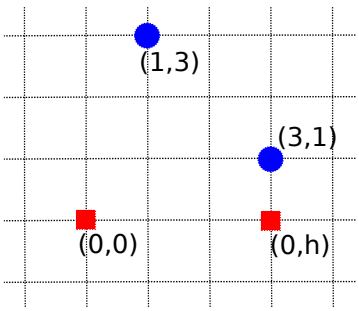
№ 2. В листовой вершине m находится — 200 объектов первого класса и 300 объектов второго класса (всего два класса). Нужно сделать выбор между двумя правилами, одно из которых генерирует поддеревья с числом объектов $(200, 200)$ и $(0, 100)$, а другое — с числом объектов $(100, 0)$ и $(100, 300)$. Вычислите значение критериев информативности: $Q_E(x)$ (ошибка классификации), $Q_G(x)$ (индекс джини), $Q_H(x)$ (энтропийный) и заполните таблицу 1. Какое из правил будет выбрано при каждом из критериев информативности?

	$Q_E(x)$	$Q_G(x)$	$Q_H(x)$
Правило 1			
Правило 2			

Таблица 1: значения критерия информативности

№ 3. На рисунке 2 изображена выборка из четырёх объектов и двух классов. Величина $h \leq 0$ — параметр. Будем строить разделяющую гиперплоскость методом опорных векторов.

a) При каких значениях параметра h выборка будет разделима? b) Опишите при каждом значении h множество опорных объектов и множество периферийных объектов. c) Как будет меняться наклон оптимальной разделяющей гиперплоскости в предыдущей задаче при изменении параметра h? d) Как зависит от h ширина разделяющей полосы, соответствующей оптимальной разделяющей гиперплоскости?



№ 4. Являются ли ядрами: a) $K(x,y) = exp(3 < x, y >) + < y + x, 2y + x >$;
b) $K(x,y) = \text{ch}(< x, y >) + 3\,\text{sh}(< x, y >)$?

№ 5. Для каждой пары (алгоритм, гиперпараметр) скажите, повышается лисклонность метода к переобучению при увеличении параметра. Кратко поясните каждый ответ.

1. (K-NN, число соседей K)
2. (решающее дерево, минимальное количество элементов в листе)
3. (линейный SVM,C)

№ 6. Рассмотрим вершину m и объекты R_m , попавшие в нее при построении дерева. Сопоставим в соответствие вершине m алгоритм $a(x)$, который выбирает класс случайно, причем класс k выбирается с вероятностью $p_{mk} = \frac{1}{|R_m|} \sum_{x_i \in R_m} [y_i = k]$. Покажите, что матожидание частоты ошибок этого алгоритма на объектах из R_m равно индексу Джини.