Приложение 1 к РПД Анализ данных и машинное обучение
09.03.02 Информационные системы и технологии
Направленность (профиль) – Программно-аппаратные комплексы
Форма обучения – очная
Год набора - 2020

МЕТОДИЧЕСКИЕ УКАЗАНИЯ ДЛЯ ОБУЧАЮЩИХСЯ ПО ОСВОЕНИЮ ДИСЦИПЛИНЫ (МОДУЛЯ)

1.	Кафедра	Информатики и вычислительной техники
2.	Направление подготовки	09.03.02 Информационные системы и технологии
3.	Направленность (профиль)	Программно-аппаратные комплексы
4.	Дисциплина (модуль)	Анализ данных и машинное обучение
5.	Форма обучения	очная
6.	Год набора	2020

1. Методические рекомендации

Приступая к изучению дисциплины, обучающемуся необходимо внимательно ознакомиться с тематическим планом занятий, списком рекомендованной литературы. Следует уяснить последовательность выполнения индивидуальных учебных заданий. Самостоятельная работа обучающегося предполагает работу с научной и учебной литературой, умение создавать тексты. Уровень и глубина усвоения дисциплины зависят от активной и систематической работы на лекциях, изучения рекомендованной литературы, выполнения контрольных письменных заданий.

При изучении дисциплины обучающиеся выполняют следующие задания:

- изучают рекомендованную научно-практическую и учебную литературу;
- выполняют задания, предусмотренные для самостоятельной работы.

Основными видами аудиторной работы обучающихся являются лекции и лабораторные занятия.

Каждый обучающийся перед началом занятий записывается преподавателем на электронный курс по данному предмету, к которому можно получить доступ через сеть Интернет. Курс поддерживается системой дистанционного обучения *moodle* (модульная объектно-ориентированная динамическая учебная среда), к которой может получить доступ зарегистрированный пользователь через сеть Интернет. Адрес курса в системе *moodle* МАГУ: http://moodle.arcticsu.ru/course/view.php?id=46 [1].

В рамках данного курса в системе *moodle*, организовано:

- взаимодействие обучающихся между собой и с преподавателем: для чего используются форумы и чаты.
- передача знаний в электронном виде: с помощью файлов, архивов, веб-страниц, лекций.
- проверка знаний и обучение с помощью тестов и заданий: результаты работы обучающиеся могут отправлять в текстовом виде или в виде файлов.
- совместная учебная и исследовательская работа обучающихся по определенной теме: с помощью встроенных механизмов: семинаров, форумов и пр.

1 Для получения доступа к курсу необходима регистрация в системе и запись на курс.

- журнал оценок: в котором учитывается успеваемость обучающихся по балльной системе.

Таким образом, самостоятельная работа обучающегося организуется через систему дистанционного обучения *moodle* МАГУ. Так же данная система используется преподавателем и в процессе проведения аудиторных занятий, для: проведения тестов, предоставления презентаций лекций и методических рекомендаций к выполнению лабораторных работ, учета успеваемости учащихся.

1.1. Методические рекомендации по организации работы обучающихся во время проведения лекционных занятий

В ходе лекций преподаватель излагает и разъясняет основные, наиболее сложные понятия темы, а также связанные с ней теоретические и практические проблемы, дает рекомендации на семинарское занятие и указания на самостоятельную работу.

Знакомство с дисциплиной происходит уже на первой лекции, где от обучающегося требуется не просто внимание, но и самостоятельное оформление конспекта. При работе с конспектом лекций необходимо учитывать тот фактор, что одни лекции дают ответы на конкретные вопросы темы, другие – лишь выявляют взаимосвязи между явлениями, помогая обучающемуся понять глубинные процессы развития изучаемого предмета как в истории, так и в настоящее время.

Конспектирование лекций – сложный вид вузовской аудиторной работы, предполагающий интенсивную умственную деятельность обучающегося. Конспект является полезным тогда, когда записано самое существенное и сделано это самим обучающимся. Не надо стремиться записать дословно всю лекцию. Такое «конспектирование» приносит больше вреда, чем пользы. Целесообразно вначале понять основную мысль, излагаемую лектором, а затем записать ее. Желательно запись осуществлять на одной странице листа или оставляя поля, на которых позднее, при самостоятельной работе с конспектом, можно сделать дополнительные записи, отметить непонятные места.

