*Визуализация биомедицинских данных*

Домашняя работа №3

Вам предстоит выполнить задания ниже в RMarkdown документе. После чего результат

(Rmd и результат knit’а) загрузить в ваш GitHub репозиторий.[[1]](#footnote-0) Домашнее задание сдается ссылкой на ваш репозиторий.

**Deadline: 17 ноября 2024 года**

Домашнее задание оценивается по системе зачёт/незачёт. Зачёт ставится при выполнении любых 9 заданий. Любые спорные ситуации при оценке решаются в пользу студента.

**Задания**

1. Загрузите датасет very\_low\_birthweight.RDS (лежит в папке домашнего задания).

Это данные о 671 младенце с очень низкой массой тела (<1600 грамм), собранные в Duke University Medical Center доктором Майклом О’Ши c 1981 по 1987 г.[[2]](#footnote-1) Описание переменных см. [здесь](https://hbiostat.org/data/repo/cvlbw). Переменными исхода являются колонки '*dead*', а также время от рождения до смерти или выписки (выводятся из '*birth*' и *'exit'*. 7 пациентов были выписаны до рождения).

Сделайте копию датасета, в которой удалите колонки с количеством пропусков больше 100, а затем удалите все строки с пропусками.

1. Постройте графики плотности распределения для числовых переменных. Удалите выбросы, если таковые имеются. Преобразуйте категориальные переменные в факторы. Для любых двух числовых переменных раскрасьте график по переменной *‘inout’.*
2. Проведите тест на сравнение значений колонки ‘*lowph*’ между группами в переменной *inout*. Вид статистического теста определите самостоятельно. Визуализируйте результат через библиотеку 'rstatix'. Как бы вы интерпретировали результат, если бы знали, что более низкое значение *lowph* ассоциировано с более низкой выживаемостью?
3. Сделайте новый датафрейм, в котором оставьте только континуальные или ранговые данные, кроме '*birth*', *'year'* и *'exit'*. Сделайте корреляционный анализ этих данных. Постройте два любых типа графиков для визуализации корреляций.
4. Постройте иерархическую кластеризацию на этом датафрейме.
5. Сделайте одновременный график heatmap и иерархической кластеризации. Интерпретируйте результат.
6. Проведите PCA анализ на этих данных. Проинтерпретируйте результат. Нужно ли применять шкалирование для этих данных перед проведением PCA?
7. Постройте biplot график для PCA. Раскрасьте его по значению колонки *'dead'*.
8. Переведите последний график в 'plotly'. При наведении на точку нужно, чтобы отображалось id пациента.
9. Дайте содержательную интерпретацию PCA анализу. Почему использовать колонку *'dead'* для выводов об ассоциации с выживаемостью некорректно?
10. Приведите ваши данные к размерности в две колонки через UMAP. Сравните результаты отображения точек между алгоритмами PCA и UMAP.
11. *Давайте самостоятельно увидим, что снижение размерности – это группа методов, славящаяся своей неустойчивостью.* Измените основные параметры UMAP (n\_neighbors и min\_dist) и проанализируйте, как это влияет на результаты.
12. *Давайте самостоятельно увидим, что снижение размерности – это группа методов, славящаяся своей неустойчивостью.* Пермутируйте 50% и 100% колонки *'bwt'*. Проведите PCA и UMAP анализ. Наблюдаете ли вы изменения в куммулятивном проценте объяснённой вариации PCA? В итоговом представлении данных на биплотах для PCA? Отличается ли визуализация данных?
13. *Давайте проведем анализ чувствительности.* Проведите анализ, как в шагах 4-6 для оригинального с удалением всех строк с пустыми значениями (т.е. включая колонки с количеством пропущенных значений больше 100), а затем для оригинального датафрейма с импутированием пустых значений средним или медианой. Как отличаются получившиеся результаты? В чем преимущества и недостатки каждого подхода?
14. *Давайте проведем анализ чувствительности.* Сделайте то же, что в пункте 14, но для методов снижения размерности – PCA и UMAP. Проанализируйте результаты.

1. Есть два способа сделать это: [первый](https://stackoverflow.com/questions/39814916/how-can-i-see-output-of-rmd-in-github/39816334#39816334) лёгкий и не совсем корректный (но результат будет правильным), второй сложнее, зато поможет вам понять, как выстроить весь цикл работы в репозитории (детали хорошо объяснены в [этом видео](https://www.youtube.com/watch?v=MdmnE3AnkQE&feature=youtu.be) (спасибо Екатерине Фокиной за находку) ). Во втором случае общая идея в том, что вы создаете и клонируете свой репозиторий, а потом настраиваете R, чтобы делать коммиты удобнее).   
    [↑](#footnote-ref-0)
2. Источник: <https://hbiostat.org/data/repo/vlbw> [↑](#footnote-ref-1)