МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение высшего образования
«Дальневосточный федеральный университет»
(ДВФУ)

ШКОЛА ЦИФРОВОЙ ЭКОНОМИКИ

СОГЛАСОВАНО
Руководитель ОП
_____ Е.В. Пустовалов
«_24_» __июня___2018 г.

«УТВЕРЖДАЮ»
Проректор по развитию
_____ Д.И. Земцов
«_27_» __июня___2018 г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

«МАТЕМАТИЧЕСКИЕ МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ»

направления 09.04.01 Информатика и вычислительная техника
Магистерская программа «Технологии виртуальной и дополненной реальности»
Форма подготовки очная

курс 1 семестр 1,2
лекции 72 час.
практические занятия 0 час.
лабораторные работы 0 час.
всего часов аудиторной нагрузки 36 час.
самостоятельная работа 144 час.
в том числе на подготовку к экзамену – 72 час.
контрольные работы программой не предусмотрены
курсовая работа/проект – не предусмотрено
зачет – не предусмотрено учебным планом
экзамен 1, 2 семестр

Рабочая программа составлена в соответствии с требованиями федерального государственного образовательного стандарта высшего образования по направлению подготовки 09.04.01 – Информатика и вычислительная техника, утвержденного приказом Министерства образования и науки Российской Федерации от 30.10.2014 № 1420

Рабочая программа рассмотрена и утверждена на заседании Дирекции Школы цифровой экономики 24 июня 2018 г., протокол №2

Составитель(и): д.ф.-м.н. Нурминский Е.А., ст. пр. Кленин А.С.

Оборотная сторона титульного листа РПД

I. Рабочая программа пересмотрена на заседании Дирекции Школы цифровой экономики:
Протокол от «_____» _____ 20___ г. № _____

Заместитель директора ШЦЭ
по учебной и воспитательной работе _____ _____
(подпись) (И.О. Фамилия)

II. Рабочая программа пересмотрена на заседании Дирекции Школы цифровой экономики:
Протокол от «_____» _____ 20___ г. № _____

Заместитель директора ШЦЭ
по учебной и воспитательной работе _____ _____
(подпись) (И.О. Фамилия)

АННОТАЦИЯ

Математические методы машинного обучения

Рабочая программа учебной дисциплины «Математические методы машинного обучения» предназначена для студентов, обучающихся по направлению подготовки 09.04.01 Информатика и вычислительная техника (уровень магистратуры), профиль «Технологии виртуальной и дополненной реальности».

Рабочая программа разработана на основе макета рабочей программы учебной дисциплины для образовательных программ высшего образования – программ бакалавриата, специалитета, магистратуры ДВФУ, утверждённого приказом ректора ДВФУ от 08.05.2015 № 12-13-824.

Дисциплина «Математические методы машинного обучения» входит в вариативную часть блока «Дисциплины (модули)» (Б1.Б.02.01) учебного плана подготовки магистров.

Общая трудоемкость освоения дисциплины составляет 6 зачетных единиц, 216 часа. Дисциплина реализуется на 1 курсе в 1 и 2 семестре.

Семестр	Аудиторные занятия	Самостоятельная работа	Контроль	Форма контроля	Всего по дисциплине	
	Лекции				Часы	Зачетные единицы
1 семестр	36	36	36	Экзамен	108	3
2 семестр	36	36	36	Экзамен	108	3
Всего	**72**	**72**	**72**		**216**	**6**

Цель – изучение основных разделов теории машинного обучения (Machine Learning) и овладение навыками практического решения задач интеллектуального анализа данных – майнинга данных (Data Mining).

Задачи:

- Изучить основные инструменты математического анализа, линейной алгебры, методов оптимизации и теории вероятностей;
- Получить базовые навыки программирования на языках C++ и Python применительно к работе с большими объемами данных;

- Изучить основные модели машинного обучения и методики оценки их качества;
- Изучить основные способы организации искусственных нейронных сетей;
- Овладеть методологией управления data-science проектами;
- Научиться строить модели машинного обучения для решения профессиональных задач.

