Домашнее задание №3
Кластеризация

CMF 2016

Задание нужно сделать в **R** и оформить максимально подробно, с пояснениями, красивыми графиками и вот этим всем. Срок сдачи — до 23:59 18.10, то есть до конца вторника :}

Загрузите встроенный датасет Ирисы Фишера (*data(iris)*) и используйте известные вам алгоритмы кластеризации, чтобы разбить данные на подмножества. Там три типа цветочков, но мы об этом до поры до времени забудем.

1 k-means

Сначала используйте алгоритм k-means:

- постройте график внутригрупповой ошибки в зависимости от количества кластеров;
- выберите оптимальное количество кластеров и объясните свой выбор;
- проведите кластеризацию с тремя кластерами, сопоставьте результат с истинными типами цветочков и посчитайте процент объектов, которые попали не в свой кластер.

2 Иерархическая кластеризация

А теперь попробуем иерархическую кластеризацию:

- используйте метод как минимум с двумя метриками, постройте дендраграммы, сравните их;
- выберите оптимальное на ваш взгляд разбиение.

3 И снова PCA

Применим PCA (метод главных компонент), чтобы визуализировать результаты кластеризации:

- используйте метод главных компонент, чтобы сократить пространство признаков до двух;
- постройте диаграмму рассеяния (scatterplot);
- окрасьте точки на диаграмме в три цвета в зависимости от того, в какой кластер они попали при использовании алгоритма кластеризации (обоих или того, который вам больше нравится), а также отметьте центры кластеров (если выбрали k-means);
- сделайте это красиво!

В этом задании можно пользоваться встроенными функциями! Главное – сделать это правильно, а затем – правильно интерпретировать полученные результаты.