Техническое задание на курсовую работу студентов 1-го года магистратуры МФТИ по курсу «Восстановление зависимостей с использованием эмпирических данных».

Составил: доктор технических наук, ведущий научный сотрудник Борисов Н.М.

**Вариант №1. Анализ больших данных текущей пандемии ковид-19.**

Материалы и методы: динамика заболеваемости и смертности пандемии ковид-19 на сайте <https://www.worldometers.info/coronavirus/>,

1. Оценить достоверность данных по разным странам и регионам, используя, например, критерий Пирсона (хи-квадрат).
2. На данных, которые прошли тест валидности, оценить (методом максимального правдоподобия) наиболее вероятные значения параметров в модели Кермака – МакКендрика (модель S-R-R, 1927 г.) которую использовала ВОЗ для рекомендаций для введения противоэпидемических карантинов.
3. Для разных стран, которые вводили карантины разной степени строгости, оценить достоверность прогнозов ВОЗ по развитию пандемии, на которых были основаны такие рекомендации, как карантины, масочный режим и т.д.

**Вариант №2. Сравнение эффективности оптимизации «по Дарвину» и «по Ламарку» для методов Монте-Карло моделирования случайных блужданий**.

Моделирование методом Монте-Карло случайных блужданий во многих случаях требует значительных методов оптимизации.

Рассмотрим кинетическое уравнение для таких блужданий частицы в рассеивающей среде,

где – плотность столкновений частиц на i-м акте рассеяния, – функция-ядро переноса частицы из точки **x′** фазового пространства в точку ***x***.

Тогда *оптимизацией блужданий по Дарвину* называется на каждой итерации процедура рулетки-расщепления каждой истории случайного блуждания пропорционально отношению фитнес-функций (функций приспособленности) *r*=. Если *r* < 1, то с вероятностью 1- *r* частица гибнет, а с вероятностью *r* – выживает («русская рулетка»). Если *r* > 1, то частица размножается на [*r*] потомков, за каждым из которых мы следим в моделировании (расщепление). Подробнее о приеме «рулетка/расщепление» см. в главе 2 диссертации (Борисов, 1999).

*Оптимизация блужданий по Ламарку* использует, в отличие от оптимизации по Дарвину, опыт предыдущих итерацией:

Одной из реализаций оптимизации по Ламарку является *алгоритм муравейника*. <https://ru.wikipedia.org/wiki/%D0%9C%D1%83%D1%80%D0%B0%D0%B2%D1%8C%D0%B8%D0%BD%D1%8B%D0%B9_%D0%B0%D0%BB%D0%B3%D0%BE%D1%80%D0%B8%D1%82%D0%BC#:~:text=%D0%9C%D1%83%D1%80%D0%B0%D0%B2%D1%8C%D0%B8%D0%BD%D1%8B%D0%B9%20%D0%B0%D0%BB%D0%B3%D0%BE%D1%80%D0%B8%D1%82%D0%BC%20(%D0%B0%D0%BB%D0%B3%D0%BE%D1%80%D0%B8%D1%82%D0%BC%20%D0%BE%D0%BF%D1%82%D0%B8%D0%BC%D0%B8%D0%B7%D0%B0%D1%86%D0%B8%D0%B8%20%D0%BF%D0%BE%D0%B4%D1%80%D0%B0%D0%B6%D0%B0%D0%BD%D0%B8%D0%B5%D0%BC,%D0%B7%D0%B0%D0%B4%D0%B0%D1%87%20%D0%BF%D0%BE%D0%B8%D1%81%D0%BA%D0%B0%20%D0%BC%D0%B0%D1%80%D1%88%D1%80%D1%83%D1%82%D0%BE%D0%B2%20%D0%BD%D0%B0%20%D0%B3%D1%80%D0%B0%D1%84%D0%B0%D1%85>.

**Задача:** сравнить эффективности оптимизации по Дарвину (рулетка-расщепление) с оптимизацией по Ламарку (алгоритм муравейника) для случайных блужданий в двух классах задач:

1. Дискретные блуждания по графу в 1024 вершинами и рандомными направленными ребрами. Мера близости – количество переходов из начала (вершина А) в конец (вершина B).
2. Блуждание моноэнергетической частицы в трехмерном рассеивающем пространстве. Область источника (А) поглотителя (B) разделены 10-20 длинами свободного пробега. Диаметр обеих областей – 0.5 длины свободного пробега. Рассеивание происходит по закону обратного синуса половинного угла; ср. ф-лу 2.18 на стр. 42 диссертации (Борисов, 1999).

**Вариант 3: исследование метода индивидуальной фильтрации генно-экспрессионных признаков в задачах предсказания методами машинного обучения ответа онкологических больных на химиотерапию.**

Материалы и методы: недавно был предложен (Tkachev et al., 2019, 2020) метод FLOating Window Projective Separator (FloWPS) для индивидуальной фильтрации (flexible data trimming) генно-экспрессионных признаков при предсказании методами машинного обучения ответа онкологических пациентов на химотерапию. Данный метод предотвращает экстраполяцию в пространстве признаков и повышает точность предсказаний для глобальных методов машинного обучения, таких как метод опорных векторов, случайного леса, биномиальный байесовский метод и многослойный перцептрон.

