Министерство науки и высшего образования Российской Федерации

ФГБОУ ВО «БАЙКАЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

УТВЕРЖДАЮ
Проректор по учебной работе
д.э.н., доц. Бубнов В.А

_____

25.06.2021г.

Рабочая программа дисциплины
Б1.О.13. Современные технологии обработки массовых данных

Направление подготовки: 01.04.05 Статистика
Направленность (профиль): Экспертная бизнес-аналитика
Квалификация выпускника: магистр
Форма обучения: заочная

Курс	2
Семестр	22
Лекции (час)	30
Практические (сем, лаб.) занятия (час)	0
Самостоятельная работа, включая подготовку к экзаменам и зачетам (час)	114
Курсовая работа (час)	
Всего часов	144
Зачет (семестр)	
Экзамен (семестр)	22

Иркутск 2021

Программа составлена в соответствии с ФГОС ВО по направлению 01.04.05 Статистика.

Автор В.Р. Абдуллин

Рабочая программа обсуждена и утверждена на заседании кафедры математических методов и цифровых технологий

Заведующий кафедрой С.С. Ованесян

Дата актуализации рабочей программы: 30.06.2022

1. Цели изучения дисциплины

Целью освоения дисциплины является формирование знаний и умений, связанных с поиском новых, нетривиальных, практически полезных и доступных для интерпретации человеком знаний, скрытых в больших объемах, накопленных сырых данных средствами автоматического анализа. Знания и умения приобретаемые в процессе изучения дисциплины имеют широкую сферу применения: рекомендательные системы, системы медицинской диагностики, задачи привлечения и удержания клиентов, кредитный скоринг, категоризация текстовых документов и т.д.

2. Перечень планируемых результатов обучения по дисциплине (модулю), соотнесенных с планируемыми результатами освоения образовательной программы

Компетенции обучающегося, формируемые в результате освоения дисциплины

Код компетенции по ФГОС ВО	Компетенция
ОПК-3	Способен анализировать статистические данные с применением методов математической и дескриптивной статистики и вероятностных методов анализа числовой и нечисловой информации

Структура компетенции

Компетенция	Формируемые ЗУНы
ОПК-3 Способен анализировать статистические данные с применением методов математической и дескриптивной статистики и вероятностных методов анализа числовой и нечисловой информации	З. Знать методы математической и дескриптивной статистики и вероятностных методов анализа числовой и нечисловой информации У. Уметь анализировать статистические данные с помощью методов математической статистики и современных технологий обработки массовых данных Н. Владеть навыками решения задач с помощью методов математической статистики и современных технологий обработки массовых данных

3. Место дисциплины (модуля) в структуре образовательной программы

Принадлежность дисциплины - БЛОК 1 ДИСЦИПЛИНЫ (МОДУЛИ): Обязательная часть.

Предшествующие дисциплины (освоение которых необходимо для успешного освоения данной): "Базы данных", "Математическая статистика", "Прикладная эконометрика"

4. Объем дисциплины (модуля) в зачетных единицах с указанием количества академических часов, выделенных на контактную работу обучающихся с преподавателем (по видам учебных занятий) и на самостоятельную работу обучающихся

Общая трудоемкость дисциплины составляет 4 зач. ед.,144 часов.

Вид учебной работы	Количество часов
Контактная(аудиторная) работа	
Лекции	30
Практические (сем, лаб.) занятия	0

Самостоятельная работа, включая подготовку к экзаменам и зачетам	114
Всего часов	144

5. Содержание дисциплины (модуля), структурированное по темам (разделам) с указанием отведенного на них количества академических часов и видов учебных занятий

5.1. Содержание разделов дисциплины

№ п/п	Раздел и тема дисциплины	Семе-стр	Лек-ции	Семинар Лаборат. Практич.	Само-стоят. раб.	В интера-ктивной форме	Формы текущего контроля успеваемости
1	Анализ данных и машинное обучение	22	2		10		
2	Логические методы классификации	22	4		14		
3	Метрические методы классификации	22	2		14		
4	Линейные методы классификации	22	4		14		Лабораторная работа №1
5	Понижение размерности и метод главных компонент	22	4		14		Лабораторная работа №2
6	Композиции алгоритмов	22	4		16		Лабораторная работа №3
7	Нейронные сети	22	4		16		
8	Кластеризация и визуализация	22	6		16		Лабораторная работа №4
	ИТОГО		30		114		

