Санкт-Петербургский государственный университет

Р А Б О Ч А Я П Р О Г Р А М М А

учебной дисциплины

Интеллектуальный анализ данных

Data Mining

Язык(и) обучения

русский

Трудоёмкость (границы трудоёмкости) в зачетных единицах: 3

Регистрационный номер рабочей программы: 001131

Санкт-Петербург

2018

Раздел 1. Характеристики учебных занятий

1.1. Цели и задачи учебных занятий

Основной целью освоения дисциплины «Интеллектуальный анализ данных» является формирование у обучающихся устойчивого понимания основных понятий и базовых алгоритмов Data Mining, приобретение практических навыков работы с инструментами, применяемыми в области интеллектуального анализа данных.

Поставленная цель достигается путем решения следующих задач курса:

1) ознакомить студентов с основными задачами, решаемыми в области анализа данных, базовыми алгоритмами Data Mining, областями применения алгоритмов;
2) способствовать развитию практических навыков работы с инструментами, применяемыми в области Data Mining;
3) ознакомить с основными тенденциями развития подходов в области интеллектуального анализа данных.

1.2. Требования к подготовленности обучающегося к освоению содержания учебных занятий (пререквизиты)

Освоение дисциплин: дискретная математика, теория вероятностей, линейная алгебра, основы программирования.

Знание одного из высокоуровневых языков программирования (Java, Python).

1.3. Перечень результатов обучения (learning outcomes)

После изучения курса обучающиеся должны

знать: основную схему обработки данных, различные виды шкал; методы выделения ассоциативных правил; методы эволюционного программирования; метрические, логические, линейные методы классификации; нейросетевые и композиционные методы классификации и регрессии; методы кластеризации;

уметь: осуществлять выбор между различными методами с учетом решаемой задачи; решать задачи обучения с учителем, обучения без учителя; использовать существующие инструментальные средства для решения задач интеллектуального анализа данных.

владеть: средствами и приемами анализа данных для решения задач классификации, регрессии, кластеризации, поиска ассоциативных правил; навыками работы с существующими инструментальными средствами для решения задач интеллектуального анализа данных.

1.4. Перечень активных и интерактивных форм учебных занятий

При изложении части тем, по желанию лектора, применяется мультимедиа-проектор для проведения презентаций.

В рамках данного курса используются такие *активные формы работы,* как:

- подготовка доклада по предложенной преподавателем теме;
- выполнение индивидуальных заданий по реализации изученных алгоритмов;

и *интерактивные формы обучения:*

- интерактивные лекции;
- метод групповой работы при решении задач во время практических занятий;
- экспертная оценка другими обучающимися результатов решения индивидуальных заданий и коллективное обсуждение полученных результатов.

Общий объем активных и интерактивных форм учебных занятий составляет 20 часов.

Раздел 2. Организация, структура и содержание учебных занятий

2.1. Организация учебных занятий

2.1.1. Профиль Технологии баз данных

Трудоёмкость, объёмы учебной работы и наполняемость групп обучающихся																	
Период обучения (модуль)	Контактная работа обучающихся с преподавателем												Самостоятельная работа				Объём активных и интерактивных форм учебных занятий	Трудоёмкость
	лекции	семинары	консультации	практические занятия	лабораторные работы	контрольные работы	коллоквиумы	текущий контроль	промежуточная аттестация	итоговая аттестация	под руководством преподавателя	в присутствии преподавателя	сам.раб. с использованием методических материалов	текущий контроль (сам.раб.)	промежуточная аттестация (сам.раб.)	итоговая аттестация (сам.раб.)		
ОСНОВНАЯ ТРАЕКТОРИЯ																		
очная форма обучения																		
Семестр 3	12		2	20					2				48		24		20	3
ИТОГО	**12**		**2**	**20**					**2**				**48**		**24**		**20**	**3**

Формы текущего контроля успеваемости, виды промежуточной и итоговой аттестации			
Период обучения (модуль)	Формы текущего контроля успеваемости	Виды промежуточной аттестации	Виды итоговой аттестации (только для программ итоговой аттестации и дополнительных образовательных программ)
ОСНОВНАЯ ТРАЕКТОРИЯ			
очная форма обучения			
Семестр 3		экзамен	

2.2. Структура и содержание учебных занятий

Период обучения (модуль): **Семестр 3**

№ п/п	Наименование темы (раздела, части)	Вид учебных занятий	Количество часов
1	**Введение в интеллектуальный**	лекции	1