Конспект лекции лучше подразделять на пункты, соблюдая красную строку. Этому в большой степени будут способствовать вопросы плана лекции, предложенные преподавателям. Следует обращать внимание на акценты, выводы, которые делает лектор, отмечая наиболее важные моменты в лекционном материале замечаниями «важно», «хорошо запомнить» и т.п. Можно делать это и с помощью разноцветных маркеров или ручек, подчеркивая термины и определения.

Целесообразно разработать собственную систему сокращений, аббревиатур и символов. Однако при дальнейшей работе с конспектом символы лучше заменить обычными словами для быстрого зрительного восприятия текста.

Работая над конспектом лекций, всегда необходимо использовать не только учебник, но и ту литературу, которую дополнительно рекомендовал лектор. Именно такая серьезная, кропотливая работа с лекционным материалом позволит глубоко овладеть теоретическим материалом.

Электронные конспекты презентаций лекций доступны для просмотра и скачивания обучающимся в электронной образовательной среде *moodle* МАГУ на странице курса: «Анализ данных и машинное обучение».

1.2. Методические рекомендации по подготовке к практическим занятиям (лабораторным /семинарам)

Подготовку к каждому практическому занятию обучающийся должен начать с ознакомления с его планом, отражающим содержание предложенной темы. Продумывание и изучение вопросов плана основывается на проработке текущего материала лекции, и изучения

рекомендованной обязательной и дополнительной литературы. Новые понятия по изучаемой теме необходимо проработать и внести в глоссарий.

Результат такой работы должен проявиться в способности обучающегося свободно ответить на теоретические вопросы практикума, его выступлении и участии в коллективном обсуждении вопросов изучаемой темы, правильном выполнении лабораторных заданий и контрольных работ.

В процессе подготовки к практическим занятиям, обучающимся необходимо обратить особое внимание на самостоятельное изучение рекомендованной литературы. При всей полноте конспектирования лекции в ней невозможно изложить весь материал из-за лимита аудиторных часов. Поэтому самостоятельная работа с учебниками, учебными пособиями, научной, справочной литературой, материалами периодических изданий и Интернета является наиболее эффективным методом получения дополнительных знаний, позволяет значительно активизировать процесс овладения информацией, способствует более глубокому усвоению изучаемого материала, формирует у обучающихся свое отношение к конкретной проблеме.

Лабораторные занятия завершают изучение наиболее важных тем учебной дисциплины. Они служат для закрепления изученного материала, развития умений и навыков подготовки докладов, сообщений, приобретения опыта устных публичных выступлений, ведения дискуссии, аргументации и защиты выдвигаемых положений, а также для контроля преподавателем степени подготовленности обучающихся по изучаемой дисциплине. На лабораторных занятиях обучающиеся совместно с преподавателем обсуждают выданные им задания, задают интересующие их вопросы и выполняют на компьютерах самостоятельно или в группах свои задания, используя программное обеспечение представленное в рабочей программе. Каждое выполненное задание обучающийся обязан оформить в виде отчета и защитить его. Методические рекомендации к лабораторным заданиям доступны для обучающегося в системе *moodle* МАГУ на сайте курса: «Анализ данных и машинное обучение».

Семинар предполагает свободный обмен мнениями по избранной тематике. Он начинается со вступительного слова преподавателя, формулирующего цель занятия и характеризующего его основную проблематику. Затем, как правило, заслушиваются сообщения обучающихся. Обсуждение сообщения совмещается с рассмотрением намеченных вопросов. Сообщения, предполагающие анализ публикаций по отдельным вопросам семинара, заслушиваются обычно в середине занятия. Поощряется выдвижение и обсуждение альтернативных мнений. В заключительном слове преподаватель подводит итоги обсуждения и объявляет оценки выступавшим обучающимся. В целях контроля подготовленности обучающихся и привития им навыков краткого письменного изложения своих мыслей преподаватель в ходе семинарских занятий может осуществлять текущий контроль знаний в виде тестовых заданий.

При подготовке к семинару обучающиеся имеют возможность воспользоваться консультациями преподавателя. Кроме указанных тем обучающиеся вправе, по согласованию с преподавателем, избирать и другие интересующие их темы.

Качество учебной работы обучающихся преподаватель оценивает с использованием технологической карты дисциплины, размещенной на сайте филиала МАГУ.