5.2. Лекционные занятия, их содержание

№ п/п	Наименование разделов и тем	Содержание
1	Постановка задачи машинного обучения	Задача обучения по прецедентам. Признаковое описание объектов. Обучение и применение модели. Проблема переобучения. Эмпирические оценки обобщающей способности алгоритма. Межотраслевой стандарт решения задач машинного обучения
2	Метод ближайших соседей	Метод ближайших соседей. Метод окна Парзена. Метрические методы классификации в задаче восстановления регрессии. Обнаружение выбросов
3	Логистическая регрессия	Логистическая регрессия. Применение логистической регрессии в задаче кредитного скоринга. Регуляризованная логистическая регрессия
4	Линейная регрессия и метод главных компонент	Решение задачи многомерной линейной регрессии с помощью сингулярного разложения. Гребневая регрессия. Метод LASSO. Метод главных компонент
5	Бэггинг и случайный лес	Простое голосование классификаторов. Бэггинг и метод случайных подпространств. Случайный лес
6	Нейронные сети	Линейная модель нейрона. Часто используемые функции

№ п/п	Наименование разделов и тем	Содержание
		активации. Линейные алгоритмы классификации и регрессии. Реализация логических функций. Метод обратного распространения ошибки. Стандартные эвристики
7	Иерархическая кластеризация	Агломеративная иерархическая кластеризация. Формула Ланса-Уильямса. Визуализация кластерной структуры. Свойства иерархической кластеризации

5.3. Семинарские, практические, лабораторные занятия, их содержание

№ раздела и темы	Содержание и формы проведения
2	Решающие деревья. Бинарные решающие деревья. Алгоритм построения решающего дерева. Обработка пропусков. Достоинства и недостатки решающих деревьев. Усечение дерева
4	Метод опорных векторов. Метод стохастического градиента. Достоинства и недостатки. Проблема переобучения. Метод опорных векторов. Обобщение для нелинейного случая
5	Линейная регрессия и метод главных компонент. Решение задачи многомерной линейной регрессии с помощью сингулярного разложения. Гребневая регрессия. Метод LASSO. Метод главных компонент
6	Градиентный бустинг. Линейные композиции для классификации и регрессии. Градиентный бустинг. Параметрическая аппроксимация градиентного шага. Алгоритм градиентного бустинга. Стохастический градиентный бустинг. Градиентный бустинг над деревьями
7	Нейронные сети. Линейная модель нейрона. Часто используемые функции активации. Линейные алгоритмы классификации и регрессии. Реализация логических функций. Метод обратного распространения ошибки. Стандартные эвристики
8	Метод k–средних. Постановка задачи кластеризации. Типы кластерных структур. Метод k–средних. EM алгоритм. Построение начального приближения для метода k–средних. Недостатки метода k–средних и способы их устранения
8	Нелинейные методы понижения размерности. Распознавание рукописных цифр. Многообразия, вложенные в пространство признаков. Постановка задачи нелинейного понижения размерности. Визуализация данных. Многомерное шкалирование. SNE и t-SNE. Визуализация рукописных цифр. Внутренняя структура данных

6. Фонд оценочных средств для проведения промежуточной аттестации по дисциплине (полный текст приведен в приложении к рабочей программе)

