МИНОБРНАУКИ РОССИИ
федеральное государственное бюджетное образовательное учреждение
высшего образования
«Санкт-Петербургский государственный экономический университет»

УТВЕРЖДАЮ
Проректор по учебной и методической работе
_____/Шубаева В.Г./
«20» июня 2018 г.

МЕТОДЫ АНАЛИЗА ДАННЫХ

Рабочая программа дисциплины

Направление подготовки	01.03.02 Прикладная математика и информатика
Направленность (профиль) программы	Прикладная математика и информатика в экономике и управлении
Уровень высшего образования	бакалавриат
Форма обучения	очная

Составители:
_____/ д.т.н., профессор Фридман Г.М.
_____/ ассистент Ивахненко Д.А.

Санкт-Петербург
2018

СОДЕРЖАНИЕ

1. ЦЕЛЬ И ЗАДАЧИ ДИСЦИПЛИНЫ 3
2. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЫ ..3
3. ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ ПО ДИСЦИПЛИНЕ 3
4. ОБЪЕМ И СТРУКТУРА ДИСЦИПЛИНЫ 4
5. СОДЕРЖАНИЕ РАЗДЕЛОВ И ТЕМ ДИСЦИПЛИНЫ 4
6. ЗАНЯТИЯ СЕМИНАРСКОГО ТИПА 6
7. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ДЛЯ ОБУЧАЮЩИХСЯ 7
7.1. Методические указания для обучающегося по освоению дисциплины 7
7.2. Организация самостоятельной работы 7
8. ОБРАЗОВАТЕЛЬНЫЕ ТЕХНОЛОГИИ 8
9. РЕСУРСНОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ 9
9.1. Учебно-методическое и информационное обеспечение дисциплины 9
9.2. Материально-техническое обеспечение учебного процесса 10
10. ОСОБЕННОСТИ ОСВОЕНИЯ ДИСЦИПЛИНЫ ДЛЯ ИНВАЛИДОВ И ЛИЦ С ОГРАНИЧЕННЫМИ ВОЗМОЖНОСТЯМИ ЗДОРОВЬЯ 10
11. ФОНД ОЦЕНОЧНЫХ СРЕДСТВ ДЛЯ ПРОВЕДЕНИЯ ТЕКУЩЕЙ И ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ ОБУЧАЮЩИХСЯ ПО ДИСЦИПЛИНЕ 10

1. ЦЕЛЬ И ЗАДАЧИ ДИСЦИПЛИНЫ

Цель дисциплины: познакомить обучающихся с современными методами анализа данных, теоретическими основами и алгоритмами машинного обучения для решения широкого круга практических задач.

Задачи:

- исследование области применения методов анализа данных,
- освоение процесса проведения анализа данных различной природы и построения выводов,
- построение математических моделей машинного обучения на основе проведенного анализа, интерпретация результатов для принятия эффективных управленческих решений.

2. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЫ

Дисциплина Б1.В.06 «Методы анализа данных» относится к вариативной части Блока 1 является обязательной для освоения обучающимся после выбора обучающимся направленности (профиля) программы.

3. ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ ПО ДИСЦИПЛИНЕ

Планируемые результаты обучения по дисциплине, соотнесенные с планируемыми результатами освоения образовательной программы представлены в таблице 3.1.

Таблица 3.1 – Планируемые результаты обучения по дисциплине, соотнесенные с планируемыми результатами освоения образовательной программы

Код и наименование компетенции выпускника	Этапы формирования компетенций	Планируемые результаты обучения/индикаторы достижения компетенций (показатели освоения компетенции)
ОПК-2 способностью приобретать новые научные и профессиональные знания, используя современные образовательные и информационные технологии	Второй уровень (углубленный) (ОПК-2)-1	Декомпозиция 1 Знать: основные методы и алгоритмы анализа данных, сферы их применения З2(I) (ОПК-2) Уметь: применять методы анализа данных для получения новых научных и профессиональных знаний У2(I) (ОПК-2) Владеть: современными инструментальными средствами анализа данных В2(I) (ОПК-2)
ОПК-4 способностью решать стандартные задачи профессиональной деятельности на основе информационной и библиографической культуры с применением информационно-коммуникационных технологий и с учетом основных требований	Уровень 2 (углубленный) (ОПК-4)-1	Декомпозиция 1 Знать: методы сбора, обработки и хранения данных, а также основные методы формирования научного знания из данных З2(I) (ОПК-4) Уметь: использовать научные и методические ресурсы сети интернет для разработки программного обеспечения и программной документации с учетом требований информационной безопасности У2(I) (ОПК-4) Владеть: компьютерными средствами для решения прикладных задач в профессиональной деятельности В2(I) (ОПК-4)