1.3. Методические рекомендации по работе с литературой

Работу с литературой целесообразно начать с изучения общих работ по теме, а также учебников и учебных пособий. Далее рекомендуется перейти к анализу монографий и статей, рассматривающих отдельные аспекты проблем, изучаемых в рамках курса, а также официальных материалов и неопубликованных документов (научно-исследовательские работы, диссертации), в которых могут содержаться основные вопросы изучаемой проблемы.

Работу с источниками надо начинать с ознакомительного чтения, т.е. просмотреть текст, выделяя его структурные единицы. При ознакомительном чтении закладками отмечаются те страницы, которые требуют более внимательного изучения.

В зависимости от результатов ознакомительного чтения выбирается дальнейший способ работы с источником. Если для разрешения поставленной задачи требуется изучение некоторых фрагментов текста, то используется метод выборочного чтения. Если в книге нет подробного оглавления, следует обратить внимание ученика на предметные и именные указатели.

Избранные фрагменты или весь текст (если он целиком имеет отношение к теме) требуют вдумчивого, неторопливого чтения с «мысленной проработкой» материала. Такое чтение предполагает выделение: 1) главного в тексте; 2) основных аргументов; 3) выводов. Особое внимание следует обратить на то, вытекает тезис из аргументов или нет.

Необходимо также проанализировать, какие из утверждений автора носят проблематичный, гипотетический характер и уловить скрытые вопросы.

Понятно, что умение таким образом работать с текстом приходит далеко не сразу. Наилучший способ научиться выделять главное в тексте, улавливать проблематичный характер утверждений, давать оценку авторской позиции – это сравнительное чтение, в ходе которого обучающийся знакомится с различными мнениями по одному и тому же вопросу, сравнивает весомость и доказательность аргументов сторон и делает вывод о наибольшей убедительности той или иной позиции.

Если в литературе встречаются разные точки зрения по тому или иному вопросу из-за сложности прошедших событий и правовых явлений, нельзя их отвергать, не разобравшись. При наличии расхождений между авторами необходимо найти рациональное зерно у каждого из них, что позволит глубже усвоить предмет изучения и более критично оценивать изучаемые вопросы. Знакомясь с особыми позициями авторов, нужно определять их схожие суждения, аргументы, выводы, а затем сравнивать их между собой и применять из них ту, которая более убедительна.

Следующим этапом работы с литературными источниками является создание конспектов, фиксирующих основные тезисы и аргументы. Можно делать записи на отдельных листах, которые потом легко систематизировать по отдельным темам изучаемого курса. Другой способ – это ведение тематических тетрадей-конспектов по одной какой-либо теме. Большие специальные работы монографического характера целесообразно конспектировать в отдельных тетрадях. Здесь важно вспомнить, что конспекты пишутся на одной стороне листа, с полями и достаточным для исправления и ремарок межстрочным расстоянием (эти правила соблюдаются для удобства редактирования). Если в конспектах приводятся цитаты, то непременно должно быть дано указание на источник (автор, название, выходные данные, № страницы). Впоследствии эта информации может быть использована при написании текста реферата или другого задания.

Таким образом, при работе с источниками и литературой важно уметь:

- сопоставлять, сравнивать, классифицировать, группировать, систематизировать информацию в соответствии с определенной учебной задачей;
- обобщать полученную информацию, оценивать прослушанное и прочитанное;
- фиксировать основное содержание сообщений; формулировать, устно и письменно, основную идею сообщения; составлять план, формулировать тезисы;
- готовить и презентовать развернутые сообщения типа доклада;
- работать в разных режимах (индивидуально, в паре, в группе), взаимодействуя друг с другом;
- пользоваться реферативными и справочными материалами;
- контролировать свои действия и действия своих товарищей, объективно оценивать свои действия;

- обращаться за помощью, дополнительными разъяснениями к преподавателю, другим обучающимся.
- пользоваться лингвистической или контекстуальной догадкой, словарями различного характера, различного рода подсказками, опорами в тексте (ключевые слова, структура текста, предваряющая информация и др.);
- использовать при говорении и письме перифраз, синонимичные средства, слова-описания общих понятий, разъяснения, примеры, толкования, «словотворчество»;
- повторять или перефразировать реплику собеседника в подтверждении понимания его высказывания или вопроса;
- обратиться за помощью к собеседнику (уточнить вопрос, переспросить и др.);
- использовать мимику, жесты (вообще и в тех случаях, когда языковых средств не хватает для выражения тех или иных коммуникативных намерений).

