# Практическая работа № 4

# «Модельные методы рекомендательных систем. Метод главных компонент»

по дисциплине «Разработка обеспечивающих подсистем систем поддержки принятия решений»

**Цели:** приобрести навыки реализации модельных коллаборативных методов рекомендательных систем, основанных на поиске скрытых факторов методом главных компонент.

**Задачи:**

1) создать программную реализацию модельного метода рекомендательной системы (РС), основанную на поиске главных компонент, включающую:

– актуальную предметную область для применения РС (вроде маркетплейса, медиа ресурсов, соц. сетей, экономической сферы и т.д.) и набор начальных данных для неё (опрос покупателей, статистка за временной период и т.д., включающий набор признаков, наиболее важных и наиболее коррелирующих друг с другом на ваш взгляд см. Примечание 1);

– по возможности отцентрировать выборку (для минимизации потерь при дальнейших расчётах, так как при таком сдвиге средние значения признаков будут стремиться к нулю), при большой разнице порядков признаков необходимо выполнить их стандартизацию или масштабирование, т.к. дисперсия величин может получить слишком большой или слишком малый размер, из-за чего понижение размерности и проецирование может получится некорректным и иметь слишком малую точность при восстановлении первоначальных данных;

– на основе модифицированной выборки построить ковариационную матрицу признаков, найти её собственные векторы и рассчитать её собственные значения (см. Примечание 2), определив новый базис для начальной выборки;

– выполнить снижение размерности по новому базису для начальных характеристик путём проекции на каждый из полученных собственных векторов, с помощью восстановления начальных данных и проверки полученного результата восстановления с настоящими начальными данными выбрать вектор, проекция на который имеет наименьшую функцию потерь (см. Примечание 3).

2) в качестве дополнительного задания (на доп. баллы) выполнить:

– реализовать центроидный метод и сравнить полученные результаты с результатами метода главных компонент (см. Примечание 4);

– полученные вышеописанными методами результаты (с составлением новых матриц) подвергнуть анализу анамнестических методов, марковских цепей или сингулярного разложения (однако стоит учитывать, что при некоторых условиях сам метод главных компонент превращается в задачу сингулярного разложения).

ПРИМЕЧАНИЕ:

1. В задачах анализа данных стараются избегать сильно коррелирующих признаков, так как они дают практически одинаковую информацию, из-за чего по сути тратится лишнее время на анализ «того же самого», аналогичная ситуации возникает при обучении моделей (в том числе и нейронных сетей) на коррелирующих данных, что может вызывать переобучение на входных данных, снижение точности результатов при дообучении модели и т.д., однако в задачах понижения размерности и поиска главных компонент в целом сильная корреляция данных наоборот позволяет привести несколько однотипных данных пространства большей размерности к комбинации этих признаков меньшей размерностей, что как раз таки и является основной задачей в настоящей работе.

2. В методе главных компонент направление максимальной дисперсии у проекции характеристик всегда совпадает с собственным вектором, имеющим максимальное собственное значение, которое равно величине этой

дисперсии (количество собственных векторов и собственных значений равны размеру матрицы, при этом значения могут повторяться).

3. Количество собственных векторов совпадает с размерностью ковариационной матрицы, что позволяет провести проекцию выборки на несколько разных векторов, что достаточно полезно, так как при разных начальных условиях (количестве точек и их взаиморасположении) направление максимизации дисперсии может быть не так очевидно (особенно, если точки формируют облако больше похожее на окружность, чем на эллипс), поэтому проекцию лучше всего делать на несколько векторов с последующим сравнением результатов восстановления данных.

4. Центроидный метод в целом при больших объёмах данных даёт результаты хуже, чем метод главных компонент, однако он проще в реализации и работает достаточно быстро, а первые выделенные главные факторы, как правило, совпадают, поэтому для достаточно простых задач и при небольшом количестве данных разница между применением этих методов не должна быть так сильно видна.