информационной безопасности		
ПК-1 способностью собирать, обрабатывать и интерпретировать данные современных научных исследований, необходимые для формирования выводов по соответствующим научным исследованиям	ПК-1	Декомпозиция 4 Знать: современные методы сбора, хранения, обработки, анализа и интерпретации данных З(IV) (ПК-1) Уметь: строить гипотезы и применять математические методы для построения практически значимых моделей У(IV) (ПК-1) Владеть: инструментальными средствами анализа данных и навыками их применения к практическим задачам В(IV) (ПК-1)

4. ОБЪЕМ И СТРУКТУРА ДИСЦИПЛИНЫ

Трудоемкость дисциплины составляет 4 зачетные единицы, 144 часа, из которых 36 часов самостоятельной работы обучающегося отводится на подготовку и защиту экзамена.

Форма промежуточной аттестации: экзамен в 7 семестре, курсовая работа 7 семестр.

Распределение фонда времени по темам дисциплины представлено в таблице 4.1.

Таблица 4.1 – Распределение фонда времени по темам дисциплины (очная форма обучения)

Номер и наименование тем	Объем дисциплины (ак. часы)			
	Контактная работа			СРО
	ЗЛТ	ПЗ	ЛР	
1	2	3	4	5
Тема 1. Введение. Постановка задачи анализа данных. Области применения методов анализа данных	2	-	-	8
Тема 2. Линейные модели в задачах анализа данных	4	8	-	12
Тема 3. Метод опорных векторов	4	10	-	12
Тема 4. Метрические методы классификации и регрессии	4	8	-	12
Тема 5. Логические методы классификации и регрессии	6	12	-	16
Тема 6. Композиции алгоритмов и градиентный бустинг	4	10	-	12
Всего по дисциплине:	**24**	**48**		**72**

*ЗЛТ – занятия лекционного типа, ПЗ – все виды занятий семинарского типа, кроме лабораторных работ, ЛР – лабораторные работы, СРО – самостоятельная работа обучающегося

5. СОДЕРЖАНИЕ РАЗДЕЛОВ И ТЕМ ДИСЦИПЛИНЫ

Тема 1. Введение. Постановка задачи анализа данных. Области применения методов анализа данных

Постановка задачи анализа данных. Методы проведения анализа данных. Предмет машинного обучения. Типы задач машинного обучения. Примеры прикладных задач. Порядок построения алгоритма решения. Понятия функции потерь и функционала качества. Сведение задачи к задаче оптимизации. Обобщающая способность алгоритма. Методы оценки качества моделей. Метрики качества.

Тема 2. Линейные модели в задачах анализа данных

Формулировка оптимизационной задачи линейной регрессии. Функционалы качества в линейных моделях классификации и регрессии. Методы обучения линейных

предсказательных моделей. Предпосылки метода наименьших квадратов. Аналитическое решение метода наименьших квадратов. Условия существования аналитического решения. Полиномиальная регрессия для задач классификации и регрессии. Проблема переобучения. Методы борьбы с переобучением. Регуляризация как метод повышения обобщающей способности модели. Гребневая и LASSO-регрессия.

Логистическая регрессия в задачах классификации. Применение логистической регрессии для анализа текстовых данных. Предположение о законе распределения целевой переменной. Обобщенная линейная модель. Понятие функции связи. Функционалы качества в задачах нелинейной регрессии.