6.1. Текущий контроль

№ п/п	Этапы формирования компетенций (Тема из рабочей программы дисциплины)	Перечень формируемых компетенций по ФГОС ВО	(ЗУНы: (З.1…З.n, У.1…У.n, Н.1…Н.n)	Контрольные задания или иные материалы, необходимые для оценки знаний, умений, навыков и (или) опыта деятельности, характеризующих этапы формирования компетенций в процессе освоения образовательной программы (Наименование оценочного средства)	Описание показателей и критериев оценивания компетенций на различных этапах их формирования, описание шкал оценивания (по 100-балльной шкале)
1	4. Линейные методы классификации	ОПК-3	З.Знать методы математической и дескриптивной статистики и вероятностных методов анализа числовой и нечисловой информации У.Уметь анализировать статистические данные с помощью методов математической статистики и современных технологий обработки массовых данных Н.Владеть навыками решения задач с помощью методов математической статистики и современных технологий обработки массовых данных	Лабораторная работа №1	Выполненная лабораторная работа оценивается в 25 баллов (25)
2	5. Понижение размерности и метод главных компонент	ОПК-3	З.Знать методы математической и дескриптивной статистики и вероятностных методов анализа числовой и нечисловой информации У.Уметь анализировать статистические данные с помощью методов математической статистики и современных технологий обработки массовых данных Н.Владеть навыками решения задач с помощью методов математической статистики и современных	Лабораторная работа №2	Выполненная лабораторная работа оценивается в 25 баллов (25)

№ п/п	Этапы формирования компетенций (Тема из рабочей программы дисциплины)	Перечень формируемых компетенций по ФГОС ВО	(ЗУНы: (З.1…З.n, У.1…У.n, Н.1…Н.n)	Контрольные задания или иные материалы, необходимые для оценки знаний, умений, навыков и (или) опыта деятельности, характеризующих этапы формирования компетенций в процессе освоения образовательной программы (Наименование оценочного средства)	Описание показателей и критериев оценивания компетенций на различных этапах их формирования, описание шкал оценивания (по 100-балльной шкале)
			технологий обработки массовых данных		
3	6. Композиции алгоритмов	ОПК-3	З.Знать методы математической и дескриптивной статистики и вероятностных методов анализа числовой и нечисловой информации У.Уметь анализировать статистические данные с помощью методов математической статистики и современных технологий обработки массовых данных Н.Владеть навыками решения задач с помощью методов математической статистики и современных технологий обработки массовых данных	Лабораторная работа №3	Выполненная лабораторная работа оценивается в 25 баллов (25)
4	8. Кластеризация и визуализация	ОПК-3	З.Знать методы математической и дескриптивной статистики и вероятностных методов анализа числовой и нечисловой информации У.Уметь анализировать статистические данные с помощью методов математической статистики и современных технологий обработки массовых данных Н.Владеть навыками решения задач с помощью методов математической	Лабораторная работа №4	Выполненная лабораторная работа оценивается в 25 баллов (25)

№ п/п	Этапы формирования компетенций (Тема из рабочей программы дисциплины)	Перечень формируемых компетенций по ФГОС ВО	(ЗУНы: (З.1…З.n, У.1…У.n, Н.1…Н.n)	Контрольные задания или иные материалы, необходимые для оценки знаний, умений, навыков и (или) опыта деятельности, характеризующих этапы формирования компетенций в процессе освоения образовательной программы (Наименование оценочного средства)	Описание показателей и критериев оценивания компетенций на различных этапах их формирования, описание шкал оценивания (по 100-балльной шкале)
			статистики и современных технологий обработки массовых данных		
				Итого	**100**

6.2. Промежуточный контроль (зачет, экзамен)

Рабочим учебным планом предусмотрен Экзамен в семестре 22.

ВОПРОСЫ ДЛЯ ПРОВЕРКИ ЗНАНИЙ:

1-й вопрос билета (40 баллов), вид вопроса: Тест/проверка знаний. Критерий: один правильный ответ на вопрос теста оценивается в 4 балла.

