Тестовое задание. Теория

Алина Ахметшина

August 2020

1 Вопросы по SQL

1.1 Какие типы соединения таблиц вы знаете?

- INNER JOIN или просто JOIN – соединение таблиц, возвращающее данные, пересекающиеся по условию (INNER JOIN table1, table2 ON condition). Возвращаются только совпадающие данные
- OUTER JOIN – возвращает все данные: и те, которые пересекаются по условию, и те, которые по условию не подходят, в них пропуски заполняются NULL. RIGHT/LEFT OUTER JOIN указывают на то, какая из таблиц будет внешней
- CROSS JOIN – возвращает перекрестное объединение двух таблиц, то есть полное соответствие каждой строки одной таблицы каждой строке другой. (при данном объединение указывать условие не нужно)
- FULL JOIN – операция, объединяющая в себе RIGHT и LEFT OUTER JOIN
- MULTI JOIN – возвращает значения при наличие нескольких условий

1.2 Приведите примеры агрегатных функций, что они считают?

В качестве аргумента агрегатные функции принимают имя столбца, а в качестве ответа выдают число.

- AVG – возвращает среднее значение
- COUNT – возвращает количество строк, если имеются значения NULL, то то же самое можно получить, передав в качестве аргумента *
- MAX/MIN – возвращает максимальное/минимальное значение
- SUM – возвращает сумму значений столбца
- VAR – возвращает статистическую дисперсию значений

1.3 Какие операторы работы с наборами вы знаете, что они выполняют?

1.4 Что такое CTE?

CTE – табличное выражение, позволяющее в рамках запроса создать таблицу, с возможностью многократно на нее ссылаться

2 Практические задания по SQL

2.1 Написать скрипты на языке SQL, чтобы получить список кредитов, которые на момент расчета имеют непогашенную задолженность, и рассчитать для каждого такого кредита:

- 1 Общую (накопленную) сумму просроченного долга непогашенную (не выплаченную) к моменту расчета

  SELECT dates, deal, SUM(Sum), fullsum

  FROM "PDCL"

  GROUP BY deal

  HAVING (fullsum>0 AND dates)

3 Вопросы по Oracle

3.1 Виды джоинов?

См. вопрос 1.1

3.2 Что произойдет про джоине по условию on a.id=1?

Произойдет соединение, результатом которого будет таблица, в которой строки будут состоять из значений строк первой таблицы и значений первой строки второй таблицы. Пример:

table1 column1	table1 column2	table2 column1
item11	item12	item21
item13	item14	item21
item16	item15	item21

3.3 Можно ли использовать оконные функции в апдейте? А в условии where?

4 Вопросы по статистике и машинному обучению

4.1 Какие описательные статистики вы знаете, какие применяли на практике? Объясните своими словами что они означают и зачем они нужны.

Стандартное отклонение, медиана, мода, дисперсия, диапазон, взвешенное среднее,среднее арифметическое.

В рамках анализа данных, полученных в ходе лабораторных практик, и практических заданий, выполненных в ходе курсов, мною были применены все вышеперечисленные описательные статистики.

Диапазон – интервал значений между максимальным и минимальным членом выборки.

Медиана характеризует то, насколько усреднен полученный упорядоченный набор. То есть для ряда упорядоченных значений половина будет больше медианы, а половина - меньше.

Мода – наиболее часто встречающееся в выборке значение Дисперсия – величина, характеризующая то, насколько сильно значение каждого элемента выборки отклоняется от среднего.

Среднеквадратическое отклонение показывает насколько сильно значения элементов рассеяны относительно ее среднего. В отличие от дисперсии имеет ту же размерность, что и сама величина.

Взвешенное среднее используется, когда некоторые значения более важны, чем остальные. При весах равных 1, совпадает со средним арифметическим.

4.2 Каким образом сравниваются статистики в двух выборках?

При сравнении двух независимых выборок, если признак измерен количественно или интервальной шкалой и его распределение близко к нормальному, то можно воспользоваться критерием Стьюдента. В противном случаем можно воспользоваться критерием Манна-Уитни.

Критерий Стьюдента проверяет гипотезу о равенстве значения средних в двух группах. При условии, что дисперсии не равны делается дополнительная поправка.

В критерии Манна-Уитни сравниваются порядковые ранги, то есть рассматриваются не средние значения в группах, а средние ранги, получаемые после построения вариационного ряда и присвоению элементам номеров.

4.3 Какие показатели определяют зависимость между двумя переменными?

Зависимость между двумя переменными определяется при помощи ковариации, коэффициентов корреляции и регрессии.

