Приложение 4

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Нижегородский государственный университет
им. Н.И. Лобачевского»

Институт информационных технологий, математики и механики

УТВЕРЖДЕНО

решением Ученого совета ННГУ

протокол № 10 от 02.12.2024 г.

Рабочая программа дисциплины

Математические методы биоинформатики

Уровень высшего образования
Бакалавриат

Направление подготовки / специальность
02.03.02 - Фундаментальная информатика и информационные технологии

Направленность образовательной программы
Инженерия программного обеспечения

Форма обучения
очная

г. Нижний Новгород

2025 год начала подготовки

1. Место дисциплины в структуре ОПОП

Дисциплина Б1.В.ДВ.07.06 Математические методы биоинформатики относится к части, формируемой участниками образовательных отношений образовательной программы.

2. Планируемые результаты обучения по дисциплине, соотнесенные с планируемыми результатами освоения образовательной программы (компетенциями и индикаторами достижения компетенций)

Формируемые компетенции (код, содержание компетенции)	**Планируемые результаты обучения по дисциплине (модулю), в соответствии с индикатором достижения компетенции**		**Наименование оценочного средства**	
	Индикатор достижения компетенции (код, содержание индикатора)	**Результаты обучения по дисциплине**	**Для текущего контроля успеваемости**	**Для промежуточной аттестации**
ПК-3: Способен создавать и исследовать новые математические модели в естественных науках, промышленности и бизнесе, с учетом возможностей современных информационных технологий и программирования и компьютерной техники	*ПК-3.1: Знает методы анализа и исследования математических моделей в области фундаментальной информатики и информационных технологий* *ПК-3.2: Умеет определять ключевые свойства и ограничения системы*	*ПК-3.1: Знает базовые математические методы исследования для решения прикладных задач биоинформатики* *ПК-3.2: Умеет определять и профессионально реализовывать необходимые для решения прикладных задач биоинформатики математические методы*	*Задачи* *Собеседование*	*Зачёт: Контрольные вопросы* *Задачи*

3. Структура и содержание дисциплины

3.1 Трудоемкость дисциплины

	очная
Общая трудоемкость, з.е.	**3**
Часов по учебному плану	**108**
в том числе	
аудиторные занятия (контактная работа):	
- занятия лекционного типа	**24**
- занятия семинарского типа (практические занятия / лабораторные работы)	**0**
- КСР	**1**
самостоятельная работа	**83**
Промежуточная аттестация	**0** **Зачёт**

3.2. Содержание дисциплины

(структурированное по темам (разделам) с указанием отведенного на них количества академических часов и виды учебных занятий)

.

Наименование разделов и тем дисциплины	Всего (часы)	в том числе			
		Контактная работа (работа во взаимодействии с преподавателем), часы из них			Самостоятельная работа обучающегося, часы
		Занятия лекционного типа	Занятия семинарского типа (практические занятия/лабораторные работы), часы	Всего	
	ооо	ооо	ооо	ооо	ооо
Основы прикладной математической статистики.	33	6		6	27
Основы машинного обучения.	36	8		8	28
Анализ транскриптомных данных.	38	10		10	28
Аттестация	0				
КСР	1			1	
Итого	108	24	0	25	83

Содержание разделов и тем дисциплины

1. Основы прикладной математической статистики. – Описательные статистики. Случайные величины и их распределения. Основные дискретные и непрерывные распределения. Корреляционный анализ. Проверка статистических гипотез. Поправка на множественные сравнения. Основы прикладной статистики в Python. Построение диаграмм размаха, гистограмм, функций плотности вероятности, планок погрешностей, диаграмм рассеяния, тепловых карт.
2. Основы машинного обучения. – Основные классы задач машинного обучения. Анализ и предварительная обработка данных. Основные алгоритмы классификации и регрессии: метод k ближайших соседей, логистическая регрессия, машина опорных векторов, деревья решений и их ансамбли, нейронные сети. Линейная регрессия. Оценка качества обучения. Методы снижения размерности и отбора признаков. Методы борьбы с переобучением. Некоторые алгоритмы кластеризации: метод k-средних, DBSCAN, иерархическая кластеризация. Построение моделей машинного обучения в Python.
3. Анализ транскриптомных данных. – Секвенирование нового поколения. Экспериментальные подходы. Выравнивания и псевдовыравнивания. Анализ дифференциальной экспрессии. Анализ функционального обогащения. Визуализация данных секвенирования. Обработка и анализ данных секвенирования на Python и R.

