

**Санкт-Петербургский филиал федерального государственного
автономного образовательного учреждения высшего образования
"Национальный исследовательский университет
"Высшая школа экономики"**

Факультет Санкт-Петербургская школа экономики и менеджмента
Национального исследовательского университета
«Высшая школа экономики»

Департамент прикладной математики и бизнес-информатики

**Рабочая программа дисциплины
Математические основы анализа данных**

для образовательной программы «Анализ больших данных в бизнесе, экономике и обществе»
направления подготовки 01.04.02 «Прикладная математика и информатика»
уровень магистратура

Разработчик(и) программы
Сироткин А.В., к.ф.-м.н., доцент, avsirotkin@hse.ru

Согласована менеджером ОП «Анализ больших данных в бизнесе, экономике и обществе»
«30» августа 2016г.

Е.С. Авдониной _____

Утверждена Академическим руководителем образовательной программы

А.В. Сироткин _____

«30» августа 2016г.

Санкт-Петербург, 2016

Настоящая программа не может быть использована другими подразделениями университета и другими вузами без разрешения кафедры-разработчика программы.



1 Область применения и нормативные ссылки

Настоящая рабочая программа дисциплины устанавливает минимальные требования к знаниям и умениям студента, а также определяет содержание и виды учебных занятий и отчетности.

Программа предназначена для преподавателей, ведущих дисциплину «Математические основы анализа данных», учебных ассистентов и студентов направления подготовки 01.04.02 «Прикладная математика и информатика», обучающихся по образовательной программе «Анализ больших данных в бизнесе, экономике и обществе».

Рабочая программа дисциплины разработана в соответствии с:

- Образовательным стандартом НИУ ВШЭ
<http://www.hse.ru/data/2016/11/02/1111123560/01.04.02%20Прикладная%20математика%20и%20информатика.pdf>;
- Образовательной программой «Анализ больших данных в бизнесе, экономике и обществе», направление подготовки 01.04.02 «Прикладная математика и информатика»;
- Объединенным учебным планом университета по образовательной программ «Анализ больших данных в бизнесе, экономике и обществе».

2 Цели освоения дисциплины

Целями освоения дисциплины "Математические основы анализа данных " являются

- изучение продвинутых и углублённых методов современной теории вероятностей, теории случайных процессов, математической статистики, в том числе и статистики случайных процессов, а так же методов линейной алгебры в задачах оптимизации и регрессии.
- формирование представления об аналитических методах теории вероятностей, математической статистики и теории случайных процессов, о моделях, лежащих в основе распределений случайных процессов,
- освоение ряда алгоритмов для моделирования случайных процессов и статистической оценки параметров процессов.

3 Компетенции обучающегося, формируемые в результате освоения дисциплины

В результате освоения дисциплины студент осваивает следующие компетенции:

Компетенция	Код по ОС ВШЭ	Уровень формирования компетенции	Дескрипторы – основные признаки освоения (показатели достижения результата)	Формы и методы обучения, способствующие формированию и развитию компетенции	Форма контроля уровня сформированности компетенции
Профессиональные компетенции					
Инструментальные компетенции					
Способен анализировать и воспроизводить смысл междисциплинарных текстов с использованием языка и аппарата прикладной	ПК-10	РБ, СД	Способен читать научные статьи и литературу и использовать методы, описанные там, для решения прикладных задач.	Семинарские занятия, домашние задания, самостоятельная работа студентов	Домашняя работа



Компетенция	Код по ОС ВШЭ	Уровень формирования компетенции	Дескрипторы – основные признаки освоения (показатели достижения результата)	Формы и методы обучения, способствующие формированию и развитию компетенции	Форма контроля уровня сформированности компетенции
математики.					
Способен создавать междисциплинарные тексты с использованием языка и аппарата прикладной математики.	ПК-11	РБ, СД	Способен описать результаты и процесс решения поставленной задачи в виде научного отчета.	Семинарские занятия, домашнее задание	Домашняя работа
Способен описывать проблемы и ситуации профессиональной деятельности, используя язык и аппарат прикладной математики при решении междисциплинарных проблем.	ПК-17	РБ, СД	Умеет сводить прикладные задачи к задачам оптимизации и регрессии.	Лекции, семинарские занятия	Домашняя работа, контрольная работа, экзамен
Способен понимать и применять в исследовательской и прикладной деятельности современный математический аппарат.	ПК-18	РБ, СД	Решает задачи с привлечением аппарата теории вероятностей. Умеет применять регрессионные модели к реальным данным	Лекции, семинарские занятия	Домашняя работа, контрольная работа, экзамен

4 Место дисциплины в структуре образовательной программы

Настоящая дисциплина относится к базовой части цикла дисциплин магистерской программы «Анализ больших данных в бизнесе, экономике и обществе».

