АВТОНОМНАЯ НЕКОММЕРЧЕСКАЯ ОБРАЗОВАТЕЛЬНАЯ ОРГАНИЗАЦИЯ ВЫСШЕГО ОБРАЗОВАНИЯ «НАУЧНО-ТЕХНОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ «СИРИУС» (АНОО ВО «УНИВЕРСИТЕТ «СИРИУС»)

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ (МОДУЛЯ)

Машинное обучение: нейронные сети в биоинформатике

Уровень образования:	высшее образование – программа магистратуры
Направление подготовки:	06.04.01 Биология
	09.04.03 Прикладная информатика
Направленность (профиль):	Биоинформатика
Форма обучения:	очная
Язык преподавания:	русский

1. Общая характеристика дисциплины

1.1. Цель дисциплины: формирование у студентов теоретических знаний и практических навыков численных методов и методов оптимизации для решения различных прикладных задач математической биологии и биоинформатики.

1.2. Задачи дисциплины: развитие у студентов математического мышления, умения ставить, исследовать и решать сложные задачи прикладной математики и информатики.

1.3. Общая трудоемкость дисциплины: 3 з.е.

1.4. Планируемые результаты обучения по дисциплине:

Формируемые компетенции (код компетенции, формулировка)	Планируемые результаты обучения по дисциплине (индикаторы достижения компетенций)
ПК-1. Способен решать актуальные задачи фундаментальной и прикладной математики.	ИПК-1.1. Знает фундаментальные основы математики, биологии и других естественных наук
	ИПК-1.2. Применяет фундаментальные знания математики, биологии и других естественных наук для постановки и решения исследовательских и практических задач
	ИПК-1.3. Анализирует современные проблемы в области биоинформатики, биоинженерии, биотехнологии и фарминдустрии, формулирует гипотезы и вырабатывает подходы для решения исследовательских и практических задач
ПК-2. Способен разрабатывать оригинальные алгоритмы и программные средства, в том числе с использованием современных информационных технологий, для решения профессиональных задач в области биоинформатики, биоинженерии, биотехнологии и фарминдустрии	ИПК-2.1. Знает современные алгоритмы, средства разработки и программные средства, а также принципы написания программ на различных языках программирования
	ИПК-2.2. Осуществляет анализ и выбор методов решения профессиональных задач на основе теоретических знаний в области информационных технологий
	ИПК-2.3. Разрабатывает оригинальные алгоритмы и программные средства для решения профессиональных задач в области биоинформатики, биоинженерии, биотехнологии и фарминдустрии
ПК-5. Способен применять методы математического и компьютерного исследования при разработке интеграционных решений на основе знаний фундаментальных математических и компьютерных наук и навыками проблемно-	ИПК-5.1. Знает математические алгоритмы и принципы определения необходимых системных и программных средств для решения профессиональных задач
	ИПК-5.2. Определяет необходимые системные и программные средства для разработки и отладки прикладного программного обеспечения в современных специализированных программных комплексах

задачной формы представления научных знаний.	ИПК-5.3. Реализует новые алгоритмы в современных специализированных программных комплексах

2. Структура и содержание дисциплины

2.1. Объем дисциплины и виды учебной деятельности:

Виды учебной деятельности	2 семестр	Всего
Контактная работа обучающихся с преподавателем, всего ч.	76	76
Лекционные занятия, ч.	36	36
Практические (семинарские) занятия, ч.	38	38
Лабораторные занятия, ч.	x	x
Промежуточная аттестация – экзамен, ч	x	x
Промежуточная аттестация – зачет с оценкой, ч	x	x
Промежуточная аттестация – зачет, ч	2	2
Самостоятельная работа обучающихся, всего ч.	32	32
Общая трудоемкость, ч.	108	108
Общая трудоемкость, з.е.	3	3

2.2. Структура дисциплины по разделам (темам) и видам учебной деятельности:

Наименования разделов (тем) дисциплины	Лекционные занятия, ч	Практические (семинарские) занятия, ч	Лабораторные занятия, ч	Промежуточная аттестация, ч	Самостоятельная работа, ч	Всего, ч	Форма текущего контроля / промежуточной аттестации
Раздел 1. Введение. Краткая история дисциплины. Общая постановка задачи оптимизации. Примеры	2	2				4	письменное домашнее задание
Раздел 2. Предмет и задачи машинного обучения и анализа данных. Основные принципы, задачи и подходы, использование в различных областях науки и индустрии. Основные этапы эволюции алгоритмов машинного обучения.	2	4			2	8	письменное домашнее задание
Раздел 3. Общий вид метрического классификатора.	4	4			4	16	письменное домашнее задание