Тема 3. Метод опорных векторов

Понятие отступа. Разделяющая гиперплоскость. Определение функции потерь в задаче классификации. Оценка сверху для пороговой функции потерь. Формулировка задачи линейного программирования для нахождения оптимальной разделяющей полосы. Линейно разделимый случай. Решение задачи классификации методом множителей Лагранжа. Типы двойственных переменных в методе опорных векторов. Ядра и спрямляющие пространства. Решение задачи регрессии методом опорных векторов. Определение коэффициентов линейной модели через двойственные переменные.

Тема 4. Метрические методы классификации и регрессии

Понятия компактности и функции расстояния в метрическом классификаторе. Обобщенный метрический классификатор. Виды функций расстояния в прикладных задачах. Особенности метрических методов. Методы масштабирования исходных данных. Метрические методы в задачах регрессии.

Метод ближайших соседей для решения задач классификации и регрессии. Метод окна Парзена. Понятие ширины окна. Выбор функции ядра. Метод окна Парзена переменной ширины. Вычислительная сложность метрических методов. Проклятие размерности. Алгоритмы отбора эталонных объектов.

Тема 5. Логические методы классификации и регрессии

Понятие логической закономерности. Конъюнкции пороговых правил. Бинарное решающее дерево. Понятие критерия информативности. Ошибка классификации. Энтропийный критерий. Критерий Джини. Выбор оптимального разбиения. Алгоритмы построения решающего дерева: ID3, C4, C4.5. Обработка пропусков в деревьях решений. Методы обработки категориальных переменных. Кодирование категориальных переменных. One-hot-encoding. Кодирование двоичным представлением. Кодирование статистическими методами. Пороговые правила для вещественных переменных. Регуляризация решающих деревьев. Алгоритм CART. Деревья регрессии. Дисперсионный критерий. Оценка наименьших квадратов. Выбор глубины решающего дерева.

Тема 6. Композиции алгоритмов и градиентный бустинг

Понятие композиции алгоритмов. Теорема Кондорсе о жюри присяжных. Оценка распределения методом бутстрэп. Построение доверительного интервала. Оценка качества на небольших выборках. Метод бэггинга Бреймана. Метод случайных подпространств. Случайный лес в задачах классификации и регрессии. Обобщающая способность случайного леса.

Градиентный бустинг. Базовый алгоритм. Алгоритм AdaBoost. Метод градиентного спуска в задачах классификации и регрессии. Регуляризация в методе градиентного бустинга. Преобразование данных для улучшения качества предсказаний. Настройка гиперпараметров модели.

6. ЗАНЯТИЯ СЕМИНАРСКОГО ТИПА

Таблица 6.1 – Практические занятия / Семинарские занятия \ Лабораторные работы

№ темы	Тема занятия	Вид занятия / Оценочное средство
1	2	3
2	Занятие 1. Вывод формулы обучения модели линейной регрессии. Построение модели численными методами. Построение модели полиномиальной регрессии. Анализ коэффициентов при переменных модели. Добавление регуляризатора в модель. Оценка качества модели регрессии на отложенной выборке и на кросс-валидации. Занятие 2. Обработка текстовых данных. Вывод формулы обучения модели логистической регрессии для задачи проставления меток текстового документа. Построение математической модели проставления меток и оценка качества результатов.	ПЗ: Расчетно-графическая работа
3	Занятие 3. Построение оптимальной разделяющей полосы методом опорных векторов. Определение коэффициентов линейной модели. Добавление регуляризатора в модель. Исследование зависимости полученного классификатора от параметра регуляризации. Выбор оптимального параметра регуляризации. Занятие 4. Выбор функции ядра для построения разделяющей полосы в случае, когда в исходном пространстве классы не являются линейно разделимыми. Визуализация полученных результатов.	ПЗ: Расчетно-графическая работа КТ№1
4	Занятие 5. Решение задачи классификации методом ближайших соседей. Выбор функции расстояния и числа соседей по критерию скользящего контроля. Решение задачи регрессии методом ближайших соседей. Определение функции расстояния в задаче классификации текстовых документов. Занятие 6. Решение задачи классификации методом окна Парзена фиксированной ширины. Выбор ширины окна по критерию скользящего контроля. Исследование зависимости качества классификации от выбора функции ядра.	ПЗ: Расчетно-графическая работа
5	Занятие 7. Построение бинарного решающего дерева. Выбор предиката в корневой вершине. Оценка прироста информации всех возможных разбиений с помощью энтропийного критерия. Оценка прироста информации всех возможных разбиений с помощью критерия Джини. Сравнение полученных результатов. Занятие 8. Построение дерева регрессии. Выбор пороговых правил. Оценка прироста информативности всех возможных разбиений с помощью дисперсионного критерия. Оценка наименьших квадратов прироста информативности. Определение оптимального начального разбиения. Занятие 9. Обработка пропусков в данных. Применение методов кодирования категориальных переменных в задачах анализа данных. Применение метода one-hot-encoding для кодирования категориальных переменных. Кодирование двоичным представлением категорий. Сравнение качества алгоритмов.	ПЗ: Расчетно-графическая работа
6	Занятие 10. Решение задачи о жюри присяжных. Формирование выборки методом бутстрэп. Оценка параметров распределения методом бутстрэп. Построение доверительного интервала. Решение задачи классификации методом бэггинга над решающими деревьями. Занятие 11. Решение задачи классификации построения случайного леса. Выбор числа деревьев. Сравнение качества на обучающей и отложенной выборках в зависимости от числа деревьев. Решение задачи регрессии алгоритмом случайного леса.	ПЗ: Расчетно-графическая работа КТ№2