4. Учебно-методическое обеспечение самостоятельной работы обучающихся

Самостоятельная работа обучающихся включает в себя подготовку к контрольным вопросам и заданиям для текущего контроля и промежуточной аттестации по итогам освоения дисциплины приведенным в п. 5.

Самостоятельная работа обучающихся включает в себя подготовку к контрольным вопросам и заданиям для текущего контроля и промежуточной аттестации по итогам освоения дисциплины.

1. Мхитарян В.С. Анализ данных: учебник для академического бакалавриата. – М.: Юрайт, 2018. – 490 с.
2. Маккинли У. Python и анализ данных. – М.: ДМК Пресс, 2015. – 481 с.
3. Мюллер А., Гвидо С. Введение в машинное обучение с помощью Python. Руководство для спе-циалистов по работе с данными. – Из-во Вильямс, 2017. – 480 с.
4. Korpelainen E., Tuimala J., Somervuo P., Huss M., Wong G. RNA-seq Data Analysis: A Practical Approach. – Chapman and Hall/CRC, 2014. – 322 p.
5. Доусон М. Программируем на Python. – 3-е изд. – СПб: Питер, 2014. – 416 с.
6. Плас Дж.В. Python для сложных задач: наука о данных и машинное обучение. - СПб.: Питер, 2018. — 576 с.: ил.
7. Неделько В.М. Основы статистических методов машинного обучения. - Новосибирск: Изд-во НГТУ, 2010. – 72 с.

5. Фонд оценочных средств для текущего контроля успеваемости и промежуточной аттестации по дисциплине (модулю)

5.1 Типовые задания, необходимые для оценки результатов обучения при проведении текущего контроля успеваемости с указанием критериев их оценивания:

5.1.1 Типовые задания (оценочное средство - Задачи) для оценки сформированности компетенции ПК-3:

Загрузите из файла heart.csv данные о сердечных заболеваниях. Выполните следующие задания (на языке программирования Python в Jupyter Notebook):

1) Вычислите описательные статистики для количественных признаков (среднее значение, медиана, мода, размах, дисперсия, среднеквадратичное отклонение, 1й/2й/3й квартили, межквартильный размах).

2) Постройте гистограммы для признаков age, trestbps, chol, thalach, oldpeak.

3) Постройте диаграммы размаха для признаков age, trestbps, chol, thalach, oldpeak.

4) Постройте на одном графике две кривые PDF (probability density function) для признака chol. Одна PDF для мужчин, другая – для женщин.

5) Для признаков, которые не были указаны в п.2-3, постройте полигоны частот.

6) Сгруппируйте данные по полу и вычислите для каждой группы среднее значение признака chol, применив функцию агрегации. Изобразите результаты в виде столбчатой диаграммы, где столбцы должны соответствовать полу, а высота столбцов - соответствующим средним значениям признака chol. Добавьте к каждому столбцу планку погрешности, отражающую среднеквадратичное отклонение.

7) Постройте следующие диаграммы рассеяния:

- trestbps от age
- chol от age
- thalach от age
- oldpeak от age

Изобразите точки на диаграммах разными цветами в зависимости от пола. Попробуйте визуально определить, коррелируют ли рассматриваемые переменные с возрастом. Проверьте свои предположения, вычислив коэффициенты корреляции Спирмена.

8) Проверьте признаки age, trestbps, chol, thalach, oldpeak на нормальность с помощью критерия Шапиро-Уилка.