1.4. Методические рекомендации по подготовке к сдаче зачета

Подготовка к зачету способствует закреплению, углублению и обобщению знаний, получаемых, в процессе обучения, а также применению их к решению практических задач. Готовясь к зачету, обучающийся ликвидирует имеющиеся пробелы в знаниях, углубляет, систематизирует и упорядочивает свои знания. На зачете обучающийся демонстрирует то, что он приобрел в процессе изучения дисциплины.

В условиях применяемой в МАГУ балльно-рейтинговой системы подготовка к зачету включает в себя самостоятельную и аудиторную работу обучающегося в течение всего периода изучения дисциплины и непосредственную подготовку в дни, предшествующие экзамену по разделам и темам дисциплины.

При подготовке к зачету обучающимся целесообразно использовать не только материалы лекций, а и рекомендованные основную и дополнительную литературу.

При подготовке к промежуточной аттестации целесообразно:

- внимательно изучить перечень вопросов и определить, в каких источниках находятся сведения, необходимые для ответа на них;
- внимательно прочитать рекомендованную литературу;
- составить краткие конспекты ответов (планы ответов).

Качество учебной работы обучающихся преподаватель оценивает с использованием технологической карты дисциплины, размещенной на сайте филиала МАГУ.

1.5. Методические рекомендации по созданию презентации

Алгоритм создания презентации:

1 этап – определение цели презентации

2 этап – подробное раскрытие информации,

3 этап − основные тезисы, выводы.

Следует использовать 10-15 слайдов. При этом:

- первый слайд – титульный. Предназначен для размещения названия презентации, имени докладчика и его контактной информации;
- на втором слайде необходимо разместить содержание презентации, а также краткое описание основных вопросов;
- оставшиеся слайды имеют информативный характер.

Обычно подача информации осуществляется по плану: тезис – аргументация – вывод.

Требования к оформлению и представлению презентации:

1. Читабельность (видимость из самых дальних уголков помещения и с различных устройств), текст должен быть набран 24-30-ым шрифтом.
2. Тщательно структурированная информация.
3. Наличие коротких и лаконичных заголовков, маркированных и нумерованных списков.
4. Каждому положению (идее) надо отвести отдельный абзац.
5. Главную идею надо выложить в первой строке абзаца.
6. Использовать табличные формы представления информации (диаграммы, схемы) для иллюстрации важнейших фактов, что даст возможность подать материал компактно и наглядно.
7. Графика должна органично дополнять текст.
8. Выступление с презентацией длится не более 10 минут;

1.6. Методические рекомендации по подготовке доклада

Алгоритм создания доклада:
1 этап – определение темы доклада
2 этап – определение цели доклада
3 этап − подробное раскрытие информации
4 этап − формулирование основных тезисов и выводов.

1.7. Методические рекомендации по составлению глоссария

1. Внимательно прочитайте и ознакомьтесь с текстом. Вы встретите в нем много различных терминов, которые имеются по данной теме.
2. После того, как вы определили наиболее часто встречающиеся термины, вы должны составить из них список. Слова в этом списке должны быть расположены в строго алфавитном порядке, так как глоссарий представляет собой не что иное, как словарь специализированных терминов.
3. После этого начинается работа по составлению статей глоссария. Статья глоссария - это определение термина. Она состоит из двух частей: 1. точная формулировка термина в именительном падеже; 2. содержательная часть, объемно раскрывающая смысл данного термина.

При составлении глоссария важно придерживаться следующих правил:

- стремитесь к максимальной точности и достоверности информации;
- старайтесь указывать корректные научные термины и избегать всякого рода жаргонизмов. В случае употребления такового, давайте ему краткое и понятное пояснение;
- излагая несколько точек зрения в статье по поводу спорного вопроса, не принимайте ни одну из указанных позиций. Глоссарий - это всего лишь констатация имеющихся фактов;
- также не забывайте приводить в пример контекст, в котором может употреблять данный термин;
- при желании в глоссарий можно включить не только отельные слова и термины, но и целые фразы.

1.8. Методические рекомендации для занятий в интерактивной форме

В учебном процессе, помимо чтения лекций и аудиторных занятий, используются интерактивные формы. В сочетании с внеаудиторной работой это способствует формированию и развитию профессиональных навыков обучающихся.