	Занятие 12. Решение задачи классификации алгоритмом AdaBoost. Оценка качества построенного алгоритма. Реализация алгоритма градиентного бустинга для задачи регрессии. Выбор гиперпараметров модели.	

*ПЗ – практические занятия, СЗ – семинарские занятия, ЛР – лабораторные работы

7. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ДЛЯ ОБУЧАЮЩИХСЯ

7.1. Методические указания для обучающегося по освоению дисциплины

Для формирования четкого представления об объеме и характере знаний и умений, которыми надо будет овладеть по дисциплине в самом начале учебного курса обучающийся должен ознакомиться с учебно-методической документацией:

– рабочей программой дисциплины: с целями и задачами дисциплины, ее связями с другими дисциплинами образовательной программы, перечнем знаний и умений, которыми в процессе освоения дисциплины должен владеть обучающийся,

– порядком проведения текущего контроля успеваемости и промежуточной аттестации;

– графиком консультаций преподавателей кафедры.

Систематическое выполнение учебной работы на занятиях лекционных и семинарских типов, а также выполнение самостоятельной работы позволит успешно освоить дисциплину.

В процессе освоения дисциплины обучающимся следует:

– слушать, конспектировать излагаемый преподавателем материал;

– ставить, обсуждать актуальные проблемы курса, быть активным на занятиях;

– задавать преподавателю уточняющие вопросы с целью уяснения теоретических положений;

– выполнять задания практических занятий полностью и установленные сроки.

При затруднениях в восприятии материала следует обратиться к основным литературным источникам. Если разобраться в материале не удалось, то обратится к лектору (по графику его консультаций) или к преподавателю на занятиях семинарского типа.

Обучающимся, пропустившим занятия (независимо от причин), не имеющим письменного решения задач или не подготовившимся к данному занятию, рекомендуется не позже, чем в 2 - недельный срок явиться на консультацию к преподавателю и отчитаться по теме.

7.2. Организация самостоятельной работы

Под самостоятельной работой обучающихся понимается планируемая работа обучающихся, направленная на формирование указанных компетенций, выполняемая во внеаудиторное время по заданию и при методическом руководстве преподавателя, без его непосредственного участия.

Методическое обеспечение самостоятельной работы при наличии обучающихся лиц с ограниченными возможностями представляется в формах, адаптированных к ограничениям их здоровья.

Виды самостоятельной работы по дисциплине представлены в таблице 7.2.1.