Коэффициент регрессии показывает на сколько единиц увеличивается вторая переменная при изменении первой на 1.

4.4 Каким образом делается вывод о наличие линейной зависимости между двумя переменными?

Наличие линейной зависимости между двумя переменными определяется при помощи коэффициентов корреляции Пирсона и Спирмена.

Величина абсолютного значения коэффициента Пирсона характеризует наличие линейной зависимости между двумя переменными.

В коэффициенте корреляции Спирмена для оценки линейной зависимости используются не сами значения коэффициентов, а их ранги.

4.5 Если вы знаете более одного показателя, то расскажите в каких случаях применяется каждый из названных показателей.

Коэффициент корреляции Спирмена применяется, если обе переменные ранжированы, а Пирсона, если обе переменные заданы в метрических шкалах и их распределение близко к нормальному.

4.6 Чем отличается линейный коэффициент корреляции от корреляции Пирсона?

Коэффициент корреляции Пирсона = линейный коэффициент корреляции

$$r_{x,y} = \frac{\sum_i (x_i - M_x)(y_i - M_y)}{\sqrt{\sum_i (x_i - {}_x)^2 (y_i - M_y)^2}} \tag{1}$$

4.7 Опишите своими словами что означает коэффициент детерминации.

Коэффициент детерминации характеризует качество используемой модели. То есть чем ближе R^2 к 1, тем большую долю дисперсии, объясняет модель.

4.8 Какие методы регрессионного анализа вы знаете, какие применяли?

Линейная, нелинейная регрессии, ridge, lasso.

Многомерную линейную регрессию в задании по предсказанию выручки компании в зависимости от уровня ее инвестиции в рекламу.

4.9 Какие методы классификаций вы знаете, какие применяли?

Метод ближайших соседей, метод решающих деревьев, метод окна Парзена, метод потенциальных функций, генетический алгоритм, персептрон.

Применяла метод k ближайших соседей и метод решающих деревьев.

Логистическую регрессию в задаче классификации изображения, метод решающих деревьев в задаче о предсказании выжили пассажир Титаника или нет. Метод k ближайших соседей в задаче на предсказания сорта винограда по результатам химического анализа.

4.10 В чем отличие между классификацией и кластеризацией?

В задачах классификации имеется конечное число известных классов, данный тип относится к классу обучения с учителем, то есть имеются объекты и ответы, и по данным парам необходимо восстановить общую зависимость.

Кластеризация относится к классу обучения без учителя, то есть имеется набор объектов, которые необходимо сгруппировать, однако количество групп (кластеров) и то какие это группы зачастую неизвестно. Вследствие чего, сложно измерить качество решения, в то время как для задачи классификации существуют различные метрики качества, например precision-recall, F мера.

4.11 Какие методы анализа временных рядов вы знаете, какие применяли?

Метод линии тренда, простое скользящее среднее, взвешенное скользящее среднее, экспоненциальное сглаживание, метод экспоненциального сглаживания с учетом тренда, метод прогнозирования сезонных изменений с помощью моделей SARMA и SARIMA или Модели Хольта-Винтерса (тройное экспоненциальное сглаживание).

4.12 В чем отличие метода главных компонент и факторного анализа?

Факторный анализ - метод, позволяющий классифицировать признаки и уменьшить число признаков, оставив из избыточного набора только необходимые, а также сжать количество признаков в более компактный набор с сохранением информации.

Метод главных компонент - единственный статистически обоснованный метод факторного анализа, в ходе которого коррелированные признаки заменяются на некоррелированные, построенные на их основе.

В отличие от МГК, который имеет под собой математическое обоснование, Факторный анализ позволяет только структурировать пространство признаков и не дает возможности обобщить это на генеральную совокупность.

4.13 Что такое функционал качества, напишите функции потерь для методов регрессионного анализа, классификации.

Функционал ошибки - это характеристика качества для данной задачи, на вход принимается алгоритм и выборка и возвращается характеристика того, насколько хорошо данный алгоритм работает на этой выборке.

$$Q(a, X^l) = \frac{1}{l}\sum_{i=1}^{l} \mathscr{L}(a, x_i) \tag{2}$$

Функционал качества алгоритма a на выборке X^l

$\mathscr{L}(a, x) = 0, 1$ - бинарная функция потерь

$\mathscr{L}(a, x) = |a(x) - y(x)|$ - абсолютная ошибка

$\mathscr{L}(a, x) = (a(x) - y(x))^2$ - квадратичная функция потерь

$\mathscr{L}(a, x) = -(a \cdot \log a(x) - (1 - x) \cdot \log(1 - a(x)))$ - логарифмическая функция потерь

4.14 В чем заключаются проблемы недообучения и переобучения модели?