Критерии оценивания (оценочное средство - Задачи)

Оценка	Критерии оценивания
зачтено	Выполнены все или большая часть этапов решения задачи или задача решена с незначительными недочетами. Результаты работы представлены преподавателю в срок.
не зачтено	Выполнены не все практические задания или выполнены не в полном объеме (представлено не полное описание этапов выполнения заданий, получен неверный ответ, результаты работы не представлены преподавателю).

5.1.2 Типовые задания (оценочное средство - Собеседование) для оценки сформированности компетенции ПК-3:

1. Какие существуют основные описательные статистики?

2. Что представляет собой диаграмма размаха («ящик с усами», boxplot)?

3. Что такое гистограмма и функция плотности вероятности?

4. Что такое корреляция и какие типы корреляции бывают?

5. В чём заключается разница между коэффициентами корреляции Пирсона и Спирмена?

6. Какие критерии нормальности Вам известны?

7. В чём заключается суть дисперсионного анализа (ANOVA)?

8. Что такое FDR-коррекция?

9. В чём заключается суть метода перекрёстного контроля?

10. В чём заключается суть метода k ближайших соседей?

11. В чём заключается суть логистической регрессии?

12. В чём заключается суть машины опорных векторов?

13. В чём заключается суть дерева решений?

14. Какие существуют способы объединения деревьев решений в ансамбль?

15. Что такое нейронная сеть?

16. Что такое глубокое обучение?

17. В чём заключается суть метода k средних?

18. Что такое иерархическая кластеризация?

19. Какие существуют экспериментальные подходы секвенирования нового поколения?

20. Какие существуют инструменты для выравнивания и псевдовыравнивания?

21. В чём заключается суть анализа дифференциальной экспрессии?

22. Какие инструменты можно использовать для анализа функционального обогащения набора генов?

Критерии оценивания (оценочное средство - Собеседование)

Оценка	Критерии оценивания
зачтено	Студент дал развернутый ответ на все вопросы без существенных ошибок.
не зачтено	При ответе студент допускает грубые ошибки в основном материале.

5.2. Описание шкал оценивания результатов обучения по дисциплине при промежуточной аттестации

Шкала оценивания сформированности компетенций

<table>
<tr><td rowspan="2">Уровень сформированности компетенций (индикатора достижения компетенций)</td><td>плохо</td><td>неудовлетворительно</td><td>удовлетворительно</td><td>хорошо</td><td>очень хорошо</td><td>отлично</td><td>превосходно</td></tr>
<tr><td colspan="2">не зачтено</td><td colspan="5">зачтено</td></tr>
</table>

Знания	Отсутствие знаний теоретического материала. Невозможность оценить полноту знаний вследствие отказа обучающегося от ответа	Уровень знаний ниже минимальных требований. Имели место грубые ошибки	Минимально допустимый уровень знаний. Допущено много негрубых ошибок	Уровень знаний в объеме, соответствующем программе подготовки. Допущено несколько негрубых ошибок	Уровень знаний в объеме, соответствующем программе подготовки. Допущено несколько несущественных ошибок	Уровень знаний в объеме, соответствующем программе подготовки. Ошибок нет.	Уровень знаний в объеме, превышающем программу подготовки.
Умения	Отсутствие минимальных умений. Невозможность оценить наличие умений вследствие отказа обучающегося от ответа	При решении стандартных задач не продемонстрированы основные умения. Имели место грубые ошибки	Продемонстрированы основные умения. Решены типовые задачи с негрубыми ошибками. Выполнены все задания, но не в полном объеме	Продемонстрированы все основные умения. Решены все основные задачи с негрубыми ошибками. Выполнены все задания в полном объеме, но некоторые с недочетами	Продемонстрированы все основные умения. Решены все основные задачи. Выполнены все задания в полном объеме, но некоторые с недочетами.	Продемонстрированы все основные умения. Решены все основные задачи с отдельными несущественными недочетами, выполнены все задания в полном объеме	Продемонстрированы все основные умения. Решены все основные задачи. Выполнены все задания, в полном объеме без недочетов
Навыки	Отсутствие базовых навыков. Невозможность оценить наличие навыков вследствие отказа обучающегося от ответа	При решении стандартных задач не продемонстрированы базовые навыки. Имели место грубые ошибки	Имеется минимальный набор навыков для решения стандартных задач с некоторыми недочетами	Продемонстрированы базовые навыки при решении стандартных задач с некоторыми недочетами	Продемонстрированы базовые навыки при решении стандартных задач без ошибок и недочетов	Продемонстрированы навыки при решении нестандартных задач без ошибок и недочетов	Продемонстрирован творческий подход к решению нестандартных задач