	анализ данных. Основная схема обработки данных. Шкалы. Визуализация данных. Основные задачи интеллектуального анализа данных.	практические занятия	0
		по методическим материалам	6
2	**Деревья решений. Продукционные правила.** Построение деревьев решений для задачи классификации. Алгоритм ID3. Алгоритм C4.5. Принципы выделения продукционных правил по дереву. Среда Weka. Язык программирования R. Задача «Куда поехать отдыхать?»	лекции	2
		практические занятия	3
		по методическим материалам	7
3	**Поиск ассоциативных правил.** Задача поиска ассоциативных правил. Алгоритм APriory. Алгоритм построения префиксного FP-дерева. Алгоритм выделения частых наборов по FP-дереву. Задача «Супермаркет».	лекции	2
		практические занятия	3
		по методическим материалам	7
4	**Генетические алгоритмы.** История эволюционного программирования. Основные этапы генетических алгоритмов. Скрещивание. Мутация. Отбор. Решение задачи поиска частых наборов с помощью генетического программирования. Задача «Супермаркет». Решение задачи классификации с помощью генетических алгоритмов. Задача «Зоопарк».	лекции	2
		практические занятия	4
		по методическим материалам	7
5	**Нейронные сети.** Основные понятия нейронных сетей. Виды нейронных сетей. Алгоритм обратного распространения. Применение нейронных сетей для решения задач интеллектуального анализа данных. Задача «Паркинсон».	лекции	2
		практические занятия	4
		по методическим материалам	7
6	**Композиции классификаторов. Кластеризация.** Понятие композиции классификаторов. Бустинг. Алгоритм AdaBoost. Бэггинг. Метод случайных подпространств. Постановка задачи кластеризации. Графовые методы ее решения. Статистические методы решения задачи кластеризации. Построение таксономий. Задача «Спам».	лекции	2
		практические занятия	3
		по методическим материалам	7
7	**Преобразование данных.** Выбор атрибутов. Дискретизация	лекции	1
		практические занятия	3

	атрибутов. Проекция данных. Сэмплирование. Очистка данных. Сведение задачи многоклассовой классификации к бинарной. Задача «Сбербанк».	по методическим материалам	7

Раздел 3. Обеспечение учебных занятий

3.1. Методическое обеспечение

3.1.1. Методические указания по освоению дисциплины

Самостоятельная работа студентов включает в себя самостоятельную проработку рассмотренных на аудиторных занятиях материалов; выполнение практических заданий по изучаемым разделам; выбор темы для детального изучения, подготовку презентации по ней, выступление с докладом перед группой; изучение рекомендованной литературы и ресурсов в сети интернет. Время и место самостоятельной работы выбираются студентами по своему усмотрению с учетом рекомендаций преподавателя.

Подготовка к семинару заключается в следующем:

- внимательно изучить материал предыдущего семинара;
- целесообразно составить краткий конспект или схему, отображающую смысл и связи основных понятий данного раздела, включенных в него тем, а затем, полезно изучить выдержки из литературы;
- узнать тему предстоящего семинара (по тематическому плану, по материалам, размещенным в системе дистанционного обучения Blackboard);
- ознакомиться с учебным материалом по учебным пособиям;
- записать возможные вопросы, которые вы зададите преподавателю в ходе аудиторного занятия.

Подготовка к практическим занятиям:

- внимательно изучить материал семинаров, относящихся к данному заданию, ознакомиться с учебным материалом по учебным пособиям;
- выполнить практические задания домашней работы;
- уяснить, какие учебные элементы остались для неясными и сформулировать вопросы, которые необходимо задать преподавателю на занятии или консультации;
- готовиться можно индивидуально или парами.

Изучение дисциплины заканчивается зачетом. При непосредственной подготовке к зачету рекомендуется тщательно разобрать все основные определения и алгоритмы. Отметить пробелы в знаниях, которые следует ликвидировать в ходе консультации.

3.1.2. Методическое обеспечение самостоятельной работы

Самостоятельная работа студентов в рамках дисциплины «Интеллектуальный анализ данных» является важным компонентом обучения. Данной программой предусмотрены виды деятельности студента, которые направляются и корректируются преподавателем, и виды учебной деятельности, которые осуществляются студентом самостоятельно в рамках плана изучения данной учебной дисциплины.

К группе видов и форм самостоятельной работы студентов относятся:

- подготовка доклада по предложенной преподавателем теме;

– выполнение индивидуальных заданий по реализации изученных алгоритмов.

Для организации самостоятельной работы студентов рекомендуется предоставить презентации с изучаемым материалом, проводить консультации во время аудиторных занятий.