Интерактивное обучение представляет собой способ познания, осуществляемый в формах совместной деятельности обучающихся, т.е. все участники образовательного процесса взаимодействуют друг с другом, совместно решают поставленные проблемы, моделируют ситуации, обмениваются информацией, оценивают действие коллег и свое собственное поведение, погружаются в реальную атмосферу делового сотрудничества по разрешению проблем.

В курсе изучаемой дисциплины «Анализ данных и машинное обучение» в интерактивной форме часы используются в виде: групповой дискуссии, заслушивании и обсуждении подготовленных обучающимися докладов с презентациями по тематике дисциплины.

Тематика занятий с использованием интерактивных форм

№ п/п	Тема	Интерактивная форма	Часы, отводимые на интерактивные формы	
			Лекции	Практические занятия
1.	Большие данные и машинное обучение.	Групповая дискуссия Доклад с презентацией	-	2
2.	Метрические методы классификации	Групповая дискуссия	-	2
3.	Логические методы классификации	Групповая дискуссия	-	2
4.	Линейные методы классификации	Групповая дискуссия		2
ИТОГО			**8 часов**	

2. Планы практических занятий:

Лабораторная работа № 1. Предобработка данных в Pandas.

План:

1. Использование Phyton для анализа данных.
2. Основные библиотеки Phyton: Scikit-learn, NumPy, SciPy, matplotlib, pandas.
3. Дистрибутив Anaconda.
4. Pandas: базовые методы.
5. Pandas: индексация и извлечение данных.
6. Pandas: применение функций к ячейкам, столбцам и строкам.
7. Pandas: группировка данных.
8. Pandas: таблицы сопряженности.
9. Pandas: сводные таблицы.
10. DataFrame в Pandas.

Литература: [3].

Вопросы для групповой дискуссии:

1. Почему язык Phyton?
2. Для чего используется библиотека NumPy?
3. Для чего используются библиотеки SciPy, matplotlib, pandas?
4. Как выполняется индексация и извлечение данных в Pandas?
5. Как выполняется группировка данных в Pandas?
6. Для чего применяются таблицы сопряженности в Pandas?
7. Как выполняется загрузка данных в DataFrame в библиотеке Pandas?
8. Как выполняется доступ к столбцам DataFrame в библиотеке Pandas?

Задание для самостоятельной работы

1. Анализ данных по доходу населения UCI Adult. В задании предлагается с помощью Pandas ответить на несколько вопросов по данным репозитория UCI Adult. Список вопросов:

- Каков средний возраст (признак age) женщин?
- Какова доля граждан Германии (признак native-country)?
- Постройте гистограмму распределения (bar plot) образования людей (признак education).
- Каковы средние значения и среднеквадратичные отклонения возраста тех, кто получает более 50K в год (признак salary) и тех, кто получает менее 50K в год?
- Правда ли, что люди, которые получают больше 50k, имеют как минимум высшее образование? (признак education - Bachelors, Prof-school, Assoc-acdm, Assoc-voc, Masters или Doctorate)
- Выведите статистику возраста для каждой расы (признак race) и каждого пола. Используйте groupby и describe. Найдите таким образом максимальный возраст мужчин расы Amer-Indian-Eskimo.
- Среди кого больше доля зарабатывающих много (>50K): среди женатых или холостых мужчин (признак marital-status)? Женатыми считаем тех, у кого marital-status начинается с Married (Married-civ-spouse, Married-spouse-absent или Married-AF-spouse), остальных считаем холостыми.
- Какое максимальное число часов человек работает в неделю (признак hours-per-week)? Сколько людей работают такое количество часов и каков среди них процент зарабатывающих много?
- Посчитайте среднее время работы (hours-per-week) зарабатывающих мало и много (salary) для каждой страны (native-country).

2. Анализ данных по пассажирам Титаника. В задании предлагается с помощью Pandas ответить на несколько вопросов по данным репозитория UCI Titanic. Список вопросов:

- Какое количество мужчин и женщин ехало на корабле?
- Какой части пассажиров удалось выжить? Посчитайте долю выживших пассажиров.
- Какую долю пассажиры первого класса составляли среди всех пас- сажиров?
- Какого возраста были пассажиры?
- Посчитайте среднее и медиану возраста пассажиров.
- Коррелируют ли число братьев/сестер с числом родителей/детей? Посчитайте корреляцию Пирсона между признаками SibSp и Parch.
- Какое самое популярное женское имя на корабле? Извлеките из полного имени пассажира (колонка Name) его личное имя (First Name).

