Программа учебной дисциплины «Анализ данных»

Утверждена
Академическим руководителем

_____ Н.В. Асеева

_____ 20_____

Автор	Маслова Е.А.
Число кредитов	8
Контактная работа (час.)	32
Самостоятельная работа (час.)	272
Курс	2
Формат изучения дисциплины	без использования онлайн курса

I. ЦЕЛЬ, РЕЗУЛЬТАТЫ ОСВОЕНИЯ ДИСЦИПЛИНЫ И ПРЕРЕКВИЗИТЫ

Целями освоения дисциплины «Анализ данных» является знакомство с основными понятиями анализа данных, развитие навыков анализа данных, овладение основными алгоритмами анализа данных.

В результате освоения дисциплины студент должен:

знать:

- основные понятия анализа данных;

уметь:

- анализировать данные, выбирать адекватные методы анализа;

владеть:

- навыками применения основных алгоритмов анализа данных.

Изучение дисциплины базируется на следующих дисциплинах:

- Математический анализ;
- Геометрия и алгебра;
- Дискретная математика;

Основные положения дисциплины должны быть использованы в дальнейшем при изучении следующих дисциплин:

1 Теория принятия решений.

II. СОДЕРЖАНИЕ УЧЕБНОЙ ДИСЦИПЛИНЫ

Тема 1. Основные задачи анализа данных.

Данные. Типы данных. Анализ данных. Классические задачи анализа данных: снижение размерности, кластеризация, классификация.

Тема 2. Методы снижения размерности. Сингулярное разложение и метод главных компонент.

Проблема уменьшения размерности. Задачи наилучшей аппроксимации матрицы заданной размерности матрицами той же размерности фиксированного ранга. Выбор матричной нормы.

SVD разложение. Сингулярные числа матрицы. Теорема Шмидта – Мирского (Эккарта-Юнга). Оценка погрешности в l^2 матричной норме и в норме Фробениуса.

Метод главных компонент как вариант SVD разложения. Вычисление главных компонент. QR - алгоритм. Матрица нагрузок как матрица корреляций наблюдений и новых факторов. Погрешность аппроксимации как изменение общей вариации данных.

Проблема выбора числа главных компонент. Анализ вариации в методе главных компонент. Интерпретация главных компонент. Поиск структуры в матрице нагрузок.

Тема 3. Методы кластеризации.

Проблема разбиения объектов на группы по степени близости объектов в группах. Расстояния в пространстве объектов. Расстояния между кластерами: метод ближайшего соседа; метод наиболее удаленных соседей; метод попарных средних; метод взвешенных попарных средних; центроидный метод; взвешенный центроидный метод; метод Варда. Таблица расстояний между объектами.

Алгоритмы иерархической кластеризации (снизу вверх и сверху вниз). Проблема выбора адекватного числа кластеров. Графическое представление иерархических алгоритмов кластеризации. Дендрограмы.

Функционал качества кластеризации W(S;c). Задача кластеризации как задача дискретно-непрерывной оптимизации (разбиения и центры). Алгоритм к-средних. Два этапа каждого шага алгоритма: разбиение по принципу ближайшего центра, вычисление центра группы.

Достоинства и недостатки алгоритма к-средних. Особенности алгоритма для различных расстояний. Пошаговый алгоритм к-средних (incremental k-means). Проблема останова.

Алгоритм PAM (partition around medoids). Сравнение с алгоритмом к-средних.

Генетические алгоритмы кластеризации. Основные понятия и детали алгоритма.

Тема 4. Методы классификации

Постановка задачи классификации, простейшие классификаторы (NN-классификатор, ближний сосед, k-NN классификатор). Оценка качества классификаторов: обучающая выборка, тестовая выборка, ошибки 0-1 классификатора (TP true positive, FP false positive, TN true negative, FN false negative), таблица ошибок.

Основное тождество вариаций (одномерный случай): Общая вариация=внутригрупповая вариация+межгрупповая вариация. Отношение Фишера, как мера возможности разделения данных на группы (возможность классификации). Проекция данных на одномерное подпространство. Вычисление отношения Фишера в проекции.

Задача о «наилучшей» проекции (наилучшая возможность разделения данных после проекции). Максимальное значение отношения двух квадратичных форм. Обобщенная задача на собственные значения. Дискриминантная функция Фишера. Классификатор на основе дискриминантной функции Фишера.

Классификация с двумя классами (да и нет). Линейные классификаторы. Линейная регрессия по методу наименьших квадратов. Логистическая регрессия. Правило классификации по методу линейной регрессии. Дискриминантное правило Фишера, как частный случай классификации по методу линейной регрессии.

Метод опорных векторов (SVM), как линейный классификатор. Правило классификации по методу опорных векторов. Задача оптимизации для поиска разделяющей гиперплоскости. Вычисление опорных векторов.

Деревья решений. Описание общего подхода. Классификаторы линейной регрессии, Фишера и метода опорных векторов как простейшие деревья решений. Достоинства и недостатки деревьев решений в сравнении с линейными классификаторами.