Компетенция: ОПК-3 Способен анализировать статистические данные с применением методов математической и дескриптивной статистики и вероятностных методов анализа числовой и нечисловой информации

Знание: Знать методы математической и дескриптивной статистики и вероятностных методов анализа числовой и нечисловой информации

1. Алгоритм SAG
2. Алгоритм построения решающего дерева
3. Бэггинг и случайный лес
4. Градиентные методы численной минимизации и алгоритм SG
5. Градиентный бустинг
6. Градиентный бустинг: модификации и эвристики
7. Гребневая регрессия
8. Иерархическая кластеризация
9. Кластеризация и визуализация
10. Композиции алгоритмов
11. Линейная регрессия
12. Линейные методы классификации
13. Логистическая регрессия
14. Машинное обучение в прикладных задачах
15. Метод LASSO
16. Метод ближайших соседей
17. Метод обратного распространения ошибки
18. Метод окна Парзена
19. Метод опорных векторов
20. Метод опорных векторов. Обобщение для нелинейного случая
21. Метод стохастического градиента. Достоинства и недостатки
22. Метод стохастического градиента. Постановка задачи
23. Метрики качества классификации

24. Метрические методы классификации
25. Метрические методы классификации в задаче восстановления регрессии
26. Многоклассовая классификация
27. Нейронные сети
28. Нелинейные методы понижения размерности
29. Обзор алгоритмов
30. Обнаружение выбросов
31. Обработка пропусков. Достоинства и недостатки решающих деревьев
32. Оценивание качества
33. Понижение размерности и метод главных компонент
34. Предобработка данных
35. Применение классификации в решении задач частичного обучения
36. Применение кластеризации в решении задач частичного обучения
37. Примеры применения машинного обучения
38. Проблема переобучения. Методология решения задач машинного обучения
39. Работа с категориальными и текстовыми признаками
40. Работа с числовыми признаками
41. Регуляризованная логистическая регрессия
42. Решающие деревья
43. Решение задачи многомерной линейной регрессии с помощью сингулярного разложения
44. Способы устранения недостатков решающих деревьев
45. Стандартные эвристики
46. Формальная постановка задачи машинного обучения
47. Частичное обучение
48. Этапы анализа данных

ТИПОВЫЕ ЗАДАНИЯ ДЛЯ ПРОВЕРКИ УМЕНИЙ:

2-й вопрос билета (30 баллов), вид вопроса: Задание на умение. Критерий: полнота и правильность выполнения задания.

Компетенция: ОПК-3 Способен анализировать статистические данные с применением методов математической и дескриптивной статистики и вероятностных методов анализа числовой и нечисловой информации

Умение: Уметь анализировать статистические данные с помощью методов математической статистики и современных технологий обработки массовых данных

Задача № 1. Важность признаков
Задача № 2. Выбор метрики
Задача № 3. Выбор числа соседей
Задача № 4. Градиентный бустинг над решающими деревьями
Задача № 5. Нормализация признаков
Задача № 6. Опорные объекты
Задача № 7. Размер случайного леса

ТИПОВЫЕ ЗАДАНИЯ ДЛЯ ПРОВЕРКИ НАВЫКОВ:

3-й вопрос билета (30 баллов), вид вопроса: Задание на навыки. Критерий: полнота и правильность выполнения задания.

Компетенция: ОПК-3 Способен анализировать статистические данные с применением методов математической и дескриптивной статистики и вероятностных методов анализа числовой и нечисловой информации

Навык: Владеть навыками решения задач с помощью методов математической статистики и современных технологий обработки массовых данных
Задание № 1. Анализ текстов
Задание № 2. Прогноз оклада по описанию вакансии
Задание № 3. Составление фондового индекса
Задание № 4. Уменьшение количества цветов изображения

ОБРАЗЕЦ БИЛЕТА

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«БАЙКАЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»
(ФГБОУ ВО «БГУ»)

Направление - 01.04.05 Статистика
Профиль - Экспертная бизнес-аналитика
Кафедра математических методов и цифровых технологий
Дисциплина - Современные технологии обработки массовых данных

ЭКЗАМЕНАЦИОННЫЙ БИЛЕТ № 1

1. Тест (40 баллов).
2. Градиентный бустинг над решающими деревьями (30 баллов).
3. Составление фондового индекса (30 баллов).