Таблица 7.2.1 – Организация самостоятельной работы обучающегося

№ темы	Вид самостоятельной работы
1	2
1	Изучение литературы по математическим методам анализа данных и машинного обучения.
2	Самостоятельное задание: построение модели линейной регрессии в Wolfram Mathematica и Python, анализ полученных результатов. Построение модели полиномиальной регрессии. Выбор оптимальной линейной модели. Построение математической модели проставления меток для текстовых данных. Подготовка к экзамену.
3	Самостоятельное задание: решение задачи классификации методом опорных векторов в Wolfram Mathematica и Python. Выбор оптимального параметра регуляризации. Построение классификатора методом опорных векторов для линейно неразделимого случая. Подготовка к экзамену.
4	Самостоятельное задание: решение задач классификации и регрессии методом ближайших соседей в Wolfram Mathematica и Python. Выбор функции расстояния и числа соседей по критерию скользящего контроля. Решение задачи классификации методом парзеновского окна фиксированной ширины. Выбор ширины окна и функции ядра. Подготовка к экзамену.
5	Самостоятельное задание: компьютерная реализация решающего дерева для задач классификации и регрессии в Wolfram Mathematica. Улучшение качества решения путем дополнительных преобразований над данными. Построение решающего дерева в Python, выбор гиперпараметров. Подготовка к экзамену.
6	Самостоятельное задание: оценка параметров распределения методом бутстрэп в Wolfram Mathematica. Решение задачи классификации методом бэггинга над решающими деревьями в Python. Построение алгоритма случайного леса. Исследование зависимости качества решения от числа деревьев. Решение задачи классификации алгоритмом градиентного бустинга, выбор гиперпараметров алгоритма. Подготовка к экзамену.

Каждый вид СРО, указанный в таблице 7.2.1 обеспечен методическими материалами.

8. ОБРАЗОВАТЕЛЬНЫЕ ТЕХНОЛОГИИ

В преподавании дисциплины «Методы анализа данных» используются как традиционные образовательные технологии, так и технологии, предусматривающие активные и интерактивные методы обучения.

Все практические занятия проводятся в компьютерной аудитории в форме выполнения расчетно-графических работ. Активно используются обсуждения вариантов постановки задач, методов решения, возможных подходов к их компьютерной реализации. Совместно с обучающимися разрабатывается пошаговый план решения поставленной задачи. Решение ИЗК (индивидуальное задание, выполняемое на компьютере) реализуется средствами Wolfram Mathematica и Python.

Проводится обсуждение адекватности вариантов полученного решения. Проводится совместное обсуждение корректности и возможности применения метода к другим практическим задачам.

9. РЕСУРСНОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

9.1. Учебно-методическое и информационное обеспечение дисциплины

Таблица 9.1.1 – Учебно-методическое обеспечение дисциплины

Библиографическое описание издания (автор, заглавие, вид, место и год издания, кол. стр.)	Основная/ дополнительная литература	Книгообеспеченность	
		Кол-во. экз. в библ. СПбГЭУ	Электронные ресурсы
Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике : учебное пособие .— 11-е изд., перераб. и доп. — Москва : Юрайт, 2019 .— 404 с. — Имеются другие года издания. — Сведения также доступны по Интернету: ЭБС Юрайт.	основная	160	ЭБС Юрайт
Гуриков С.Р. Основы алгоритмизации и программирования на Python : учебное пособие. — Электрон. дан. – М.: ФОРУМ: ИНФРА-М, 2019. – 343 с.	основная	–	ЭБС ZNANIUM
Статистические методы анализа данных : учебник / Л.И. Ниворожкина, С.В. Арженовский, А.А. Рудяга [и др.] ; под общ. ред. Л.И. Ниворожкиной . — Электрон. дан. - М.; Издательский Центр РИОР: ООО "Научно-издательский центр ИНФРА-М", 2016. - 333 с.	дополнительная	–	ЭБС ZNANIUM
Анализ данных : учебник / В.С. Мхитарян [и др.] ; под ред. В.С. Мхитаряна. — Электрон. дан. — М. : Издательство Юрайт, 2019 — 490 с.	дополнительная	–	ЭБС Юрайт

Таблица 9.1.2 – Перечень современных профессиональных баз данных (СПБД)