Шкала оценивания при промежуточной аттестации

Оценка		Уровень подготовки
зачтено	**превосходно**	Все компетенции (части компетенций), на формирование которых направлена дисциплина, сформированы на уровне не ниже «превосходно», продемонстрированы знания, умения, владения по соответствующим компетенциям на уровне выше предусмотренного программой
	отлично	Все компетенции (части компетенций), на формирование которых направлена дисциплина, сформированы на уровне не ниже «отлично».
	очень хорошо	Все компетенции (части компетенций), на формирование которых направлена дисциплина, сформированы на уровне не ниже «очень хорошо»
	хорошо	Все компетенции (части компетенций), на формирование которых направлена дисциплина, сформированы на уровне не ниже «хорошо».
	удовлетворительно	Все компетенции (части компетенций), на формирование которых направлена дисциплина, сформированы на уровне не ниже «удовлетворительно», при этом хотя бы одна компетенция сформирована на уровне «удовлетворительно»

не зачтено	**неудовлетворительно**	Хотя бы одна компетенция сформирована на уровне «неудовлетворительно».
	плохо	Хотя бы одна компетенция сформирована на уровне «плохо»

5.3 Типовые контрольные задания или иные материалы, необходимые для оценки результатов обучения на промежуточной аттестации с указанием критериев их оценивания:

5.3.1 Типовые задания (оценочное средство - Контрольные вопросы) для оценки сформированности компетенции ПК-3

1. Описательные статистики. Построение диаграмм размаха, гистограмм, функций плотности вероятности, планок погрешностей на Python.

2. Случайные величины и их распределения. Основные дискретные и непрерывные распределения.

3. Корреляционный анализ. Проверка статистических гипотез. FDR-коррекция.

4. Алгоритм предварительной обработки данных для задач машинного обучения.

5. Алгоритм k ближайших соседей.

6. Логистическая регрессия. Логистическая функция, функция softmax.

7. Машина опорных векторов. Ядра и спрямляющие пространства. Случай с более чем двумя классами.

8. Деревья решений. Алгоритм CART.

9. Ансамбли решающих правил. Баггинг. Случайный лес. Экстремально случайные деревья.

10. Ансамбли решающих правил. Бустинг. AdaBoost. Градиентный бустинг деревьев решений.

11. Нейронные сети. Персептрон Розенблатта. Алгоритм обратного распространения ошибки.

12. Линейная регрессия. Метод наименьших квадратов. Проверка статистической значимости модели. Коэффициент детерминации.

13. Экспериментальная оценка качества обучения и выбор параметров модели. Обучающая, проверочная и тестовая выборки. Метод перекрестного (скользящего) контроля. Метрики качества алгоритмов машинного обучения.

14. Метод сокращения размерности: метод главных компонент.

15. Методы отбора признаков: методы фильтрации, встроенные методы, методы обертки.

16. Методы борьбы с переобучением.

17. Задача кластеризации. Метод k средних. Метод DBSCAN. Алгоритмы иерархической кластеризации.

18. Секвенирование нового поколения. Экспериментальные подходы.

19. Алгоритмы выравнивания и псевдовыравнивания.

20. Анализ дифференциальной экспрессии генов.

21. Анализ функционального обогащения набора генов.

Критерии оценивания (оценочное средство - Контрольные вопросы)

Оценка	Критерии оценивания
зачтено	Студент ответил на большую часть вопросов возможно с незначительными недочетами.
не зачтено	При ответе студент допускает грубые ошибки в основном материале и решении стандартных задач.