При переобучении у алгоритма, показывающего хорошие результаты на обучающей выборке, не получается выявить в ней закономерности и применить их для классификации новых объектов.

При недообучении алгоритм плохо описывает и саму обучающую выборку и новые данные.

4.15 Из-за чего возникает переобучение, как обнаружить и как от него избавиться?

Переобучение возникает при слишком сильной подгонке алгоритма под обучающую выборку, одним из признаков являются большие веса, соответствующие признакам. Также переобучение может возникнуть из-за мультиколлинеарности или линейной зависимости признаков.

Оценить качество алгоритма и найти долю ошибок при классификации можно при помощи отделения части обучающей выборки, на которой не должно проходить обучение, и после проверить на ней обученный алгоритм. Данный метод является частным случаем кросс-валидации, при которой выборка разделяется на k частей примерно одинакового размера, после чего каждый блок используется в качестве тестового, в то время как остальные - в качестве обучающей выборки.

Для устранения переобучения можно упростить семейство алгоритмов, т.е. брать более простые. Или воспользоваться регуляризацией.

4.16 Какие методы поиска параметров вы знаете? Приведите пример: модель - ф-ия потерь - алгоритм обучения(поиск параметров)

4.17 Что такое регуляризация? Какие типы вы знаете, зачем она нужна?

Регуляризация - это способ борьбы с переобучением модели, основанный на наложении ограничения на абсолютное значение весов признаков.

Мне известны два типа регуляризаторов: Ridge(L2) и Lasso(L1). В L2 к основному функционалу вводится дополнительное слагаемое, которое штрафует избыточное увеличение нормы векторов коэффициентов.

$$Q(w, X) + \lambda ||w||^2 \to \min_w \tag{3}$$

Где λ - коэффициент регуляризации

Lasso это другой подход к регуляризации, когда вместо минимизации суммы квадратов коэффициентов мы минимизируем сумму их модулей. В отличае от Ridge он не гладкий, следовательно на нем нельзя реализовать метод градиентного спуска.

$$Q(w, X) + \lambda ||w|| \to \min_w \tag{4}$$

4.18 В чем отличие метода наименьших квадратов и принципа максимума правдоподобия?

Метод максимума правдоподобия ищет оценки неизвестных параметров. Оценка выбирается с помощью максимизации функции правдоподобия.

В случае линейной регрессии задача поиска МНК оценки эквивалента задаче поиска ОМП.

$$y_i = \sum_j \beta_j X_{j,i} + \xi_i, \quad \xi_i \sim \mathcal{N}(0, \sigma^2)$$

Функция правдоподобия для нормального распределения:

$$C \cdot \exp\left(-\sum_i \frac{\xi_i^2}{2\sigma^2}\right) = C \cdot \exp\left(-\sum_i \frac{(y_i - \sum_j \beta_j X_{j,i})^2}{2\sigma^2}\right),$$

где C — константа.

$$arg\max_\beta \; C \cdot \exp\left(-\sum_i \frac{(y_i - \sum_j \beta_j X_{j,i})^2}{2\sigma^2}\right) = arg\min_\beta \sum_i \left(y_i - \sum_j \beta_j X_{j,i}\right)^2$$

4.19 Объясните просто, что такое собственный вектор и определитель матрицы?

Собственный вектор матрицы — ненулевой вектор, переходящий в ему коллинеарный, при умножении на него данной матрицы, число строк которой равно числу элементов вектора.

Определитель матрицы — ориентированный объем многомерного параллелепипеда, простроенного на векторах-столбцах этой матрицы. Понятие определителя имеет смысл только для квадратных матриц.

4.20 Объясните просто, что такое интеграл и производные 1, 2, 3 порядка?

Интеграл — площадь фигуры под графиком кривой.

$$\int f(x)dx = F(x) + C \tag{5}$$

$$\frac{dF}{dx} = f(x) \tag{6}$$

1 производная — предел отношения приращения функции к приращению аргумента.
$f'(x0) = \lim_{x \to x_0} \frac{\Delta f(x)}{\Delta x}$
2 производная, буквально, является скоростью изменения скорости изменения функции:
$f''(x) = (f'(x))' = \frac{d^2 f}{dx^2}$
3 производная:
$f'''(x) = (f''(x))' = ((f'(x))')'$
Геометрический смысл 1 и 2 производных:

- Первая производная — это tg угла наклона касательной к прямой. По знаку первой производной на определенном интервале можно определить возрастание или убывание функции
- знак 2 производной характеризует выпуклость или вогнутость функции.