Изучение данной дисциплины базируется на следующих дисциплинах обучения в бакалавриате: «Математический анализ», «Теория вероятностей и математическая статистика», «Основы математических исследований».

Для освоения учебной дисциплины, студенты должны владеть следующими знаниями и компетенциями:

- способен анализировать и воспроизводить смысл междисциплинарных текстов с использованием языка и аппарата прикладной математики;
- способен создавать междисциплинарные тексты с использованием языка и аппарата прикладной математики;
- способен описывать проблемы и ситуации профессиональной деятельности, используя язык и аппарат прикладной математики при решении междисциплинарных проблем;
- способен понимать и применять в исследовательской и прикладной деятельности современный математический аппарат.

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

- «Современные методы анализа данных»,
- «Финансовые рынки и финансовая математика»,
- «Анализ социальных и экономических сетей»,



- «Продвинутые методы эконометрики»

5 Тематический план учебной дисциплины

ОБЪЕМ ДИСЦИПЛИНЫ - 3 зачетные единицы.

№	Название раздела	Всего часов	Аудиторные часы		Самостоятельная работа
			Лекции	Семинары	
1	Элементы линейной алгебры в задачах оптимизации и регрессии	24	4	4	16
2	Классические статистические методы	44	6	6	32
3	Случайные процессы	46	6	6	34
ИТОГО		114	16	16	82

6 Формы контроля знаний студентов

Тип контроля	Форма контроля	1 модуль								Параметры
		Сентябрь				Октябрь				
Текущий	Домашняя работа						*			Домашняя работа в форме написания программ решающих поставленные задачи.
	Контрольная работа							*		Письменная работа продолжительностью 80 минут.
Итоговый	Экзамен							*		Письменный экзамен - 120 минут

7 Критерии оценки знаний, навыков

Критерии оценки контрольной работы:

Контрольная работа ($O_{кр}$) проводится в письменной форме и содержит не менее 4 задач. Максимальный балл за отдельную задачу указывается для каждой задачи индивидуально и может быть от 1 до 3 в зависимости от сложности задачи. Сумма максимальных баллов за все задачи равна 10. Полный балл выставляется за полное решение задачи, при этом допускаются незначительные арифметические ошибки. Частичный балл (1 для задач с максимальной оценкой 2 балла или 1 или 2 для задач с максимальной оценкой 3 балла) выставляется за решение, содержащее ключевые элементы, но либо не доведенное до конца, либо содержащее существенные ошибки. 0 баллов выставляется при отсутствии решения или за решения, содержащее грубые ошибки, показывающие, что студент не владеет методами решения данной задачи. Итоговая оценка за контрольную представляет собой сумму баллов за все задачи. Максимальная оценка за контрольную работу – **10 баллов**.



Критерии оценки домашней работы:

Домашняя работа ($O_{др}$) представляет собой написание программ решающих поставленные задачи. Домашняя работа состоит из двух задач, решение каждой из которых оценивается от 0 до 5 баллов.

5 баллов соответствует полностью решенной задаче, проходящей все подготовленные преподавателем тесты.

3-4 балла выставляется за задачу, если она проходит большую часть тестов и в целом решена с пониманием правильного метода решения, но допущены незначительные ошибки, при исправлении которых будут пройдены все тесты.

1-2 балла выставляется за задачу, которая корректно проходит ряд тестов и содержит частичное решение поставленной задачи, но со значительными ограничениями по условиям применения.

0 баллов выставляется при отсутствии решения, или если решение не проходит ни одного теста.

Оценка за домашнюю работу представляет собой сумму баллов за каждую из двух задач. Максимальная оценка за домашнюю работу. – **10 баллов.**

Критерии оценки за экзамен:

Экзамен ($O_{экз}$) проводится в письменной форме и содержит 5 задач, охватывающих основные пройденные темы.

Каждая задача оценивается в 0, 1 или 2 балла.

2 балла за задачу ставится если она полностью решена.

1 балл ставится за задачу, если в решении приведены ключевые моменты для решения задачи или указан метод которым задача должна быть решена, но решение не доведено до конца.

0 баллов выставляется при отсутствии решения или за решения содержащее грубые ошибки, показывающие, что студент не владеет методами решения данной задачи.

