Задачи к лекции “Метод наименьших квадратов”

Задача 1

Пусть дана выборка точек y_i. Решите задачу МНК, моделируя данные постоянной величиной $\breve{y}$, что отвечает минимизации функции потерь

$$\mathcal{L} = \sum_{i=1}^{l} (y_i - \breve{y})^2 \quad \to \quad \min_{\breve{y}} . \tag{1}$$

Задача 2

Для четырех выборок из *квартета Энскомба* вычислите выборочные дисперсии x и y координат, а также коэффициент линейной корреляции Пирсона. Изобразите выборки на графиках. Данные можно получить в системе jupyter с помощью библиотеки `seaborn`, вызвав метод `load_dataset('anscombe')`.

Задача 3

На лекции обсуждалось, что метод наименьших квадратов — это способ поставить задачу о решении *переопределенной* системы $Xw = y$, которая имеет явный ответ, выражающийся через левую псевдообратную матрицу для X. Для *недоопределенной* системы $Xw = y$ (имеющей бесконечно много решений) можно поставить задачу о поиске решения с минимальной l_2-нормой весов $\|w\|^2 = w^T w$. Решите такую задачу и покажите, что ответ выражается через правую псевдообратную матрицу для X. Считайте, что прямоугольная матрица X имеет полный ранг (максимально возможный).

Задача 4

Обработайте какую-нибудь лабораторную работу (например, из курса общей физики), требующую проведения прямой по экспериментально полученным точкам. Для решения задачи регрессии рекомендуется использовать библиотеку `scikit-learn (sklearn)` или `scipy`.

Задача 5

На лекции обсуждался учет влияния систематической погрешности путем усреднения решения задачи МНК по гауссовому нормальному распределению для y-координат точек выборки: $\tilde{y}_i \sim \mathcal{N}(y_i, s^2)$, где погрешность по оси ординат считалась равной s. Обобщите этот вывод на случай, когда каждая точка имеет свою y-погрешность s_i. Для этого проведите усреднение по многомерному нормальному распределению для $\tilde{y}_i$ с произвольной симметричной матрицей ковариации A^{-1}:

$$\tilde{y} \sim \frac{1}{(2\pi)^{l/2} \det A} \exp\left(-\frac{(\tilde{y} - y)^T A (\tilde{y} - y)}{2}\right), \tag{2}$$

где $y = \begin{pmatrix} y_i & \dots & y_l \end{pmatrix}^T$, а $\tilde{y} = \begin{pmatrix} \tilde{y}_i & \dots & \tilde{y}_l \end{pmatrix}^T$.

1. Покажите, что распределение (2) правильно нормированно. *Указание*: Выполните замену координат $\tilde{y} - y = Sz$, где матрица S диагонализует A.

2. Вычислите неприводимые парные корреляторы $\langle\langle \tilde{y}_i \tilde{y}_j \rangle\rangle$, усредняя по распределению (2). *Указание*: Сделайте замену $\tilde{y} - y = Y$. Для вычисления гауссового интеграла с предэкспонентой вычислите интеграл $\int d^l Y \, \exp\left(-Y^T A Y/2 + J^T Y\right)$ и выполните дифференцирование по параметрам J_i (компоненты вектора J).

3. Оцените погрешности параметров модели w_α, следуя вычислению, приведенному на лекции, и используя корреляторы, полученные в предыдущем пункте.
4. Запишите решение в частном случае диагональной матрицы $A = \mathrm{diag}\,(A_1, \ldots, A_l)$. Как следует выбирать величины A_i для моделирования y-погрешности i-ой точки, равной s_i?

Задача 6*

Выполните оценку погрешности весов w_α, учитывая систематическую погрешность x-координат точек выборки, усреднив решение задачи МНК по гауссовому нормальному распределению $\tilde{x}_i \sim \mathcal{N}(x_i, s_i^2)$. Для простоты считайте погрешности для каждой точки равными: $s_i = s$ и пренебрегайте погрешностью y-координат. *Указание*: разложите аналитическое решение задачи МНК в ряд Тейлора по отклонениям $\tilde{x}_i - x_i$, считая такое разложение допустимым.