5.3.2 Типовые задания (оценочное средство - Задачи) для оценки сформированности компетенции ПК-3

Загрузите из файла rnaseq.csv данные экспрессии генов. Выполните следующие задания (на языке программирования Python в Jupyter Notebook):

1) Выполните нормализацию признаков.

2) Разбейте данные на обучающую и тестовую выборку.

3) Обучите метод *k* ближайших соседей. Постройте графики зависимости ошибки классификации на обучающей и тестовой выборках от *k*.

4) Выполните процедуру перекрестного контроля (5-fold, 10-fold, LOO) с обучающей выборкой. Постройте графики зависимости CV-ошибки от числа используемых соседей *k*. Выберите наилучшую модель и проверьте ее качество на тестовой выборке.

5) Примените к рассматриваемым данным метод главных компонент для сокращения размерности пространства признаков. Спроецируйте данные в новое пространство.

6) Обучите метод *k* ближайших соседей на спроецированных данных, вычислите ошибки классификации на обучающей и тестовой выборках в зависимости от *k*.

Критерии оценивания (оценочное средство - Задачи)

Оценка	Критерии оценивания
зачтено	Выполнены все или большая часть этапов решения задачи с соблюдением всех требований, указанных в условии задачи, или задача решена с незначительными недочетами. Результаты работы представлены преподавателю в срок. Ответ на вопросы по заданию изложен четко и логично.

Оценка	Критерии оценивания
не зачтено	Выполнены не все этапы решения задачи или выполнены не в полном объеме (представлено не полное описание этапов выполнения заданий, получен неверный ответ, не соблюдены требования, указанные в условии задачи, результаты работы не представлены преподавателю). Ответ на вопросы по заданию неполный и поверхностный.

6. Учебно-методическое и информационное обеспечение дисциплины (модуля)

Основная литература:

1. Кремер Наум Шевелевич. Теория вероятностей и математическая статистика = Probability theory and mathematical statistics : учеб. для студентов вузов, обучающихся по экон. специальностям. - 2-е изд., доп. и перераб. - М. : Юнити-Дана, 2006. - 573 с. - Библиогр. список: с. 533 - 534. - ISBN 5-238-00573-3 : 210.00., 58 экз.

Дополнительная литература:

1. Неделько Виктор Михайлович. Основы теории вероятности : Учебное пособие. - Новосибирск : Новосибирский государственный технический университет (НГТУ), 2011. - 116 с. - Профессиональное образование. - ISBN 978-5-7782-1701-0., https://e-lib.unn.ru/MegaPro/UserEntry?Action=FindDocs&ids=609079&idb=0.

Программное обеспечение и Интернет-ресурсы (в соответствии с содержанием дисциплины):

1. Anaconda3 (дистрибутив Python): https://www.anaconda.com/download.
2. Программный пакет R: http://cran.r-project.org/.

7. Материально-техническое обеспечение дисциплины (модуля)

Учебные аудитории для проведения учебных занятий, предусмотренных образовательной программой, оснащены мультимедийным оборудованием (проектор, экран), техническими средствами обучения, компьютерами, специализированным оборудованием: Лицензионное и свободно распространяемое ПО: операционные системы семейства Microsoft Windows, – лицензия по подписке Microsoft Imagine; Anaconda3 (дистрибутив Python), – https://www.anaconda.com/download (в свободном доступе): программный пакет R, – http://cran.r-project.org/ (в свободном доступе).

Помещения для самостоятельной работы обучающихся оснащены компьютерной техникой с возможностью подключения к сети "Интернет" и обеспечены доступом в электронную информационно-образовательную среду.

Программа составлена в соответствии с требованиями ОС ННГУ по направлению подготовки/специальности 02.03.02 - Фундаментальная информатика и информационные технологии.

Автор(ы): Вершинина Ольга Сергеевна, кандидат физико-математических наук.

Рецензент(ы): д.т.н., профессор НГТУ им. Р.Е. Алексеева Ломакина Л.С..

Заведующий кафедрой: Иванченко Михаил Васильевич, доктор физико-математических наук.

Программа одобрена на заседании методической комиссии от 02.12.2024, протокол № 5.