Итоговая оценка за экзамен выставляется как сумма баллов по всем пяти задачам. Максимальная оценка за экзамен – **10 баллов.**

8 Содержание дисциплины

1. Раздел 1 Элементы линейной алгебры в задачах оптимизации и регрессии

Векторы, матрицы, матрично-векторные операции. проекция. Приближенное решение систем линейных уравнений. Собственные вектора и собственные значения. Разложения матриц. Задача регрессии, как задача оптимизации и ее решение в матричном виде.

2. Раздел 2 Классические статистические методы

Сравнение гипотез. Методы параметрического и непараметрического оценивания случайных величин. p-value и его применение в проверке гипотез. Поправки на множественные сравнения. Смеси распределений и оценки их параметров.

3. Раздел 3 Случайные процессы.

Пуассоновский процесс. Марковские и не Марковские процессы. Случайные блуждания. Марковские цепи. Стационарные распределения. Процессы на графах. МСМС-методы, семплирование и алгоритм Метрополисс-Гастингсон.

9 Образовательные технологии

Для проведения занятий со студентами используются:

- презентационные технологии;
- компьютерные технологии.

9.1 Методические рекомендации преподавателю



9.2 Методические указания студентам по освоению дисциплины

При выполнении домашней работы важно обратить внимание на крайние случаи условий. Проверьте, что ваши решения корректно работают на больших примерах.

9.3 Учебно-методическое обеспечение для самостоятельной работы студентов по дисциплине

При самостоятельной работе рекомендуется дополнительно к основной программе решать задачи расположенные на <https://projecteuler.net>. Данные задачи сочетают аналитические способности и знания математики с базовыми навыками программирования, что способствует лучшему усвоению.

Дополнительно, для поддержки самостоятельной работы можно воспользоваться свободно доступным курсом "Ликбез по дискретной математике" ([https:// stepik.org/course/Ликбез-по-дискретной-математике-91](https://stepik.org/course/Ликбез-по-дискретной-математике-91)), отдельные разделы которого позволят глубже понять основной курс.

10 Оценочные средства для текущего контроля и аттестации студента

10.1 Оценочные средства для оценки качества освоения дисциплины в ходе текущего контроля

Примеры заданий для Контрольной работы.

1. [2 балла] У вас есть 2 монетки не различимые по внешним признакам. Одна честная и выпадает орлом с вероятностью 0.5. Вторая монетка выпадает орлом с вероятностью p , которую вы не знаете. Вы случайно выбрали одну монетку и подбросили ее 10 раз, после чего взяли оставшуюся монетку и тоже подбросили ее 10 раз. Найдите оценку максимального правдоподобия для параметра p , если в первом случае выпало 4 орла, а во втором -- 7.

2. [1 балл] При каких условиях на элементы матрицы перехода, Марковская сеть с двумя состояниями будет эргодической.

3. [2 балла] Найдите уравнение регрессионной прямой для точек (2,1), (3, 2) и (4, 4).

4. [3 балла] Болезнь XOR очень заразна и имеет две формы 0 и 1. Известно, что если два человека дружат и один из них болен, то время до заражения второго распределено по экспоненциальному (показательному) закону, с плотностью $f(t) = \lambda e^{-\lambda t}$. При заражении тип болезни меняется, то есть если заразивший имел тип 0, то зараженный будет иметь тип 1 и наоборот. Интенсивность распространения носителем с типом 1 в три раза больше, то есть $\lambda_1 = 3\lambda_0$.

Известно, что в момент времени T у Гриши обнаружили XOR (заразиться он мог раньше).

Считая за момент времени 0, тот когда Миша заразился типом 0. Определите тип XOR у Гриши, если известно, что Миша и Гриша дружат, имеют двух общих друзей, которые не дружат между собой и никаких других путей заражения, кроме как от Миши или двух общих друзей, у Гриши не было. Все заражения независимы и повторно заразиться нельзя.

5. [2 балла] Найдите проекцию вектора (4,0,1) на плоскость задаваемую парой векторов (0,0,1) и (1, 1, 0).

Примеры заданий для Домашней работы.

1. Реализуйте алгоритм Метрополисс-Гастингсон, для семплирования случайных графов Эрдеша-Реньи, для заданного числа вершин. Определите долю графов содержащих ровно три ребра и сравните полученный результат с теоретической оценкой. При условии, что $p=0.7$ (вероятность существования ребра), а $n=6$ (число вершин)

2.

10.2 Примеры заданий итогового контроля

Примеры экзаменационных заданий:

1. Винни-Пух и Пятачок очень любят ходить в гости. Если Винни-Пух и Пятачок гуляют, то они с вероятностью 0.4 решат погулять еще немного, а с вероятностью 0.6 пойдут в гости к Кролику. Если же Винни и Пятачок уже у Кролика, то с вероятностью 0.3 они пойдут гулять, а с вероятностью 0.7 посидят еще немного.

