

## Практика 12. Оценка погрешности регрессионной модели на скользящем контроле и по доверительному интервалу.

### Часть 1. Скользящий контроль

1. Загрузить данные по характеристикам насосных агрегатов по своему варианту.
2. Реализовать скользящий контроль LOO (leave-one-out) и оценить дисперсию шума на скользящем контроле по формуле

$$\hat{D}_{LOO} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}(x_i))^2$$

3. Оценить дисперсию шума без выделения экзаменационной выборки по формуле

$$\hat{D}_\varepsilon = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}(x_i))^2$$

4. Оценить дисперсию шума без выделения экзаменационной выборки по формуле

$$S_\varepsilon^2 = \frac{1}{n-k} \sum_{i=1}^n (y_i - \hat{y}(x_i))^2$$

5. Сравнить все три дисперсии, сделать выводы (на защите)

### Часть 2. Доверительный интервал прогноза

1. Для каждого разбиения выборки на скользящем контроле LOO рассчитать 95%-доверительный интервал прогноза нового значения.
2. Вывести количество и долю попаданий нового значения в доверительный интервал по всем разбиениям. Доля попаданий должна быть близка к 95%.

### Вопросы к защите

1. Скользящий контроль
  - a. Что такое эмпирическая модель? Какие еще бывают модели, которые нельзя назвать эмпирическими?
  - b. Что такое обучающая и экзаменационная выборка?
  - c. Эффект оптимистической смещенности при оценивании погрешности эмпирической модели по обучающей выборке. В чем заключается, к чему приводит?
  - d. Какие факторы влияют на величину оптимистической смещенности?
  - e. Что такое проблема переобучения? Как она связана с эффектом оптимистической смещенности?
  - f. Какие из использованных оценок дисперсии являются оптимистически смещенными, а какие нет?
  - g. Докажите несмещенность  $S_\varepsilon^2$
  - h. Почему  $\hat{D}_\varepsilon$  является смещенной?
  - i. Почему  $\hat{D}_{LOO}$  является несмещенной, несмотря на использование такой же формулы, как для  $\hat{D}_\varepsilon$ ?
  - j. Контроль по одному элементу и контроль по блокам (LOO, q-fold).
  - k. Расчет дисперсии на скользящем контроле LOO. Вывод формул  $\hat{D}_{LOO}$ .
  - l. Сопоставьте численные значения оценок  $\hat{D}_\varepsilon$  и  $\hat{D}_{LOO}$ . Проявился ли эффект оптимистической смещенности  $\hat{D}_\varepsilon$  относительно  $\hat{D}_{LOO}$ ?
  - m. Сопоставьте численные значения оценок  $S_\varepsilon^2$  и  $\hat{D}_{LOO}$ . Объясните отличие в числах.
2. Доверительные интервалы прогноза регрессионной модели

- a. Формулы дисперсий (ковариаций)  $\hat{\beta}, \hat{y}, e_n, e_{in}$ . Вывести формулы соответствующих оценок, с учетом того, что вместо истинной дисперсии  $D_\varepsilon$  имеется только ее оценка  $S_\varepsilon^2$ .
- b. Формирование t-статистики для расчета доверительного интервала прогноза нового значения при неизвестной дисперсии шума.
  - i. Доказать несмещенность прогноза
  - ii. Вывести МО числителя t-статистики, доказать, что оно нулевое.
  - iii. Вывод степеней свободы для знаменателя.
- c. Расчет доверительного интервала на основе t-статистики.
- d. Чем обусловлена погрешность прогноза истинного значения?
- e. Будет ли погрешность нового значения равна нулю, если  $\hat{\beta} = \beta_{\text{ист}}$ ?