fit, predict, и всё готово

Линейная регрессия

1. Пусть $X_1, \ldots, X_n$ — независимые случайные величины, где X_i распределена по нормальному закону $\mathcal{N}(a + bi, \sigma^2)$. Постройте несмещённые оценки параметров a, b.

2. Пусть в модели линейной регрессии помимо k вещественных признаков имеется категориальный признак, принимающий d различных значений. Для j-ого из них рассматривается своя линейная зависимость с n_j объектами, которые этим признаком обладают:

$$\mathbf{X}^{(j)} = Z^{(j)}\boldsymbol{\theta} + \boldsymbol{\varepsilon}^{(j)},$$

где $\mathbf{X}^{(j)} \in \mathbb{R}^{n_j}$, $Z^{(j)} \in \mathbb{R}^{n_j \times k}$. Сведите задачу к одной линейной регрессии с той же МНК-оценкой. Чем это отличается от тупого добавления $d - 1$ onehot признаков?

3. Недоопределение. В модели линейной регрессии $\mathbf{X} = Z\boldsymbol{\theta} + \boldsymbol{\varepsilon}$, где $\mathbf{X} \in \mathbb{R}^n$, $Z \in \mathbb{R}^{n \times k}$, $\theta_i \neq 0$, МНК-оценка $\widehat{\theta}$ вектора $(\theta_1, \ldots, \theta_m)^T$ строится только по первым $m < k$ столбцам матрицы Z (статистик думает, что целевая величина зависит лишь от первых m признаков, когда на самом деле от всех k). Докажите, что полученная оценка является смещённой.

4. Переопределение. В модели линейной регрессии $\mathbf{X} = Z\boldsymbol{\theta} + \boldsymbol{\varepsilon}$, где $\mathbf{X} \in \mathbb{R}^n$, $Z \in \mathbb{R}^{n \times k}$, МНК-оценка $\widehat{\theta}$ вектора $(\theta_1, \ldots, \theta_m)^T$ строится сразу по $m > k$ признакам с матрицей

$$Z' = (\mathbf{z}_1, \ldots, \mathbf{z}_k, \mathbf{z}_{k+1}, \ldots, \mathbf{z}_m) = (Z, \ldots)$$

(статистик думает, что целевая величина зависит от всех m признаков, когда на самом деле только от первых k, то есть $\theta_{k+1} = \ldots = \theta_m = 0$). Докажите, что для всех $i = 1, \ldots, k$ выполнено неравенство $\mathsf{D}_{\boldsymbol{\theta}}\widehat{\theta}_i \geq \mathsf{D}_{\boldsymbol{\theta}}\widetilde{\theta}_i$, где $\widetilde{\boldsymbol{\theta}}$ — МНК-оценка, построенная по матрице Z (иными словами, добавление бесполезных признаков увеличило дисперсию оценок нужных параметров).

5. Постройте доверительный интервал уровня доверия γ для целевой величины, отвечающего некоторому набору признаков $\mathbf{z}_0$ (то есть $\mathbf{z}_0^T\boldsymbol{\theta} + \varepsilon$, где $\varepsilon \sim \mathcal{N}(0, \sigma^2)$) для модели гауссовской линейной регрессии $\mathbf{X} = Z\boldsymbol{\theta} + \boldsymbol{\varepsilon}$ (обычно ДИ для случайной величины называют *предсказательным*).

6. Пусть $X_1, \ldots, X_n \sim \mathcal{N}(a, \sigma^2)$ и $Y_1, \ldots, Y_m \sim \mathcal{N}(b, \sigma^2)$ — независимые выборки. Сведя задачу к линейной регрессии, постройте доверительный интервал для $a - b$.

7. Докажите, что статистика $(\widehat{\boldsymbol{\theta}}, \|\mathbf{X} - Z\widehat{\boldsymbol{\theta}}\|^2)$ является **(а)** достаточной; **(б)*** полной в модели гауссовской линейной регрессии.

8.* В модели линейной регрессии $\mathbf{X} = Z\boldsymbol{\theta} + \boldsymbol{\varepsilon}$, где $\mathsf{E}\boldsymbol{\varepsilon} = \mathbf{0}$, $\mathsf{D}\boldsymbol{\varepsilon} = \sigma^2 V$ (V — п.о. матрица) найдите лучшую оценку среди линейных несмещённых в среднеквадратичном подходе.