Лабораторная работа № 2. Метрические методы классификации в Scikit-learn.

План:

1. Метрические методы классификации.
2. Признаковые описания объекта.
3. Гипотеза компактности.
4. Метрики, виды метрик.
5. Весовая Евклидова метрика, метрика Минковского.
6. Масштабирование признаков.
7. Метод k ближайших соседей.
8. Реализация kNN в классе sklearn.neighbors.KNeighborsClassifier.

9. Кросс-валидация.
10. Алгоритм выполнения кросс-валидации по блокам.
11. Вычисление ошибки на разбиениях.

Литература: [2, с. 127-139].

Вопросы для групповой дискуссии:

1. В чем идея гипотезы компактности?
2. В чем состоит смысл обучения в метрических методах?
3. Для чего используется библиотека Scikit-learn?
4. Для чего выполняют масштабирование признаков?
5. Как обычно выполняется масштабирование количественных признаков?
6. В каком классе Scikit-learn реализован метод kNN?
7. Какой параметр метода k ближайших соседей, задает число соседей для построения прогноза?
8. В чем смысл кросс-валидации?
9. Как вычисляется весовая Евклидова метрика?
10. Приведите формулу Метрики Минковского. Что является ее параметром?

Задание для самостоятельной работы:

Задание 1.

1. В этом задании нужно подобрать оптимальное значение k для алгоритма kNN. Будем использовать набор данных Wine, где требуется предсказать сорт винограда, из которого изготовлено вино, используя результаты химических анализов.
2. Выполните следующие шаги:

- Загрузите выборку Wine по адресу https://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data
- Извлеките из данных признаки и классы. Класс записан в первом столбце (три варианта), признаки — в столбцах со второго по последний. Более подробно о сути признаков можно прочитать по адресу https://archive.ics.uci.edu/ml/datasets/Wine
- Оценку качества необходимо провести методом кроссвалидации по 5 блокам (5-fold). Создайте генератор разбиений, который перемешивает выборку перед формированием блоков (shuffle=True). Для воспроизводимости результата, создавайте генератор KFold с фиксированным параметром random_state=42. В качестве меры качества используйте долю верных ответов (accuracy).
- Найдите точность классификации на кросс-валидации для метода k ближайших соседей (sklearn.neighbors.KNeighborsClassifier), при k от 1 до 50. При каком k получилось оптимальное качество? Чему оно равно (число в интервале от 0 до 1)?
- Произведите масштабирование признаков с помощью функции sklearn.preprocessing.scale. Снова найдите оптимальное k на кросс-валидации.
- Какое значение k получилось оптимальным после приведения признаков к одному масштабу? Как изменилось значение качества? Приведите ответы на вопросы.

Задание 2.

1. Нам понадобится решать задачу регрессии с помощью метода k ближайших соседей — воспользуемся для этого классом sklearn.neighbors.KNeighborsRegressor.

2. Метрика задается с помощью параметра metric, нас будет интересовать значение ’minkowski’. Параметр метрики Минковского задается с помощью параметра p данного класса.

3. Инструкция по выполнению

- Мы будем использовать в данном задании набор данных Boston, где нужно предсказать стоимость жилья на основе различных характеристик расположения (загрязненность воздуха, близость к дорогам и т.д.). Подробнее о признаках можно почитать по адресу https://archive.ics.uci.edu/ml/datasets/Housing
- Загрузите выборку Boston с помощью функции sklearn.datasets.load_boston(). Результатом вызова данной функции является объект, у которого признаки записаны в поле data, а целевой вектор — в поле target.
- Приведите признаки в выборке к одному масштабу при помощи функции sklearn.preprocessing.scale.
- Переберите разные варианты параметра метрики p по сетке от 1 до 10 с таким шагом, чтобы всего было протестировано 200 вариантов (используйте функцию numpy.linspace). Используйте KNeighborsRegressor с n_neighbors=5 и weights=’distance’ - данный параметр добавляет в алгоритм веса, зависящие от расстояния до ближайших соседей. В качестве метрики качества используйте среднеквадратичную ошибку (параметр scoring=’mean_squared_error’ у cross_val_score; при использовании библиотеки scikit-learn версии 18.0.1 и выше необходимо указывать scoring=’neg_mean_squared_error’). Качество оценивайте, как и в предыдущем задании, с помощью кросс-валидации по 5 блокам с random_state = 42, не забудьте включить перемешивание выборки (shuffle=True).
- Определите, при каком p качество на кросс-валидации оказалось оптимальным. Обратите внимание, что cross_val_score возвращает массив показателей качества по блокам; необходимо макет массив показателей качества по блокам; необходимо максимизировать среднее этих показателей.