№	Наименование СПБД
1	Электронная библиотека Grebennikon.ru – www.grebennikon.ru
2	Научная электронная библиотека eLIBRARRY – www.elibrary.ru
3	Научная электронная библиотека КиберЛеника – www.cyberleninka.ru
4	База данных ПОЛПРЕД Справочники – www.polpred.com
5	База данных OECD Books, Papers & Statistics на платформе OECD iLibrary – www.oecd-ilibrary.org

Таблица 9.1.3 – Перечень информационных справочных систем (ИСС)

№	Наименование ИСС
1	Справочная правовая система КонсультантПлюс (инсталлированный ресурс СПбГЭУ или www.consultant.ru)
2	Справочная правовая система «ГАРАНТ» (инсталлированный ресурс СПбГЭУ или www.garant.ru)
3	Информационно-справочная система «Кодекс» (инсталлированный ресурс СПбГЭУ или www.kodeks.ru)
4	Электронная библиотечная система BOOK.ru - www.book.ru

5	Электронная библиотечная система ЭБС ЮРАЙТ – www.urait.ru
6	Электронно-библиотечная система ЗНАНИУМ (ZNANIUM) – www.znanium.com
7	Электронная библиотека СПбГЭУ– opac.unecon.ru

9.2. Материально-техническое обеспечение учебного процесса

Для реализации данной дисциплины имеются специальные помещения для проведения занятий лекционного типа, занятий семинарского типа, курсового проектирования (выполнения курсовых работ), групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации, а также помещения для самостоятельной работы.

Таблица 9.2.1 – Перечень программного обеспечения (ПО)

№ п/п	Наименование ПО
1	Microsoft Windows Professional (КОНТРАКТ № 244/20 «26» июня 2020 г)
2	Microsoft Office Professional (КОНТРАКТ № 244/20 «26» июня 2020 г)
3	7-Zip (freeware)
4	Wolfram Mathematica 12 (Сублицензионный договор № 328/20 от 20.08.2020г.)
5	Python 3.7.4 (freeware)

10. ОСОБЕННОСТИ ОСВОЕНИЯ ДИСЦИПЛИНЫ ДЛЯ ИНВАЛИДОВ И ЛИЦ С ОГРАНИЧЕННЫМИ ВОЗМОЖНОСТЯМИ ЗДОРОВЬЯ

Обучение обучающихся с ограниченными возможностями здоровья при необходимости осуществляется на основе адаптированной рабочей программы с использованием специальных методов обучения и дидактических материалов, составленных с учетом особенностей психофизического развития, индивидуальных возможностей и состояния здоровья таких обучающихся (обучающегося).

В целях освоения учебной программы дисциплины инвалидами и лицами с ограниченными возможностями здоровья Университет обеспечивает:

– для инвалидов и лиц с ограниченными возможностями здоровья по зрению: размещение в доступных для обучающихся, являющихся слепыми или слабовидящими, местах и в адаптированной форме справочной информации о расписании учебных занятий; присутствие ассистента, оказывающего обучающемуся необходимую помощь; выпуск альтернативных форматов методических материалов (крупный шрифт или аудиофайлы);

– для инвалидов и лиц с ограниченными возможностями здоровья по слуху: надлежащими звуковыми средствами воспроизведение информации;

– для инвалидов и лиц с ограниченными возможностями здоровья, имеющих нарушения опорно-двигательного аппарата: возможность беспрепятственного доступа обучающихся в учебные помещения, туалетные комнаты и другие помещения кафедры, а также пребывание в указанных помещениях.

Образование обучающихся с ограниченными возможностями здоровья может быть организовано как совместно с другими обучающимися, так и в отдельных группах или в отдельных организациях.

11. ФОНД ОЦЕНОЧНЫХ СРЕДСТВ ДЛЯ ПРОВЕДЕНИЯ ТЕКУЩЕЙ И ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ ОБУЧАЮЩИХСЯ ПО ДИСЦИПЛИНЕ

Фонд оценочных средств для проведения аттестации уровня сформированности компетенций обучающихся по дисциплине оформляется отдельным документом и является приложением к рабочей программе дисциплины.