Теперь усложним задачу. Пускай к нам поступает неизвезстный образец, и нам надо его классифицировать – как ответчика либо неответчика. Для классификации мы можем использовать либо top 30 маркерных признаков, либо ядро маркерных признаков, вычисленное без неизвестного образца. Устроим теперь скользящий контроль, назначая в рамках такой процедуры каждый образец по очереди неизвестным и классифицируя его.

Результаты такой leave-one-out cross-validation для набора данных TARGET-50(Goldman et al., 2015; Walz et al., 2015) показали, что в случае отбора top 30 маркерных признаков получается garbage c AUC порядка 0.5, а в случае отбора ядра маркерных признаков, вычисленного без неизвестного образца, получается осмысленный результат: AUC = 0.69 для классического SVM AUC = 0.78 для FloWPS с p =0.90.

*Задача 1:* провести тест для оставшихся наборов данных (Borisov et al., 2020) и методов опорных векторов, случайного леса, биномиального наивного Байеса и многослойного перцептрона. Фактически это полный цикл кросс-валидации для процедуры отбора маркерных признаков.

*Задача 2:* оптимизировать число T в схеме «top T признаков», пересечение которых дает ядро маркерных признаков.

**Вариант №4: исследование и оптимизация метода универсальной гармонизации генно-экспрессионных профилей.**

В биоинформатике актуальной является проблема гармонизации и сравнения генно-экспрессионных профилей, поученных на разных установках, использующих методы высокопроизводительного секвенирования мРНК и микрочиповой гибридизации мРНК.

Недавно нами был предложен метод Shambhala (Borisov et al., 2019), опирающийся на более ранний метод XPN (Shabalin et al., 2008).

Метод Shambhala зависит от двух массивов профилей – промежуточного (P) и дефинитивного (Q).

В качестве критерия качества гармонизации будем использовать две метрики – 1) watermelon multisection metric, описанную в работе (Zolotovskaia et al., 2020): WM-метрика кластерной дендрограммы для деления по типу биологических образцов должна быть больше, чем WM-метрика кластерной дендрограммы для деления по типу экспериментальной платформы; 2) метод анализа однородности на основе теста Стьюдента (Aliev and Borisov, 2018): однородной внутри одинаковых типов образцов должна быть больше, чем внутри одинаковой платформы после гармонизации.

Исследование провести для семи типов промежуточных массивов профилей и двух типов дефинитивных массивов, попарно для каждого типа экспериментальных установок (задача 1), а также для всего массива профилей вместе (задача 2).

**Литература**

Aliev, R. O., and Borisov, N. M. (2018). The Method for Analysis of Expression Data Homogeneity Based On the Student Test. *Math. Biol. Bioinforma.* 13, 50–67. doi:10.17537/2018.13.50.

Borisov, N., Shabalina, I., Tkachev, V., Sorokin, M., Garazha, A., Pulin, A., et al. (2019). Shambhala: a platform-agnostic data harmonizer for gene expression data. *BMC Bioinformatics* 20, 66. doi:10.1186/s12859-019-2641-8.

Borisov, N., Sorokin, M., Tkachev, V., Garazha, A., and Buzdin, A. (2020). Cancer gene expression profiles associated with clinical outcomes to chemotherapy treatments. *BMC Med. Genomics* 13. doi:10.1186/s12920-020-00759-0.

Shabalin, A. A., Tjelmeland, H., Fan, C., Perou, C. M., and Nobel, A. B. (2008). Merging two gene-expression studies via cross-platform normalization. *Bioinformatics* 24, 1154–1160. doi:10.1093/bioinformatics/btn083.

Tkachev, V., Sorokin, M., Borisov, C., Garazha, A., Buzdin, A., and Borisov, N. (2020). Flexible Data Trimming Improves Performance of Global Machine Learning Methods in Omics-Based Personalized Oncology. *Int. J. Mol. Sci.* 21, 713. doi:10.3390/ijms21030713.

Tkachev, V., Sorokin, M., Mescheryakov, A., Simonov, A., Garazha, A., Buzdin, A., et al. (2019). FLOating-Window Projective Separator (FloWPS): A Data Trimming Tool for Support Vector Machines (SVM) to Improve Robustness of the Classifier. *Front. Genet.* 9, 717. doi:10.3389/fgene.2018.00717.

Zolotovskaia, M. A., Sorokin, M. I., Petrov, I. V., Poddubskaya, E. V., Moiseev, A. A., Sekacheva, M. I., et al. (2020). Disparity between Inter-Patient Molecular Heterogeneity and Repertoires of Target Drugs Used for Different Types of Cancer in Clinical Oncology. *Int. J. Mol. Sci.* 21, 1580. doi:10.3390/ijms21051580.