1. Ёжику досталась обучающая выборка для задачи классификации:
целевая переменная $y = (1, -1, -1, 1, 1)$ и предиктор $x = (1, 2, 3, 4, 5)$.

   Медвежонок выбирает одно наблюдение из пяти равновероятно, обозначим с помощью Y и X полученные случайные значения целевой переменной и предиктора. Ёжик может задать Медвежонку один вопрос вида «правда ли, что X больше константы c?».

   Какую c надо выбрать Ёжику, чтобы минимизировать $H(Y|Q)$, где Q — это ответ Медвежонка?

   Подсказка: если $h(t) = -t \ln t$, то $h(1/2) = 0.347$, $h(1/3) = 0.366$, $h(2/3) = 0.270$, $h(3/4) = 0.216$, $h(2/5) = 0.367$, впрочем, задачу можно полностью решить и без этих цифр :)

2. Ежу понятно, что математическое ожидание первой производной лог-функции правдоподобия тождественно равно нулю. Производная нуля равна нулю. Для хорошей функции правдопобия производная от ожидания первой производной равна ожиданию второй производной. Ожидание второй производной лог-функции правдоподобия со знаком минус называется информацией Фишера. Следовательно, информация Фишера всегда равна нулю.

   Найдите качественную ошибку в этом рассуждении.

3. Хорошо обученная свинья тратит на поиск одного трюфеля экспоненциальное время с ожиданием μ минут. Поиск различных трюфелей независим. Жерар Депардье, к сожалению, замерял время поиска 100 трюфелей только тремя диапазонами: от 0 до 10 минут (20 трюфелей), от 10 до 20 минут (50 трюфелей), более 20 минут (30 трюфелей).

   а) Постройте 95% асимптотический доверительный интервал для μ.

   б) С помощью LM-теста проверьте гипотезу о том, что $\mu = 10$ против альтернативной гипотезы о неравенстве для уровня значимости 5%.

   Табличное: правые 5% критические значения: $\chi^2_1 = 3.84$, $\chi^2_2 = 5.99$, $\chi^2_3 = 7.81$, $\chi^2_4 = 9.49$, функция плотности экспоненциального распределения имеет вид $\lambda \exp(-\lambda x)$.

4. Ёж проверяет всего две гипотезы H_0^i одновременно, одна из которых верна, а вторая — нет. Для неверной гипотезы P-значение распределено равномерно на $[0; 0.5]$.

   Ёж сортирует P-значения по возрастанию, $p_{(1)} \le p_{(2)}$. Затем сравнивает $p_{(1)}$ с константой 0.05. Если $p_{(1)} \ge 0.05$, то обе $H_0^{(i)}$ не отвергаются и Ёж заканчивает работу. Если $p_{(1)} < 0.05$, то $H_0^{(1)}$ отвергается и Ёж сравнивает $p_{(2)}$ с константой b. Если $p_{(2)} < b$, то $H_0^{(2)}$ отвергается, иначе $H_0^{(2)}$ не отвергается.

   Постройте график зависимости FDR (false discovery rate) от $b \in [0.05; 1]$.

5. Пчёлы бывают правильные ($b_i = \text{good}$) и неправильные ($b_i = \text{bad}$). Из одного дупла правильных пчёл можно извлечь случайное равномерное количество мёда, $(y_i \mid b_i = \text{good}) \sim U[0; a]$, где параметр a не известен и $a > 1$. Для одного дупла неправильных $(y_i \mid b_i = \text{bad}) \sim U[0; 1]$. Имеется n независимых наблюдений. Неизвестную вероятность того, что в дупле водятся правильные пчёлы, обозначим буквой π.

   Явно выпишите целевую функцию для M-шага EM-алгоритма в этой задаче, поясните по каким параметрам она оптимизируется и смысл остальных параметров.