Методические материалы

Экзамен по анализу данных – Базовый уровень

15 июня 2022 г.

- Среднее значение переменной X.
$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$
- Выборочная дисперсия переменной X (несмещённая).
$$\mathrm{sVar}(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$
- Выборочный коэффициент корреляции Пирсона для переменных X и Y.
$$\mathrm{scorr}(X, Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}.$$
- Для линейной (парной) регрессии
$$\hat{Y}_i = \hat{w}_0 + \hat{w}_1 X_i$$
оценки коэффициентов рассчитываются по формулам
$$\hat{w}_0 = \bar{Y} - \hat{w}_1\bar{X},$$
$$\hat{w}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$
- Среднеквадратичная ошибка.
$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$
- Средняя абсолютная ошибка.
$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \hat{Y}_i\right|.$$
- Логистическая функция.
$$\sigma(X) = \frac{1}{1 + e^{-X}}.$$
- Формулы для ROC-кривой.
$$\mathrm{TPR} = \mathrm{Sensitivity} = \mathrm{Recall} = \frac{TP}{TP + FN} = \frac{\sum_{i=1}^{n}\mathbb{I}(Y_i = +1)\mathbb{I}(\hat{Y}_i = +1)}{\sum_{i=1}^{n}\mathbb{I}(Y_i = +1)},$$
$$\mathrm{FPR} = 1 - \mathrm{Specificity} = \frac{FP}{FP + TN} = \frac{\sum_{i=1}^{n}\mathbb{I}(Y_i = -1)\mathbb{I}(\hat{Y}_i = +1)}{\sum_{i=1}^{n}\mathbb{I}(Y_i = -1)}.$$

- Статистика χ^2 критерия согласия Пирсона.

$$\chi^2 = \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{(O_{i,j}-E_{i,j})^2}{E_{i,j}} \overset{H_0}{\sim} \chi^2_{(R-1)(C-1)},$$

  где

  - R и C – число строк и столбцов в таблице сопряжённости.
  - $O_{i,j}$ – количество наблюдений в клетке (i,j) таблицы сопряжённости.
  - N – число наблюдений в выборке.
  - $E_{i,j} = Np_{i\cdot}p_{\cdot j}$
  - $p_{i\cdot} = \sum_{j=1}^{C}\dfrac{O_{i,j}}{N}$
  - $p_{\cdot j} = \sum_{i=1}^{R}\dfrac{O_{i,j}}{N}$

- Пусть имеется выборка независимых одинаково распределённых случайных величин $X_1, \ldots, X_n$, где $X_i \sim \mathcal{N}(\mu, \sigma^2)$. Тогда Z-статистика для проверки гипотезы

$$\begin{cases} H_0 : \mu = \mu_0, \\ H_1 : \mu \neq \mu_0 \end{cases}$$

  рассчитывается по формуле

$$Z_{obs} = \frac{\bar{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$$

- 95%-ый доверительный интервал для математического ожидания независимых одинаково распределённых случайных величин $X_1, \ldots, X_n$, где $X_i \sim \mathcal{N}(\mu, \sigma^2)$:

$$\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leqslant \mu \leqslant \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}$$

- Z-статистика для проверки гипотезы о равенстве матожиданий двух независимых нормальных выборок $X_1, \ldots, X_{n_1}$ и $Y_1, \ldots, Y_{n_2}$ с известными диспресиями:

$$Z_{obs} = \frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_1}+\frac{\sigma_Y^2}{n_2}}} \sim \mathcal{N}(0,1).$$

- t-статистика для проверки гипотезы о равенстве матожиданий двух независимых нормальных выборок $X_1, \ldots, X_{n_1}$ и $Y_1, \ldots, Y_{n_2}$ с неизвестными, но равными дисперсиями:

$$t = \frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{\hat{\sigma}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2},$$

$$\hat{\sigma} = \sqrt{\frac{(n_1-1)\hat{\sigma}_X^2+(n_2-1)\hat{\sigma}_Y^2}{n_1+n_2-2}}$$

- t-статистика для проверки гипотезы о равенстве матожиданий для нормальных парных выборок $X_1, \ldots, X_n$ (до) и $Y_1, \ldots, Y_n$ (после):

$$d_i = X_i - Y_i,$$

$$t = \frac{\bar{d}-\mu_d}{\frac{\sigma_d}{\sqrt{n}}} \sim t_{n-1},$$

$$\sigma_d = \sqrt{\frac{1}{n-1}\left(\sum_i d_i^2 - \frac{(\sum_i d_i)^2}{n}\right)}.$$

Некоторые модули

- Z-test: `from statsmodels.stats.weightstats import ztest`
- t-test:
  - `from scipy.stats import ttest_1samp`
  - `from scipy.stats import ttest_ind`
  - `from scipy.stats import ttest_rel`
- Проверка гипотезы о доле: `from scipy.stats import binom_test`
- Критерий согласия для проверки независимости: `from scipy.stats import chi2_contingency`
- Линейная регрессия: `from sklearn.linear_model import LinearRegression`
- Логистическая регрессия: `from sklearn.linear_model import LogisticRegression`
- MSE, MAE, R^2: `from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score`