Количественные методы

Кластеризация

### Вариант: Россия, 4 кластера.

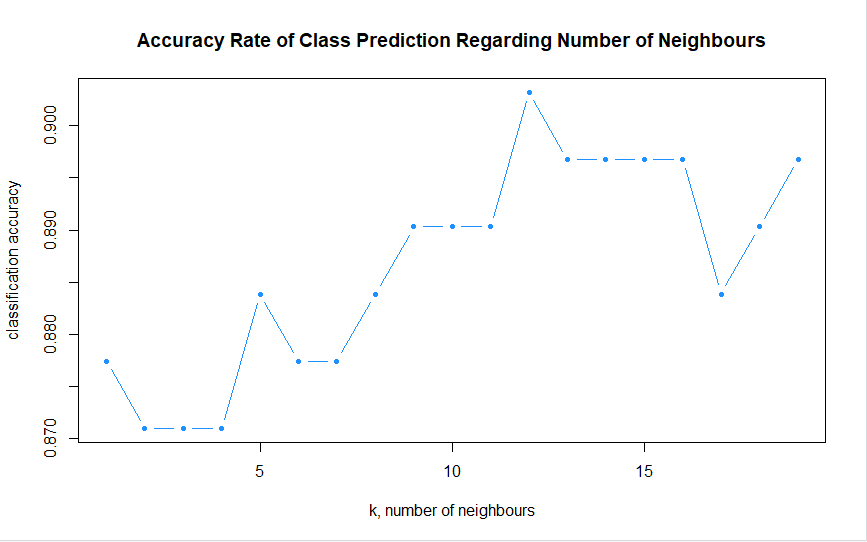
2. С помощью алгоритма kd-tree были найдены 5 ближайших соседей России по компонентам индекса счастья, а также Евклидовы расстояния в R^6 между Россией и соседями. Пять ближайших соседей России: Болгария, Латвия, Литва, Венгрия, Словакия. Из них только два развитых: Латвия и литва.

3. Время работы алгоритмов полного перебора и kd-tree поиска 5 ближайших соседей для одного наблюдения в выборке из 156 наблюдений по шести характеристикам составляло от 0.00 сек. (< 0.005 сек.) до 0.02 сек. при разных попытках инициализации алгоритма. Причём каждый раз время для двух алгоритмов совпадало. Это связано с несколькими факторами: во-первых, источник всего один, во-вторых, выборка небольшая, в-третьих, количество характеристик тоже небольшое.

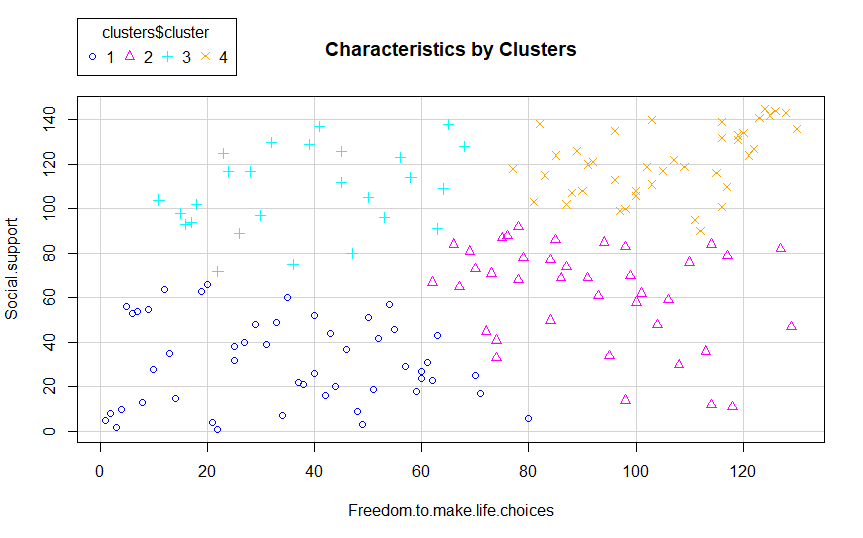
4. В качестве оценки класса страны по её 5 соседям я взял средневзвешенный класс этих соседей. Весами выступают величины, обратные Евклидовому расстоянию от страны, класс которой оценивается, до 5 её ближайших соседей по компонентам индекса счастья. То есть, если множество соседей страны: , то оценка её класса будет . Где – Евклидово расстояние по вектору рассматриваемых характеристик от страны, для которой значение предсказывается, до её i-го соседа. Предсказанное значение для России при таком методе оценивания: 0.

5. По выборке, не включающей Россию ошибка предсказания класса страны по 5 ближайшим соседям по компонентам индекса счастья составила .

Было бы хорошо провести процедуру кросс-валидации для более четкого ответа на следующий вопрос, но уже не успеваю это сделать. Так что просто рассмотрим график доли верно предсказанных классов для стран с использованием оценки вышеуказанного вида. Чтобы максимизировать точность для произвольной выборки такого размера без особой потери времени, затраченного на них, стоит взять больше 5, а даже и 10 соседей. Лично я бы брал 15.



6. Страны были разбиты на 4 кластера по компонентам «Социальная поддержка», «Свобода делать жизненный выбор». Содержательно они получились такими далее первое – для «Социальной поддержки», второе – для «Свободы делать жизненный выбор»: 1 кластер – низкая/низкая, 2 кластер – низкая/высокая, 3 кластер – высокая/низкая, 4 кластер – высокая/высокая. Россия оказалась в 3-ем кластере.



7. В результате кластеризации стран на 4 кластера по всем компонентам индекса счастья Россия попала в кластер с наименьшими индексами счастья (содержательно – удовлетворённость) по «Коррупции», «Щедрости»; с невысоким индексом по «Свободе делать жизненный выбор», и со средними по «ВВП на душу населения» и «Социальной поддержке».

8. Суммарное внутригрупповое расстояние минимизируется при размере кластера в 1 элемент. То есть на 156 кластеров, но в содержательном плане это ничего не даёт. 