Задачи для подготовки к контрольной работе

1. **Линейная регрессия.** Для задачи линейной регрессии с L_2-регуляризацией с одним признаком запишите формулу для определения веса при этом признаке с помощью оптимизации среднеквадратичной ошибки. Считайте, что свободный член $w_0 = 0$ (его не нужно настраивать).

2. **Логистическая регрессия.** Петя оценил логистическую регрессию по четырём наблюдениям и одному признаку с константой, получил $b_i = \hat{\mathcal{P}}(y_i = 1|x_i)$, но потерял последнее наблюдение:

y_i	b_i
1	0.7
-1	0.2
-1	0.3
?	?

   (a) Выпишите функцию потерь для задачи логистической регрессии.

   (b) Выпишите условие первого порядка по коэффициенту перед константой.

   (c) Помогите Пете восстановить пропущенные значения!

3. **Метрики качества классификации.**

   (a) У алгоритма $b(x)$ AUC-ROC равен 0.1. Предложите способ построить алгоритм, имеющий лучшее качество.

   (b) Вася построил алгоритм b(x), AUC-ROC которого 0.63. Петя построил алгоритм, $c(x) = b(x)/3$. Чему равен AUC-ROC для него?

4. **AUC-ROC.** Постройте ROC-кривые и посчитайте $AUC - ROC$ для алгоритма $a(x)$, здесь p — вектор предсказанных вероятностей принадлежности классу $+1$, y — истинный класс.

$$p = \Big[0.9,\ \ 0.1,\ \ 0.75,\ \ 0.56,\ \ 0.2,\ \ 0.37,\ \ 0.25\Big],$$

$$y = \Big[+1,\ \ -1,\ \ -1,\ \ +1,\ \ +1,\ \ -1,\ \ -1\Big]$$

5. **Ядра.** Полиномиальное ядро второй степени задается формулой:

$$K(x, y) = (1 + (x, y))^2.$$

   Имеются три наблюдения:

	x_1	x_2
A	1	-2
B	2	1
C	3	0

Найдите расстояние AB и косинус угла ABC в расширенном пространстве.

6. **Решающие деревья.** Имеются объекты трех типов, составляющих 60%, 20% и 20% обучающей выборки соответственно. Рассмотрим разбиение выборки на две части при построении дерева решений: левая часть состоит только из объектов первого типа, вторая из объектов второго и третьего типа. Чему равен прирост информации при таком разбиении, если мера неоднородности: 1) критерий Джини? 2) энтропия?

7. **Разложение ошибки.** Истинная зависимость имеет вид $y_i = x_i^2 + u_i$, где y_i прогнозируемая переменная, x_i- признак и u_i - случайная составляющая. Величины x_i независимы и равновероятно принимают значения 1 и 2. Величины u_i независимы и равновероятно принимают значения -1 и 1. Разложите ожидание квадрата ошибки прогноза на шум, смещение и разброс, если:

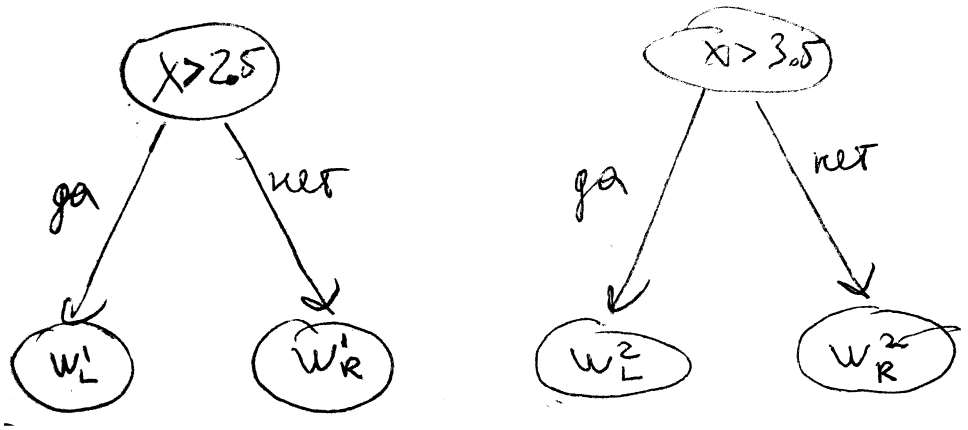
   a) По обучающей выборке строится регрессия на константу.

   b) По обучающей выборке строится регрессионое дерево.

8. **Ансамбли.** Машин-лёрнер Василий лично раздобыл выборку из четырёх наблюдений.

x_i	y_i
1	6
2	6
3	12
4	18

Два готовых дерева для леса Василий подглядел у соседа:



a) Василий решил использовать бэггинг. Первому дереву достались наблюдения номер 1, 1, 2 и 3. А второму дереву - 2, 3, 4 и 4. Прогнозы в каждом листе Василий строит минимизируя сумму квадратов ошибок. Какие прогнозы внутри обучающей выборки получит Василий с помощью своего леса?

b) Василий решил использовать бустинг с темпом обучение η. Прогнозы в каждом листе конкретного дерева Василий строит минимизируя функцию:

$$Q = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{T} w_j^2,$$

где y_i- прогнозируемое значение для i-го наблюдения, N- количество наблюдений, w_j прогноз в j-ом листе, T- количество листов на дереве. Какие прогнозы внутри обучающей выборки получит Василий при $\eta = 1$ и $\lambda = 1$?