Алгоритм K ближайших соседей. Алгоритмы отбора эталонов.							
Раздел 4. Алгоритмы кластеризации с фиксированным количеством кластеров. Алгоритмы кластеризации по плотности. Иерархическая кластеризация.	4	4			4	16	письменное домашнее задание
Раздел 5. Правила и анализ качества (точность, полнота). Анализ с помощью ROC кривой. Алгоритм построения деревьев решений. Критерий информационного выигрыша и критерий Джини. Леса решающих деревьев.	4	4			4	16	письменное домашнее задание
Раздел 6. Перцептрон и разделяющая гиперплоскость. Переход в пространство повышенной размерности. Метод опорных векторов	4	4			4	16	письменное домашнее задание
Раздел 7. Логистическая регрессия. Градиентный спуск. Нейронные сети и алгоритм обратного распространения градиента. Глубокое обучение, свертки и пулинг.	4	4			4	16	письменное домашнее задание
Раздел 8. Линейная регрессия. Полиномиальная регрессия. Смещение и дисперсия. Гребневая регрессия.	4	4			4	16	письменное домашнее задание
Раздел 9. Голосование. Бутстраппинг. Бустинг, адаптивный бустинг, градиентный бустинг.	4	4			4	16	письменное домашнее задание
Раздел 10. Монте-Карло поиск. Алгоритм симулированного отжига. Генетический алгоритм.	4	4			2	10	письменное домашнее задание

Промежуточная аттестация				2		2	Зачет
Итого	36	38	х	2	32	108	

2.3. Содержание разделов (тем) дисциплины:

Наименования разделов (тем) дисциплины
Раздел 1. Введение. Краткая история дисциплины. Общая постановка задачи оптимизации. Примеры
Раздел 2. Предмет и задачи машинного обучения и анализа данных. Основные принципы, задачи и подходы, использование в различных областях науки и индустрии. Основные этапы эволюции алгоритмов машинного обучения.
Раздел 3. Общий вид метрического классификатора. Алгоритм K ближайших соседей. Алгоритмы отбора эталонов.
Раздел 4. Алгоритмы кластеризации с фиксированным количеством кластеров. Алгоритмы кластеризации по плотности. Иерархическая кластеризация.
Раздел 5. Правила и анализ качества (точность, полнота). Анализ с помощью ROC кривой. Алгоритм построения деревьев решений. Критерий информационного выигрыша и критерий Джини. Леса решающих деревьев.
Раздел 6. Перцептрон и разделяющая гиперплоскость. Переход в пространство повышенной размерности. Метод опорных векторов
Раздел 7. Логистическая регрессия. Градиентный спуск. Нейронные сети и алгоритм обратного распространения градиента. Глубокое обучение, свертки и пулинг.
Раздел 8. Линейная регрессия. Полиномиальная регрессия. Смещение и дисперсия. Гребневая регрессия.
Раздел 9. Голосование. Бутстраппинг. Бустинг, адаптивный бустинг, градиентный бустинг.
Раздел 10. Монте-Карло поиск. Алгоритм симулированного отжига. Генетический алгоритм.

2.4. Самостоятельная работа:

Самостоятельная работа по дисциплине предусматривает: самостоятельное изучение теоретического материала, подготовку к ответам на семинарских заданиях, подготовку к текущему контролю и промежуточной аттестации, выполнение тестовых заданий по пройденным темам курса.