Составитель _____ В.Р. Абдуллин

Заведующий кафедрой _____ С.С. Ованесян

7. Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины (модуля)

а) основная литература:
1. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP. учеб. пособие [для вузов]. рек. УМО вузов по унив. политехн. образованию. 2-е изд., перераб. и доп./ А. А. Барсегян [и др.].- СПб.: БХВ-Петербург, 2008.-375 с.
2. Нестеров С.А. Интеллектуальный анализ данных средствами MS SQL Server 2008 [Электронный ресурс] / С.А. Нестеров. — Электрон. текстовые данные. — М. : Интернет-Университет Информационных Технологий (ИНТУИТ), 2016. — 303 c. — 2227-8397. — Режим доступа: http://www.iprbookshop.ru/62813.html

б) дополнительная литература:
1. Осовский С., Osowski S., Рудинский И. Д. Нейронные сети для обработки информации. Sieci neuronowe do przetwarzania informacji. Sieci neuronowe do przetwarzania informacji/ Станислав Осовский.- М.: Финансы и статистика, 2004.-343 с.
2. Полубояров В.В. Использование MS SQL Server Analysis Services 2008 для построения хранилищ данных [Электронный ресурс] / В.В. Полубояров. — 2-е изд. — Электрон. текстовые данные. — М. : Интернет-Университет Информационных Технологий (ИНТУИТ), 2016. — 663 c. — 2227-8397. — Режим доступа: http://www.iprbookshop.ru/73682.html

8. Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины (модуля), включая профессиональные базы данных и информационно-справочные системы

Для освоения дисциплины обучающемуся необходимы следующие ресурсы информационно-телекоммуникационной сети «Интернет»:
– Сайт Байкальского государственного университета, адрес доступа: http://bgu.ru/, доступ круглосуточный неограниченный из любой точки Интернет
– Электронно-библиотечная система IPRbooks, адрес доступа: http://www.iprbookshop.ru. доступ неограниченный

9. Методические указания для обучающихся по освоению дисциплины (модуля)

Изучать дисциплину рекомендуется в соответствии с той последовательностью, которая обозначена в ее содержании. Для успешного освоения курса обучающиеся должны иметь первоначальные знания теории вероятностей, математической статистики и методов оптимизации.

На лекциях преподаватель озвучивает тему, знакомит с перечнем литературы по теме, обосновывает место и роль этой темы в данной дисциплине, раскрывает ее практическое значение. В ходе лекций студенту необходимо вести конспект, фиксируя основные понятия и проблемные вопросы.

Задание на практическое (семинарское) занятие сообщается обучающимся до его проведения. На семинаре преподаватель организует обсуждение этой темы, выступая в качестве организатора, консультанта и эксперта учебно-познавательной деятельности обучающегося.

Изучение дисциплины (модуля) включает самостоятельную работу обучающегося.

Основными видами самостоятельной работы студентов с участием преподавателей являются:

• текущие консультации;

• прием и защита лабораторных работ (во время проведения занятий);

Основными видами самостоятельной работы студентов без участия преподавателей являются:

• формирование и усвоение содержания конспекта лекций на базе рекомендованной лектором учебной литературы, включая информационные образовательные ресурсы (электронные учебники, электронные библиотеки и др.);

• самостоятельное изучение отдельных тем или вопросов по учебникам или учебным пособиям;

• подготовка к семинарам и лабораторным работам.

10. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), включая перечень программного обеспечения

В учебном процессе используется следующее программное обеспечение:
– MS Office,
– ActivePython x64,
– SQL Server Data Tools (SSDT),

11. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине (модулю):

В учебном процессе используется следующее оборудование:
– Помещения для самостоятельной работы, оснащенные компьютерной техникой с возможностью подключения к сети "Интернет" и обеспечением доступа в электронную информационно-образовательную среду вуза,

– Учебные аудитории для проведения: занятий лекционного типа, занятий семинарского типа, практических занятий, выполнения курсовых работ, групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации, укомплектованные специализированной мебелью и техническими средствами обучения,
– Компьютерный класс