Домашнее задание №1

1. Пусть объекты (0,1) и (0, -1) принадлежат одному классу, а объект (2, 0) — другому. Получить аналитическую формулу разделяющей поверхности для метода одного ближайшего соседа с ℓ_2 метрикой.

2. Решить предыдущую задачу для ℓ_1 метрики.

3. Привести пример обучающего и тестового множества минимальной мощности, для которых при $k > 1$:

   (a) метод 1-ближайшего соседа даёт точность на тесте больше, чем метод k-ближайших соседей,

   (b) метод k-ближайших соседей даёт точность на тесте больше, чем метод 1-ближайшего соседа.

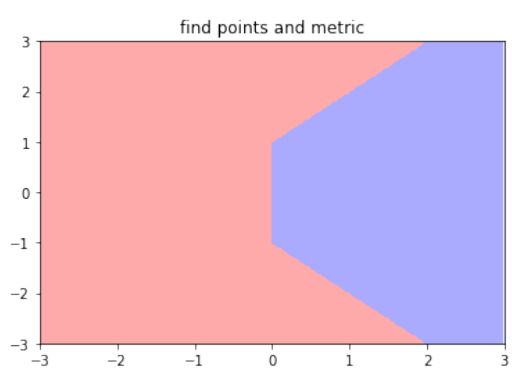
4. Привести пример обучающего и тестового множества минимальной мощности, для которых при $k > 1$:

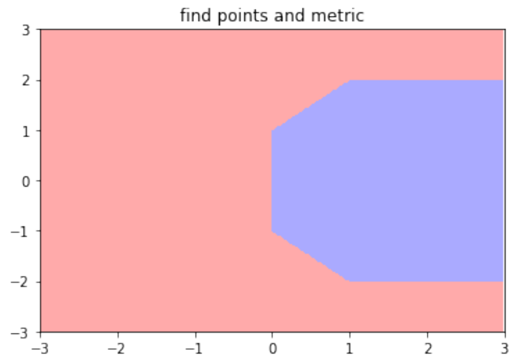
   (a) обычный метод k-ближайших соседей даёт точность на тесте больше, чем метод k-ближайших взвешенных соседей (веса, например, обратно пропорциональны расстояниям),

   (b) метод k-ближайших взвешенных соседей даёт точность на тесте больше, чем обычный метод k-ближайших соседей.

5. Для метода классификации одного ближайшего соседа получились следующие разделяющие поверхности. Примести пример обучающего множества и метрики, на которых метод мог бы выучить такое разделяющее правило.

a)



b) 

6. Опишите преимущества и недостатки k-fold валидации и LOO валидации. Приведите примеры, когда предпочтительнее использовать LOO вместо 5-fold валидации, и наоборот.

7. Для объектов $x_1, ..., x_n$ с правильными ответами $y_1, ..., y_n$ из $\mathbb{R}$ постройте константную модель $a(x) = c$ для функции потерь:

   a) $MSE\ (mean\ squared\ error) = \frac{1}{N}\sum\limits_{i=1}^{n}(y_i - c)^2$;

   b) $MAE\ (mean\ absolute\ error) = \frac{1}{N}\sum\limits_{i=1}^{n}|y_i - c|$.

8. Для объектов $x_1, ..., x_n$ с правильными ответами $y_1, ..., y_n$ из $\{0, 1\}$ постройте наилучшую константную классификационную модель $a(x) = c$ для функции потерь бинарной перекрёстной энтропии.

9. Для объектов $x_1, ..., x_n$ из $\mathbb{R}$ с правильными ответами $y_1, ..., y_n$ из $\mathbb{R}$ постройте наилучшую константную модель $a(x) = kx + b$ для функции потерь

$$MSE\ (mean\ squared\ error) = \frac{1}{N}\sum_{i=1}^{n}(y_i - kx_i - b)^2.$$

10. Пусть дана обучающая выборка $(x_1, y_1), ..., (x_n, y_n)$, где пара (x_i, y_i) — объект и правильный ответ, причем $x_i \in \mathbb{R}^2$, а $y_i \in \{0, 1\}$. Известно, что распределение для обоих классов — гауссовское с параметрами: $\mu_0 = (a, b)^T$, $\mu_1 = (-a, -b)^T$, $\Sigma_0 = \Sigma_1 = diag(\sigma_1, \sigma_2)$. Выпишите байесовский алгоритм классификации и уравнение разделяющей поверхности.

11. Доказать, что наивный байесовкий классификатор в случае бинарных признаков $f_i \in \{0, 1\}$ является линейным разделителем:

$$a(x) = a(f_1, ..., f_n) = [a_0 + a_1 f_1 + ... + f_n x_n > 0].$$

Выведите формулы для коэффициентов a_i.