3. Текущий контроль и промежуточная аттестация по дисциплине. Оценочные материалы

3.1. Текущий контроль успеваемости по дисциплине «Методы машинного обучения» проводится в течение семестра в следующих формах:

Наименования разделов (тем) дисциплины	Форма текущего контроля	Оценочные материалы
Раздел 1. Введение. Краткая история дисциплины. Общая постановка задачи оптимизации. Примеры	Устный опрос, тестирование	перечень домашних заданий
Раздел 2. Предмет и задачи машинного обучения и анализа данных. Основные принципы, задачи и подходы, использование в различных областях науки и индустрии.	Устный опрос, тестирование	перечень домашних заданий

Основные этапы эволюции алгоритмов машинного обучения.		
Раздел 3. Общий вид метрического классификатора. Алгоритм K ближайших соседей. Алгоритмы отбора эталонов.	Устный опрос, тестирование	перечень домашних заданий
Раздел 4. Алгоритмы кластеризации с фиксированным количеством кластеров. Алгоритмы кластеризации по плотности. Иерархическая кластеризация.	Устный опрос, тестирование	перечень домашних заданий
Раздел 5. Правила и анализ качества (точность, полнота). Анализ с помощью ROC кривой. Алгоритм построения деревьев решений. Критерий информационного выигрыша и критерий Джини. Леса решающих деревьев.	Устный опрос, тестирование	перечень домашних заданий
Раздел 6. Перцептрон и разделяющая гиперплоскость. Переход в пространство повышенной размерности. Метод опорных векторов	Устный опрос, тестирование	перечень домашних заданий
Раздел 7. Логистическая регрессия. Градиентный спуск. Нейронные сети и алгоритм обратного распространения градиента. Глубокое обучение, свертки и пулинг.	Устный опрос, тестирование	перечень домашних заданий
Раздел 8. Линейная регрессия. Полиномиальная регрессия. Смещение и дисперсия. Гребневая регрессия.	Устный опрос, тестирование	перечень домашних заданий
Раздел 9. Голосование. Бутстраппинг. Бустинг, адаптивный бустинг, градиентный бустинг.	Устный опрос, тестирование	перечень домашних заданий
Раздел 10. Монте-Карло поиск. Алгоритм симулированного отжига. Генетический алгоритм.	Устный опрос, тестирование	перечень домашних заданий

3.2. Оценочные материалы для текущего контроля:

Критерии оценки для устного опроса

Критерий	Зачтено	Не зачтено
Процент правильных ответов	Успешно/в целом успешно применяет инновационные инструменты и методы при определении путей решения профессиональных задач.	Не применяет/не в полной мере применяет инновационные инструменты и методы при определении путей решения профессиональных задач.

Примеры домашних заданий

Домашнее задание №1 выдается студентам в одном варианте и состоит из 3 задач. Каждой задаче присвоен свой балл. Срок выполнения домашнего задания - 2 недели. Форма представления обучающимися домашнего задания -

представленные в письменном виде решения задач.

Задача 1 [3 балла]. Реализуйте алгоритм kNN классификации по k ближайшим соседям, используя простое евклидовое расстояние.

Задача 2 [3 балла]. Реализуйте алгоритм k-means для кластеризации на 2-4 кластера.

Задача 3 [4 балла]. Реализуйте алгоритм DBSCAN, найдите параметры для кластеризации на 4 кластера.

Домашнее задание №2 выдается студентам в одном варианте и состоит из 4 задач. Каждой задаче присвоен свой балл. Срок выполнения домашнего задания - 2 недели. Форма представления обучающимися домашнего задания - представленные в письменном виде решения задач.

Задача 1 [2,5 балла]. Реализуйте алгоритмы построения дерева с критерием информационного выигрыша и критерием Джини и определению класса по мажоритарному классу в листе. Найдите оптимальную глубину дерева в обоих случаях (в отрезке 2-10).

Задача 2 [2,5 балла]. Примените метод SVM (например, из библиотеки sklearn) для датасета blobs2. Визуализируйте результат (разбиение плоскости и опорные вектора) при разных вариантах ядер (линейное; полиномиальное степеней 2,3,5; RBF).

Задача 3 [2,5 балла]. Реализуйте алгоритм логистической регрессии со стохастическим градиентным спуском, обучите его на датасете spambase_old (train) и проверьте на датасете spambase_new (val). Получите ROC кривые для вариантов без нормировки и с нормировкой признаков.

Задача 4 [2,5 балла]. Модифицируйте модель из задачи 3, заменив последний нейрон на 10 нейронов, и реализовав мультиклассовую классификацию с softmax в качестве решающей функции и кросс-энтропией в качестве функции потерь и обучите на подготовленном датасете mnist.

