Задание 1 по курсу «Байесовский выбор моделей»

Общая информация

- Время сдачи задания: 8е октября, 16:00 по Москве;
- Максимальная базовая оценка за задание 50 баллов, так что при желании можно выполнять не всё;
- Оценка автора наилучшей работы удваивается (с учетом баллов сверх 50), но не более, чем до 125 баллов;
- Вопросы и само задание принимаются по почте: aduenko1@gmail.com;
- Тема письма: вопрос по заданию #1 или решение задания #1;
- Опоздание на неделю снижает оценку в 2 раза, опоздание на час на $0.5^{1/(7 \cdot 24)} = 0.41\%$;
- Работы опоздавших не участвуют в конкурсе на лучшую работу;
- Задание не принимается после его разбора и / или после объявления об этом.

Задача 1 (10 баллов). Пусть проводится эксперимент по угадыванию стороны выпадания честной монеты. Известно, что оракул прав с вероятностью $p_1 = 0.9$, а обычный человек с вероятностью $p_2 = 0.5$. Известно, что человек P оказался прав во всех $n = 10$ бросаниях. С какой вероятностью P является оракулом, если случайные человек оказывается оракулом с вероятностью 10^{-4} (3 балла)? Пусть человек P выбран не случайно, а как лучший среди 100 человек по угадыванию $k = 100$ выпадений монеты. Вывести новую априорную вероятность того, что P оракул с учётом его неслучайного выбора (7 баллов).
а) аналитически (приближенно) для случая $k \gg 1$;
б) сэмплированием для значений k от 1 до 1000. Построить график для разных k.

Задача 2 (10 баллов). Пусть имеется НОР выборка $\{x_1, \ldots, x_n\}$ из неизвестного распределения с конечной плотностью. На уровне значимости $\alpha = 0.05$ проверить гипотезу о том, что десятипроцентная квантиль этого распределения равна $m_0 = 0$.

Задача 3 (25+10 баллов). Пусть имеется выборка пар $\mathbf{z}_i = (x_i, y_i)$, $i = \overline{1, n}$,

$$\mathbf{z}_i \sim N\left(\mathbf{z}_i | (0, 0)^\mathsf{T}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right).$$

Гипотеза $\mathbf{H_0}$: $\rho = 0$
Для статистики $T(\mathbf{Z}) = \dfrac{1}{n} \sum_{i=1}^n x_i y_i$ получить
а) распределение для разных значений ρ и нарисовать плотность для $\rho = 0$ и $\rho = 0.5$ для $n = 100$ (2 балла);
б) построить критерий для проверки гипотезы $\rho = 0$ на уровне значимости $\alpha = 0.05$ (3 балла);
в) зависимость мощности данного критерия от истинного ρ
сэмплированием (5 баллов) и приближенно аналитически (5 баллов), и предложить форму (3 балла) зависимости мощности критерия от n и ρ.

Сравнить мощность (7 баллов) в зависимости от ρ со статистикой $T(\mathbf{Z}) = \dfrac{1}{2n} \sum_{i=1}^n (x_i - y_i)^2$, рассмотренной на лекции.
Какую статистику вы предложили бы для использования на практике? (+10 баллов тому, кто предложит свою и аргументированно обоснует, что она лучше).

Задача 4 (15 баллов). Пусть $\mathbf{x} = \{x_1, \ldots, x_n\}$, $n = 12$ есть НОР выборка из $\mathcal{N}(0, 1)$. Пусть $\mathbf{y} = \{y_1, \ldots, y_n\}$, $n = 12$ есть НОР выборка из $\mathcal{N}(0, 1)$, независимая от $\{x_1, \ldots, x_n\}$.

Оценить (сэмплированием или приближенно аналитически), сколько разных (и независимых) выборок $\mathbf{y}$ нужно рассмотреть K, чтобы найти ту, которая дает выборочную корреляцию с $\mathbf{x}$ не менее $\rho = 0.97$ (5 баллов). Построить график зависимости $K(\rho)$ в диапазоне от 0 до 0.99 (5 баллов). Какой прикладной вывод можно сделать из этого эксперимента помимо известного «корреляция не означает причинность» (5 баллов)?

Задача 5 (5 баллов). Привести пример, когда наивный байесовский классификатор классифицирует объекты не лучше, чем наугад, хотя генеральная совокупность (все возможные объекты) идеально разделима?

Задача 6 (15 баллов). В условиях задачи 3 при $n = 100$ сэмплировать $m = 1000$ выборок пар $\mathbf{z}_i$, $i = \overline{1,\ m}$, для 500 из которых $\rho = 0$ и $\rho = 0.2$ для оставшихся. С помощью одной из рассмотренных (или своей) статистик получить достигаемые уровни значимости $p_1, \ldots, p_m$. Для уровня значимости $\alpha = 0.05$ сравнить результаты применения отсутствия поправки на множественное тестирование, поправки Бонферрони и поправки Бенджамини-Хохберга в терминах получения ложных открытий (ложно отклоненные гипотезы) или пропуска таковых (ложно принятые гипотезы). Контролирует ли поправка Бенджамини-Хохберга FDR на уровне α и почему? (8 баллов)
Рассмотреть отдельно 1000 выборок для $\rho = 0$ и повторить эксперимент. сравнить результаты применения отсутствия поправки на множественное тестирование, поправки Бонферрони и поправки Бенджамини-Хохберга в терминах получения ложных открытий (ложно отклоненные гипотезы) или пропуска таковых (ложно принятые гипотезы). Контролирует ли поправка Бенджамини-Хохберга FDR на уровне α и почему? (7 баллов)

Задача 7 (10 баллов). В условиях задачи 6 сэмплировать $m = 1000$ выборок пар, но с ρ_m, зависящим от номера выборки. Провести те же исследования, что и в задаче 6.

$$\rho_1 = 0,\ \rho_i = \begin{cases} \rho_{i-1}, \text{ с вероятностью } 0.3, \\ 0.2 - \rho_{i-1}, \text{ с вероятностью } 0.7. \end{cases}$$