3.1.3. Методика проведения текущего контроля успеваемости и промежуточной аттестации и критерии оценивания

По результатам освоения дисциплины студентам предлагается выполнить доклад по предложенной преподавателем теме.

Промежуточная аттестация проходит в форме собеседования по пройденному материалу, с учетом результатов выполнения индивидуальных заданий и сделанного доклада.

Оценка «отлично» ставится при условии успешного выполнения всех практических заданий, своевременно сделанного доклада или полного ответа на один теоретический вопрос по курсу, знания основных определений и алгоритмов.

Оценка «хорошо» ставится при условии успешного выполнения не менее 75% практических заданий, своевременно сделанного доклада или полного ответа на один теоретический вопрос по курсу, знания основных определений и алгоритмов.

Оценка «удовлетворительно» ставится при условии успешного выполнения не менее 50% практических заданий или полного ответа на один теоретический вопрос по курсу, знания основных определений и алгоритмов.

Оценка «неудовлетворительно» ставится при условии успешного выполнения менее 50% практических заданий, неполного ответа на теоретический вопрос по курсу, отсутствии знаний основных определений и алгоритмов.

Преподаватель имеет право предоставить информацию о задолженностях студента в аттестационную комиссию.

3.1.4. Методические материалы для проведения текущего контроля успеваемости и промежуточной аттестации (контрольно-измерительные материалы, оценочные средства)

Примерный перечень тем докладов:

1. Text Mining: особенности интеллектуального анализа текстовых документов.
2. Самоорганизующиеся карты Кохонена и их применение к задачам интеллектуального анализа данных.
3. Алгоритмы интеллектуального анализа данных для решения задачи прогнозирования.
4. Средства интеллектуального анализа данных в системах управления базами данных (MS SQL Server, Oracle, IBM DB2).
5. Нечеткие нейронные сети, их архитектура и применение в системах принятия решений.
6. Нечеткие деревья решений и их сравнение с классическими (модификация на основе алгоритма C4.5).
7. Средства визуализации данных в R.
8. Современные алгоритмы кластеризации: DBScan и Affinity Propagation.

Примерный перечень теоретических вопросов по курсу:

1. Место интеллектуального анализа данных среди других областей Computer Science. Определение интеллектуального анализа данных. Основные задачи и приложения.
2. Понятие данных. Различные шкалы данных. Типы наборов данных. Проблема качества данных. Основные этапы предобработки данных.
3. Понятие данных. Различные шкалы данных. Способы преобразования данных: выбор атрибутов, дискретизация атрибутов, проекция данных, очистка данных, сэмплирование.
4. Деревья решений. Два этапа конструирования деревьев решений. Алгоритм ID3. Преимущества и недостатки метода.
5. Деревья решений. Два этапа конструирования деревьев решений. Алгоритм C4.5. Преимущества и недостатки метода.
6. Деревья решений. Два этапа конструирования деревьев решений. Алгоритм CART. Преимущества и недостатки метода.
7. Классификация на основе правил. Задача построения продукционных правил по дереву решений. Извлечение правил из данных. Алгоритмы AQ, CN2.
8. Вопросы качества классификации. Разделение выборки. Ошибки 1го и 2го рода. Метрики качества классификации, микро- и макро- усреднение оценок. Метод перекрестного контроля.
9. Задача поиска ассоциативных правил. Алгоритм APriory.
10. Задача поиска ассоциативных правил. Применение FP-дерева для решения задачи.
11. Искусственные нейронные сети для решения задач интеллектуального анализа данных. Персептрон Розенблатта и правила Хебба. Сети прямого распространения, различные функции активации нейронов сети.
12. Классификации искусственных нейронных сетей. Рекуррентные нейронные сети: сеть Хопфилда, сеть Хэмминга. Радиально-базисные сети. Самоорганизующаяся сеть Кохонена.
13. Обучение нейронных сетей. Метод градиентного спуска и его вариации.
14. Обучение нейронных сетей. Алгоритм обратного распространения ошибки.
15. Генетические алгоритмы для решения задач интеллектуального анализа данных. Основные понятия генетических алгоритмов. Вариации генетических операторов.
16. Генетические алгоритмы для решения задач интеллектуального анализа данных. Основные понятия генетических алгоритмов. Популярные модели: Genitor, CHC, островная модель, модель кооперативной коэволюции.
17. Задача кластеризации. Постановка задачи, ее виды. Графовые методы кластеризации. Вопросы качества кластеризации.
18. Задача кластеризации. Постановка задачи, ее виды. Статистические методы кластеризации. Вопросы качества кластеризации.
19. Задача кластеризации. Постановка задачи, ее виды. Иерархические методы кластеризации. Вопросы качества кластеризации.