Домашнее задание №3 выдается студентам в одном варианте и состоит из 4 задач. Каждой задаче присвоен свой балл. Срок выполнения домашнего задания - 2 недели. Форма представления обучающимися домашнего задания - представленные в письменном виде решения задач.

Задача 1 [2,5 балла]. Реализуйте алгоритм линейной регрессии, и полиномиальной регрессии (для датасета noisysine – степеней от 2 до 5, для датасета hydrodynamics – степени 2) без регуляризации.

Задача 2 [2,5 балла]. Реализуйте алгоритм гребневой регрессии и найдите оптимальный параметр регуляризации для случаев из задачи 1.

Задача 3 [2,5 балла]. Найдите максимум функции с помощью алгоритма кросс-энтропийного поиска, изображая распределение на каждом шаге. Задача 4

[2,5 балла]. Найдите лучший путь в задаче коммивояжёра с помощью алгоритма отжига

Критерии оценки для устного опроса

Критерий	Зачтено	Не зачтено
Процент правильных ответов	Успешно/в целом успешно применяет инновационные инструменты и методы при определении путей решения профессиональных задач.	Не применяет/не в полной мере применяет инновационные инструменты и методы при определении путей решения профессиональных задач.

Примерный перечень тестовых заданий:

1. Что такое машинное обучение?

a) Программирование компьютерных игр

b) Разработка программного обеспечения для офисной работы

c) Направление искусственного интеллекта, основанное на обучении моделей на основе данных

d) Изучение поведения человека

Правильный ответ: c)

2. Какой тип нейронной сети используется для обработки изображений?

a) Полносвязная нейронная сеть

b) Свёрточная нейронная сеть (CNN)

c) Рекуррентная нейронная сеть (RNN)

d) Генеративная состязательная сеть (GAN)

Правильный ответ: b)

3. Что такое биоинформатика?

a) Анализ только медицинских документов

b) Наука о применении информационных технологий к анализу биологических данных

c) Изучение биологии с помощью химии

d) Только работа с ДНК-кодом

Правильный ответ: b)

4. Какие данные могут быть использованы в биоинформатике при работе с нейронными сетями?

a) Только финансовые отчеты

b) Последовательности ДНК, белковые структуры, данные экспрессии генов

c) Только текстовые документы

d) Социологические опросы

Правильный ответ: b)

5. Какова роль нейронных сетей в задачах распознавания образцов в биоинформатике?

a) Они не используются

b) Позволяют автоматически выявлять закономерности в больших биологических данных

c) Используются только для обучения студентов

d) Не имеют практического применения

Правильный ответ: b)

Критерии оценки для тестовых заданий

Критерий	Зачтено	Не зачтено
Процент правильных ответов	Успешно/в целом успешно применяет инновационные инструменты и методы при определении путей решения профессиональных задач.	Не применяет/не в полной мере применяет инновационные инструменты и методы при определении путей решения профессиональных задач.

3.3. Формой промежуточной аттестации по дисциплине является зачет.

Результаты промежуточной аттестации оцениваются как «зачтено» и «не зачтено».

Оценка «зачтено» означает успешное прохождение промежуточной аттестации по дисциплине.

3.4. Оценочные материалы для промежуточной аттестации:

Перечень вопросов для подготовки к зачету:

1. Препроцессинг. Масштабирование. Нормировка. Полиномиальные признаки. One-hot encoding.
2. Кластеризация. kMeans, MeanShift, DBSCAN, Affinity Propagation.
3. Смещение и дисперсия (bias and variance). Понятие средней гипотезы.
4. Ансамблевые методы. Soft and Hard Voting. Bagging. Случайные леса. AdaBoost.

5. Типы обучения: с учителем, без учителя, с подкреплением, с частичным участием учителя, активное обучение.

6. Бустинг деревьев решений.

7. Ошибка внутри и вне выборки. Ошибка обобщения. Неравенство Хёфдинга. Валидация и кросс-валидация.

8. Линейная регрессия. Полиномиальная регрессия. Гребневая регрессия.

9. Размерность Вапника-Червоненкиса. Размерность Вапника-Червоненкиса для перцептрона.

10. Логистическая регрессия. Градиентный спуск.

11. Пороговые условия. Эффективность по Парето. Presicion-Recall и ROC кривые. AUC.

12. Ансамблевые методы регрессии. RANSAC. Theil-Sen. Huber.