Правила разделения, основанные на значениях одного признака (случай дискретных значений признака). Индекс Джини и энтропия конкретного значения признака, как меры способности этого значения порождать хорошую классификацию. Индекс Джини и энтропия всего признака. Построение дерева решений по правилу разбиения по одному из признаков (дискретный набор значений). Выбор признака для разбиения. Критерий останова (стоп критерий). Случай, когда признак принимает любые значения из некоторого интервала. Дискретизация. Построение дерева решений.

III. ОЦЕНИВАНИЕ

Контроль знаний студентов включает формы текущего и итогового контроля. Текущий контроль оценивается по 10-балльной шкале. По результатам текущего контроля организуются индивидуальные консультации в рамках второй половины рабочего дня преподавателя. Форма итогового контроля – экзамен. Формы итогового контроля оцениваются также по 10-балльной шкале.

Лабораторная работа:

оценка в 10 баллов проставляется в исключительных случаях самостоятельно проведенной работы, результаты которой могут в дальнейшем использоваться в учебном процессе или в исследовательской работе студента;

оценка в 8-9 баллов проставляется при самостоятельно разработанном или удачно адаптированном и отлично представленном исследовании по выбранной тематике;

оценка в 6-7 баллов проставляется при своевременно выполненном и самостоятельно представленном исследовании по выбранной тематике;

оценка в 4-5 баллов проставляется при частичном, несамостоятельном участии в выполнении работ над заданием;

оценка в 2-3 балла проставляется, когда студент не может самостоятельно представить работу или когда работа носит явные признаки заимствований (работу предлагается переделать);

оценка в 1 балл проставляется при наличии каких-либо демонстративных проявлений безграмотности и неэтичного отношения к работе.

Экзамен:

На экзамене, представляющем собой письменные ответы на вопросы и решение задачи с последующим собеседованием, оценка проставляется следующим образом:

высшая оценка в 9 баллов (10 баллов только в исключительных случаях) проставляется при отличном выполнении заданий (полных, с примерами и возможными обобщениями ответах на вопросы, при правильном решении задачи и детальном ее представлении);

почти отличная оценка в 8 баллов проставляется при полностью правильных ответах на вопросы и решении задачи, но при отсутствии примеров и обобщений, а также детального представления решаемой задачи;

оценка в 7 баллов проставляется при правильных ответах на вопросы и правильном решении задачи, но при отсутствии пояснений и обобщений, а также детального представления решаемой задачи;

оценка в 6 баллов проставляется при наличии отдельных неточностей в ответах на вопросы или неточностях в решении задачи непринципиального характера (описки и случайные ошибки);

оценка в 4-5 баллов проставляется в случаях, когда в ответах на вопросы и в решении задачи имеются существенные неточности и ошибки, свидетельствующие о недостаточном понимании изучаемой дисциплины;

оценка в 2-3 балла проставляется при наличии лишь отдельных положительных моментов в ответах на вопросы и в решении задачи;

оценка в 1 балл проставляется в тех случаях, когда наряду с неправильными ответами на вопросы и решением задачи имеют место какие-либо демонстративные проявления безграмотности или неэтичное отношение к изучаемой дисциплине.

По результатам устного собеседования с преподавателем возможны корректировки оценки в ту или иную сторону.

Результирующая оценка $О_{результ}$ учитывает итоговую накопленную оценку $О_{итоговая\ накопл.}$ и оценку, полученную за экзаменационную работу, и вычисляется по формуле

$$О_{результ} = (8/10)* О_{итог.накопл} + (2/10)*О_{экз},$$

где $О_{итоговая\ накопл.} = (5/8)*О_{практ.} + (3/8)*О_{теор}$,

$$О_{практ} = (1/3)(О_{лаб1} + О_{лаб2} + О_{лаб3}),\ О_{теор} = (1/3)(О_{КР1} + О_{КР2} + О_{КР3}),$$

Способ округления оценок – арифметический.

IV. ПРИМЕРЫ ОЦЕНОЧНЫХ СРЕДСТВ

Оценочные средства для текущего контроля студента

Примеры заданий экзамена

1. В SVD разложении матрицы данных *X* размера *(50* x *5)* матрица сингулярных чисел *S* размера *(50* x *5)* имеет вид *S= diag(4.0; 3.0; 2.0; 1.0; 0.0).* Вычислите ошибку наилучшей аппроксимации матрицы *X* матрицами ранга 2 в норме Фробениуса и в евклидовой норме.

2. При применении метода главных компонент к преобразованной (центрированной и нормированной) матрице данных *Z* размера *(40* x *5)* получены следующие результаты: квадраты сингулярных чисел (Таблица 1) и коэффициенты разложений столбцов матрицы *Z* по скрытым факторам (Таблица 2).

- сколько главных факторов достаточно? Обоснуйте ответ (например, с помощью доли объясненной вариации данных или как то иначе).

- запишите разложение столбцов матрицы Z (аппроксимация) по выбранным главным факторам.

- Поясните, какие исходные признаки объясняет первый главный фактор.