Лабораторная работа № 3. Деревья решений. Важность признаков

План:

1. Логическая закономерность.
2. Основы вопросы построения логических алгоритмов классификации.
3. Определение бинарного решающего дерева.
4. Реализация решающих деревьев в библиотеке scikit-learn.
5. Важность признаков.
6. Пропуски в данных.

Литература: [2, с. 115-127].

Вопросы для групповой дискуссии:

1. В каких классах scikit-learn реализуются решающие деревья для задач классификации и регрессии?
2. С помощью какой функции scikit-learn реализуется обучение модели решающих деревьев?
3. Какая переменная содержит массив "важностей" признаков?
4. С помощью какой функции можно проверить, является ли число nan?
5. Основы вопросы построения логических алгоритмов классификации.
6. Определение бинарного решающего дерева.
7. Реализация решающих деревьев в библиотеке scikit-learn.
8. Важность признаков.

Задание для самостоятельной работы:

1. Загрузите выборку из файла titanic.csv с помощью пакета Pandas.
2. Оставьте в выборке четыре признака: класс пассажира (Pclass), цену билета (Fare), возраст пассажира (Age) и его пол (Sex).
3. Обратите внимание, что признак Sex имеет строковые значения.
4. Выделите целевую переменную — она записана в столбце Survived.
5. В данных есть пропущенные значения — например, для некоторых пассажиров неизвестен их возраст. Такие записи при чтении их в pandas принимают значение nan. Найдите все объекты, у которых есть пропущенные признаки, и удалите их из выборки.
6. Обучите решающее дерево с параметром random_state=241 и остальными параметрами по умолчанию.
7. Вычислите важности признаков и найдите два признака с наибольшей важностью. Их названия будут ответами для данной задачи (в качестве ответа укажите названия признаков через запятую без пробелов).

Лабораторная работа № 4. Линейная классификация. Нормализация признаков.

План:

1. Линейные алгоритмы классификации.
2. Персептрон.
3. Нормализация признаков. Стандартизация признаков.
4. Реализация линейных классификаторов в библиотеке scikit-learn.
5. Метрика качества.
6. Метод опорных векторов.
7. Опорные объекты.

Литература: [2, с. 139-155].

Вопросы для групповой дискуссии:

1. Сформулируйте постановку задачи линейной классификации.
2. Как выполняется стандартизация признаков?
3. В каком классе scikit-learn реализуется персептрон?
4. Для чего используется функция sklearn.metrics.accuracy_score?
5. Каким классом удобно воспользоваться для стандартизации признаков?
8. На что направлен функционал, который он оптимизирует метод опорных векторов?
6. Какие объекты называют опорными?

Задание для самостоятельной работы:

Задание 1.

1. Загрузите обучающую и тестовую выборки из файлов perceptrontrain.csv и perceptron-test.csv. Целевая переменная записана в первом столбце, признаки — во втором и третьем.
2. Обучите персептрон со стандартными параметрами и random_state=241.
3. Подсчитайте качество (долю правильно классифицированных объектов, accuracy) полученного классификатора на тестовой выборке.

4. Нормализуйте обучающую и тестовую выборку с помощью класса StandardScaler.

5. Обучите персептрон на новых выборках. Найдите долю правильных ответов на тестовой выборке.

6. Найдите разность между качеством на тестовой выборке после нормализации и качеством до нее.

Задание 2.

1. Загрузите выборку из файла svm-data.csv. В нем записана двумерная выборка (целевая переменная указана в первом столбце, признаки — во втором и третьем).

2. Обучите классификатор с линейным ядром, параметром C=100000 и random_state=241. Такое значение параметра нужно использовать, чтобы убедиться, что SVM работает с выборкой как с линейно разделимой. При более низких значениях параметра алгоритм будет настраиваться с учетом слагаемого в функционале, штрафующего за маленькие отступы, из-за чего результат может не совпасть с решением классической задачи SVM для линейно разделимой выборки.

3. Найдите номера объектов, которые являются опорными (нумерация с единицы).