13. Перцептрон. Перцептрон с карманом.

14. Метод опорных векторов. Постановка задачи. Формулировка и решение двойственной задачи. Типы опорных векторов. Ядра.

15. Гипотезы и дихотомии. Функция роста. Точка поломки. Доказательство полиномиальности функции роста в присутствии точки поломки.

16. Деревья решений. Информационный выигрыш, критерий Джини. Регуляризация деревьев. Небрежные решающие деревья.

17. Байесовский классификатор. Типы оценки распределений признаков (Gaussian, Bernoulli, Multinomial). EM алгоритм.

18. Нейронные сети. Перцептрон Розенблатта. Функции активации. Обратное распространение градиента. Softmax.

19. Стохастическая оптимизация. Hill Climb. Отжиг. Генетический алгоритм.

20. Метрические классификаторы. kNN. WkNN. Отбор эталонов. DROP5. Kdtree.

4. Учебно-методическое и информационное обеспечение дисциплины

4.1. Перечень основной литературы:

1. П. Флах. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных Machine Learning and Knowledge Discovery in Databases / Peter A. Flach; Tijl Bie; Nello Cristianini. Springer Berlin Heidelberg. 2012

4.2. Перечень дополнительной литературы:

1. An introduction to statistical learning: with applications in R / G. James, D. Witten, T. Hastie, R. Tibshirani. – New York: Springer, 2013. – 426 с.

2. Data Analysis, Machine Learning and Knowledge Discovery / Spiliopoulou, Myra; Janning, Ruth; Schmidt-Thieme, Lars; Gesellschaft für Klassifikation. Springer International Publishing. 2014.

3. Machine Learning: A Probabilistic Perspective (Adaptive Computation and Machine Learning Series) / Murphy, Kevin P. MIT Press. 2014

4. Encyclopedia of Machine Learning and Data Mining / Sammut. Springer. 2016

4.3. Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине:

В рамках самостоятельной работы обучающиеся используют интернет-ресурсы:

1. https://collection.asdlib.org/

4.4. Перечень современных профессиональных баз данных и ресурсов информационно-телекоммуникационной сети «Интернет»:

1. https://stepik.org/course/4852/promo

5. Материально-техническое и программное обеспечение дисциплины

5.1. Материально-техническое обеспечение:

Вид аудитории	Технические средства и оборудование
Учебная аудитория для проведения лекционных занятий	Альфа 5.1 - учебная аудитория для проведения учебных занятий, предусмотренных программой магистратуры. Доска магнитно-маркерная поворотная BoardSYS Twist 100x160 ПО-15Ф 1 шт. Флипчарт 70*100 на роликах 1 шт. Стол-кафедра 1 шт. Стол аудиторный 1 шт. Столы-трансформеры Summa GA ученические 25 шт. Стулья на колесах ученические 25 шт. Ноутбук HP 1 шт. Интерактивная панель NexTouch Nextpanel 86” 1 шт. Радиосистема Arthur Forty U-9700C PSC (UHF) в комплекте. Акустическая система Behringer B215D 2 шт. Веб-камера 4K с технологией искусственного интеллекта JazzTel JT-Vintage-4K 1 шт. Комплект электронных презентаций.
Учебная аудитория для проведения практических занятий – Компьютерный класс	Бета 4.1 – учебная аудитория для проведения практических занятий (компьютерный класс). Доска магнитно-маркерная поворотная BoardSYS Twist 100x160 ПО-15Ф 1 шт. Флипчарт 70*100 на роликах 1 шт. Стол преподавателя аудиторный 1 шт. Столы и стулья ученические 42 шт. Компьютеры Lenovo ThinkCentre M920s SFF в комплекте с мониторами IIYAMA 27” и периферией – 42 шт. Интерактивная панель NexTouch Nextpanel 86” 1 шт. Радиосистема Arthur Forty U-9700C PSC (UHF) в комплекте. Акустическая система Behringer B215D 2 шт. Веб-камера 4K с технологией искусственного интеллекта JazzTel JT-Vintage-4K 1 шт. Комплект электронных презентаций.

5.2. Перечень лицензионного и свободно распространяемого программного обеспечения, в том числе российского производства:

– пакет библиотек для Python (Anaconda)
– инструмент для сборки Haskell (Stack), компилятор C++ (clang).