3. Представлена матрица расстояний и дендрограмма применения объединительного (agglomerative) иерархического алгоритма к набору из 10 объектов.

- Сколько, по вашему мнению, кластеров в наборе (обсудите разные варианты)?

- По какому методу объединялись группы объектов: метод ближнего соседа или метод дальнего соседа?

4. Разделяющий (divisive) алгоритм иерархической кластеризации на каждом шаге разбивает на две части одну из групп объектов, так чтобы расстояние между этими двумя частями было

наибольшим из всех возможных. Поясните, почему, по вашему мнению, разделяющий алгоритм применяется гораздо реже объединительного.

5. Объясните, почему для пошаговой версии алгоритма к-средних (incremental k-means) нельзя гарантировать остановки алгоритма за конечное число шагов, в отличии от пакетной версии алгоритма к-средних (batch version).

6. Некоторый классификатор, построенный по обучающей выборке, показал на тестовой выборке следующие результаты

	True Positive	False Positive	False Negative	True Negative
Результат	70	10	10	10

- Укажите долю правильно классифицированных объектов. Сколько в тестовой выборке содержится объектов класса «Yes» и класса «No»?

, сингулярное разложение и метод главных компонент.

- объясните, почему, несмотря на высокую долю правильно классифицированных объектов, этот классификатор нельзя считать хорошим? Приведите пример результата «хорошего» на ваш взгляд классификатора для этой тестовой выборки.

V. РЕСУРСЫ

5.1 Основная литература

1. Статистический анализ данных, моделирование и исследование вероятностных закономерностей. Компьютерный подход [Электронный ресурс] / Б.Ю.Лемешко, С.Б.Лемешко, С.Н.Постовалов, Е.В.Чимитова; ЭБС Znanium. - М.: НИЦ ИНФРА-М, 2015. - 890 с. – Режим доступа: http://znanium.com/bookread2.php?book=515227. – Загл. с экрана.
2. Статистические методы анализа данных [Электронный ресурс]: учебник / Л.И. Ниворожкина, С.В. Арженовский, А.А. Рудяга [и др.]; под общ. ред. д-ра экон. наук, проф. Л.И. Ниворожкиной; ЭБС Znanium. — М.: РИОР: ИНФРА-М, 2016. — 333 с. — (Высшее образование: Бакалавриат). – Режим доступа: http://znanium.com/bookread2.php?book=556760. – Загл. с экрана.

5.2 Дополнительная литература

1. Лагутин, М.Б. Наглядная математическая статистика: учебное пособие / М.Б.Лагутин. - 3-е изд; испр. - М.: БИНОМ, Лаборатория знаний, 2013. - 472 с.
2. Ратникова, Т.А. Анализ панельных данных и данных о длительности состояний: учебное пособие / Т.А.Ратникова, К.К.Фурманов; Нац. исслед. ун-т Высшая школа экономики. - М.: Изд. дом ВШЭ, 2014. - 373 с.
3. Крыштановский, А.О. Анализ социологических данных с помощью пакета SPSS: учебное пособие / А.О.Крыштановский. - М.: Изд. дом ГУ ВШЭ, 2006. - 281 с. - (Учебники Высшей школы экономики). Гриф МО РФ
4. Паклин, Н.Б. Бизнес-аналитика: от данных к знаниям / Н.Б.Паклин, В.И.Орешков. - СПб.: Питер, 2009. - 624 с. + 1опт. диск (CD-ROM): аналитическая платформа Deductor Academic.
5. Кацко, И.А. Практикум по анализу данных на компьютере: учебно-практическое пособие / И.А.Кацко, Н.Б.Паклин; под ред. проф. Г.В.Гореловой. - М.: КолосС, 2009. - 278 с. Гриф МО РФ
6. Malhotra, N.K. Basic Marketing Research / N.K.Malhotra. - 4th ed. - Edinbugrh: Pearson Education Limited, 2014. - 670 p.

5.3 Программное обеспечение

№	Наименование	Условия доступа

п/п		
1.	Microsoft Office 2013 Prof +	*Государственный контракт*
2.	Stata/SE	*Из внутренней сети университета (договор)*

5.4 Профессиональные базы данных, информационные справочные системы, интернет-ресурсы (электронные образовательные ресурсы)

№ п/п	Наименование	Условия доступа
	Профессиональные базы данных, информационно-справочные системы	
2.	Электронно-библиотечная система Юрайт	URL: https://biblio-online.ru/

5.5 Материально-техническое обеспечение дисциплины

Учебные аудитории для лекционных занятий по дисциплине обеспечивают использование и демонстрацию тематических иллюстраций, соответствующих программе дисциплины в составе:

– ПЭВМ с доступом в Интернет (операционная система, офисные программы, антивирусные программы);

– мультимедийный проектор с дистанционным управлением.

Учебные аудитории для лабораторных и самостоятельных занятий по дисциплине оснащены оборудованием, с возможностью подключения к сети Интернет и доступом к электронной информационно-образовательной среде НИУ ВШЭ.