

Задание 2. «Кластеризация»

- 1) Запустите SAS Enterprise Miner, откройте проект с именем "Prak"
- 2) В дереве проекта создайте новую диаграмму с именем Diag2
- 3) В дереве проекта добавьте набор данных из библиотеки SAMPSTO с именем Baseball. Этот набор данных содержит информацию о результатах игроков в бейсбол в 1986 году и за всю предыдущую карьеру.
- 4) В процессе подключения набора данных задайте метаданные для набора (полезно выбрать чекбокс Label при описании переменных): роль набора Raw; для переменной Name поставьте роль ID, переменной YrMajor (число лет в лиге), а также всем переменным, чье имя начинается с n (результаты 1986 года) или cr (результаты за всю карьеру) – роль Input, тип Interval. Отсальным переменным поставьте роль Rejected.
- 5) Перетащите узел с добавленным набором данных на диаграмму, подключите после него узел Cluster, а после Cluster узел Segment Profile. Обучите модель.
- 6) Сколько кластеров было сформировано узлом cluster? Используя информацию из графика CCC Plot (Summary Statistics) узла Cluster ответьте на вопрос, какое значение CCC было для выбранного числа кластеров, а также для числа кластеров на один больше и на один меньше?
- 7) Используя информацию из графиков узла Cluster (Segment Plot, Means Statistics, Input Mean Plot и другие) и из графиков узла Segment Profile ответьте на вопросы:
 - a. Вариант I. Какой кластер содержит самых опытных игроков (переменная YrMajor)?
Вариант II. Какой кластер содержит самых успешных по хитам в 1986 году игроков (переменная nHits)?
 - b. Сколько человек попало в этот кластер?
 - c. Какая переменная самая важная с точки зрения отделения этого кластера ото всей выборки (см. Segment Profile).
- 8) Выберите узел Cluster, в настройках нажмите кнопку Exported data, выберите Browse. Ответьте на вопрос, как зовут самого типичного игрока в кластере вашего варианта (у него должно быть минимальное расстояние до центра вашего кластера)?
- 9) Параллельно с существующей веткой после узла с набором данных подключите узел SOM, задав размер сетки 2 на 2 (раздел настроек Segment, Rows=2, Columns=2) и после узла SOM узел Segment Profile (после обучения SOM, откройте Edit Variables на узле Segment Profile и поставьте Use = no для переменных с координатами на решетке SOM_Dimension).
- 10) Ответьте на все те же вопросы, что заданы в пункте 7, но после кластеризации SOM.
- 11) Ответы на вопросы включите в текстовый файл (в pdf формате), запишите диаграмму в формате xml (нажав на диаграмму в дереве проекта и выбрав Save as). Перешлите текстовый файл с ответами и xml с диаграммой в качестве результата зада

Задание 2.5. «Выявление скрытых признаков» (Вариант I)

- 1) После узла с подключенными данными Baseball подключите узел Variable Clustering, в настройках поставьте в Разделе Score->Variable Selection->Best variables. Обучите модель.
- 2) На сколько групп было разбито множество переменных? Какие представители каждой группы были выбраны? Какой процент вариации описывают отобранные переменные?

Задание 2.5. «Выявление скрытых признаков» (Вариант II)

- 1) После узла с подключенными данными Baseball подключите узел Principal Components, в настройках поставьте в eigenvalue cutoff->cumulative=0.8 (80% вариации) . Обучите модель.
- 2) Сколько главных компонент получилось? Какие собственные значения у полученных главных компонент? Глядя на график Principal Component coefficient, как бы вы охарактеризовали одной фразой признак, соответствующий первой главной компоненте, с точки зрения качества игроков, глядя на коэффициенты основных характеристик в первой главной компоненте?