3.1.5. Методические материалы для оценки обучающимися содержания и качества учебного процесса

Для оценки обучающимися содержания и качества учебного процесса используется анкета-отзыв установленная локальными актами СПбГУ.

3.2. Кадровое обеспечение

3.2.1. Образование и (или) квалификация преподавателей и иных лиц, допущенных к проведению учебных занятий

К чтению лекций должны привлекаться преподаватели, имеющие ученую степень и/или ученое звание, опыт планирования и организации учебного процесса, или специалисты в этой области.

3.2.2. Обеспечение учебно-вспомогательным и (или) иным персоналом

Для технического обеспечения учебного процесса необходима возможность прибегать к помощи специалистов, ответственных за надлежащее функционирование компьютеров и программного обеспечения, а также за своевременное поддержание в рабочем состоянии другой используемой техники.

3.3. Материально-техническое обеспечение

3.3.1. Характеристики аудиторий (помещений, мест) для проведения занятий

Аудитории и помещения, предназначенные для проведения занятий по данной дисциплине должны отвечать санитарным нормам.

В аудиториях требуется наличие компьютеризированных рабочих мест для проведения совместных практических работ и демонстрации материалов курса: мультимедийный проектор, доска.

3.3.2. Характеристики аудиторного оборудования, в том числе неспециализированного компьютерного оборудования и программного обеспечения общего пользования

Для реализации программы необходим доступ преподавателей к офисной технике (персональный компьютер, копировальный аппарат, принтер), а также достаточное количество расходных материалов к ней, выделенных для использования в учебном процессе.

Минимально необходимый для реализации курса перечень материально-технического обеспечения включает: мультимедийный проектор для презентаций и демонстраций, компьютеры для проведения практических работ.

3.3.3. Характеристики специализированного оборудования

Нет специальных требований.

3.3.4. Характеристики специализированного программного обеспечения

При практической работе каждый обучающийся во время занятий и самостоятельной подготовки должен быть обеспечен рабочим местом в компьютерном классе с выходом в Интернет.

Необходим доступ к инструментам и библиотекам для разработки: Weka; R и RStudio.

3.3.5. Перечень и объёмы требуемых расходных материалов

Фломастеры цветные, губки, бумага формата А4, канцелярские товары, картриджи принтеров, диски, флеш-накопители и др. в объёме, необходимом для организации и проведения занятий, по заявкам преподавателей, подаваемым в установленные сроки

3.4. Информационное обеспечение

3.4.1. Список обязательной литературы

1. Cios, K.J., Swiniarski, R.W., Pedrycz, W., Kurgan L.A. “Data Mining: A Knowledge Discovery Approach” (2007)
(электронный доступ через библиотеку университета).
2. Bramer, M. “Principles of Data Mining” (2nd ed., 2013)
(электронный доступ через библиотеку университета).
3. Galushkin, A. I. “Neural Networks Theory” (2007)
(электронный доступ через библиотеку университета).
4. Sivanandam, S.N., Deepa, S.N. “Introduction to Genetic Algorithms” (2008)
(электронный доступ через библиотеку университета).

3.4.2. Список дополнительной литературы

1. C. Sammut, G.I. Webb (Ed.) “Encyclopedia of Machine Learning” (2011)
(электронный доступ через библиотеку университета).
2. Christmann, A., Steinwart, I. “Support Vector Machines” (2008)
(электронный доступ через библиотеку университета).
3. Hastie, T., Tibshirani, R., Friedman, J. “Elements of Statistical Learning: Data Mining, Inference and Prediction” (2nd ed., 2009)
(электронный доступ через библиотеку университета).

3.4.3. Перечень иных информационных источников

1. Машинное обучение, курс лекций: http://www.machinelearning.ru/
2. ШАД Yandex: http://shad.yandex.ru/lectures/machine_learning.xml
3. Официальный сайт Weka: http://www.cs.waikato.ac.nz/ml/weka/
4. Официальный сайт Encog: http://www.heatonresearch.com/encog
5. Сайт компании BaseGroup: http://www.basegroup.ru
6. Официальный сайт FANN: http://leenissen.dk/fann/wp/

Раздел 4. Разработчики программы

Фамилия, имя, отчество	Учёная степень	Учёное звание	Должность	Контактная информация (служебный адрес электронной почты, служебный телефон)
Романенко Елена Станиславовна	--	--	ст. преп.	e.s.romanenko@spbu.ru