Известно, что Винни и Пятачок уже очень давно ведут себя как описано выше, какова вероятность, что они сейчас в гостях у Кролика.

2. Найдите формулу регрессионной прямой для точек (1,3), (2,4), (6, 4), (0,0).

3. Есть три человека: Вася, Петя и Коля. Для любой пары вероятность что они дружат равна 0.6. Вася узнал интересную новость про экзамен. Если человек знает новость, то вечером он расскажет ее своему другу с вероятностью 0.5, при этом новость нельзя пересказать в тот же вечер когда она была узнана. Известно, что после третьего вечера Коля знает новость про экзамен, какова вероятность, что в графе дружбы ровно 2 ребра.

4. Вам предложили поиграть в простую карточную игру. Ведущий тасует колоду из трех карт, двух черной масти и одной красной, ваша задача угадать где красная. Вы начинаете эту игру предполагая, что маловероятно ($p=0.1$), что перед вами жулик (в игре с жуликом вы не можете выиграть). Однако после трех неудач, вы начинаете сомневаться. Какова в этот момент оценка вероятности, что перед вами жулик? А еще после трех неудач? Через сколько неудачных попыток, вы начнете считать, что перед вами скорее жулик, чем нет ($p > 0.5$).

5. Найдите минимальное количество подбрасываний монетки N за которое вы сможете отличить честную монетку от монетки с вероятностью выпадения орла 0.95, ошибаясь не более чем в 12.5% случаев. Укажите N и критерий. Обоснуйте свое решение.

11 Порядок формирования оценок по дисциплине

Накопленная оценка по дисциплине рассчитывается с помощью взвешенной суммы оценок за отдельные формы текущего контроля знаний следующим образом:

$$O_{\text{накопленная}} = 0,5O_{\text{дз}} + 0,5O_{\text{кр}}, \text{ где}$$

$O_{\text{дз}}$ - оценка знаний студента за домашнюю работу;

$O_{\text{кр}}$ – оценка знаний студента за контрольную работу;

Результирующая оценка по дисциплине (которая идет в диплом) рассчитывается следующим образом:

$$O_{\text{результ}} = 0,5O_{\text{накопленная}} + 0,5O_{\text{экз}}, \text{ где}$$

$O_{\text{накопленная}}$ - накопленная оценка по дисциплине;

$O_{\text{экз}}$ - оценка за экзамен.

В формулу для $O_{\text{результ}}$ подставляются значения $O_{\text{накопленная}}$ и $O_{\text{экз}}$, округленные по арифметическим правилам. $O_{\text{результ}}$ округленные по арифметическим правилам.

По усмотрению ведущего преподавателя, если это не противоречит действующим документам на момент экзамена, при получении накопленной оценки 8 баллов и более, студент может быть освобожден от экзамена. В таком случае, с согласия студента, ему выставляется результирующая оценка, равная накопленной.

Студент не получает возможность пересдать низкие результаты за домашнюю работу и/или работу на семинарских или контрольную работу, а также при пропуске соответствующих им учебных часов.

При получении неудовлетворительной оценки $O_{\text{результ}}$ (значение после округления менее 4 баллов) выставляется оценка «НЕУДОВЛЕТВОРИТЕЛЬНО».



12 Учебно-методическое и информационное обеспечение дисциплины

12.1 Основная литература

Шевцов Г.С. Линейная алгебра: теория и прикладные аспекты: Учебное пособие [Электронный ресурс] / Г.С. Шевцов. - 3-е изд., испр. и доп. - М.: Магистр: НИЦ ИНФРА-М, 2014. - Режим доступа по паролю: <http://znanium.com/bookread2.php?book=438021> (Электронно-библиотечная система «ZNANIUM»)

12.2 Дополнительная литература

Murphy K. P. Machine learning: a probabilistic perspective [Electronic Resource] / Kevin Patrick Murphy. – Cambridge: Cambridge University Press, 2012. – 1098 p. - Authorized access: <http://site.ebrary.com/lib/hselibrary/detail.action?docID=10597102> (Online Digital Library "Ebrary").

12.3 Справочники, словари, энциклопедии

Не предусмотрены.

12.4 Ресурсы информационно-телекоммуникационной сети «Интернет»

1. <http://projecteuler.net>

12.5 Программные средства

Язык программирования Python

12.6 Информационные справочные системы

Не требуются.

12.7 Дистанционная поддержка дисциплины

Не предусматривается.

13 Материально-техническое обеспечение дисциплины

При проведении лекционных и семинарских занятий преподавателем может использоваться компьютер и мультимедийный проектор.