В результате освоения дисциплины обучающийся должен:

Знать:

современное состояние исследований в области машинного обучения;
принципы построения систем машинного обучения;
модели представления и описания технологий машинного обучения.

Уметь:

проводить анализ предметной области;
определять назначение, выбирать методы и средства для построения систем машинного обучения;
строить системы машинного обучения.

Иметь навыки и (или) опыт деятельности (владеть):

использования аппарата простейшего анализ данных;
применения методов классификации информации;
реализации алгоритмов машинного обучения.

Связь курса с другими дисциплинами.

Для успешного изучения дисциплины «Математические методы машинного обучения» необходимы знания базовой программы курса «Высшая математика» и основ программирования (желательно Python).

В результате данной дисциплины у обучающихся формируются следующие общепрофессиональные и профессиональные компетенции (элементы компетенций).

Код и формулировка компетенции	**Этапы формирования компетенции**	
ОК-3 – способность к самостоятельному обучению новым методам исследования, к изменению научного и научно-производственного профиля своей профессиональной деятельности	Знает	- основные научно-технические тенденции, истории их развития в системе обучения; - современные подходы к проектированию и использованию информационных технологий и ресурсов в образовании; - основные методики организации самостоятельного обучения - основные средства информационных технологий в образовательной и профессиональной

		деятельности
	Умеет	- проводить анализ образовательных продуктов, научных трудов и публикаций для решения задач профессиональной деятельности - осуществлять методическую проработку новых знаний и методов исследования, а также адаптировать их к собственным профессиональным задачам - оценивать достижения использования информационных технологий обучения для последующей управляемости и воспроизводимости полученных результатов; - применять мультимедийные технологии в образовании и при проведении исследований.
	Владеет	- подходами в решении задач, связанных с недостаточностью профессиональных знаний и методов исследования; - способностью совершенствовать и развивать свой интеллектуальный и общекультурный уровень; - опытом исследовательской деятельности в сфере анализа информационных технологий в контексте их эффективности; - опытом образовательной деятельности в среде информационных технологий; - рефлексивной деятельности в том числе самооценки, взаимооценки, рецензирования.
ОК-6 – способность проявлять инициативу, в том числе в ситуациях риска, брать на себя всю полноту ответственности	Знает	- способы организации исследовательских и проектных работ, управления коллективом
	Умеет	- самостоятельно планировать шаги выполнения поставленных задач, - видеть личную ответственность за выполнение задания в соответствии самостоятельно определенным планом - видеть личную ответственность за принятые проектные и реализационные решения, - самостоятельно анализировать ход выполнения задания и осуществлять изменения в принятом порядке выполнения задания, в проектных и реализационных решениях с целью обеспечить выполнение задания полностью и в срок.
	Владеет	- навыками в организации исследовательских и проектных работ, в управлении коллективом
ОК-7 – способностью самостоятельно приобретать с помощью информационных технологий и использовать в практической деятельности, в том числе в новых областях знаний,	Знает	- информационные возможности современного общества (интернет): электронные библиотеки, сервисы, журналы и др. ресурсы для освоения раздела дисциплины, получения новой информации или дополнительных умений.
	Умеет	- Умеет найти информацию в глобальных компьютерных сетях на заданную тему, используя надёжные и достоверные источники.

непосредственно не связанных со сферой деятельности	Владеет	- основными методами поиска и обработки информации; - навыками принимать самостоятельное решение о ценности, необходимости, важности и надёжности найденной информации пропорционально затрачиваемому времени на её поиск в условиях ограниченности временных рамок.
ОПК-5 – владение методами и средствами получения, хранения, переработки и трансляции информации посредством современных компьютерных технологий, в том числе в хранения, переработки и трансляции информации посредством современных компьютерных технологий, в том числе в глобальных компьютерных сетях	Знает	- основные методы и средства получения, хранения, переработки и трансляции информации посредством современных компьютерных технологий, в том числе в глобальных компьютерных сетях
	Умеет	- использовать основные методы и средства получения, хранения, переработки и трансляции информации посредством современных компьютерных технологий, в том числе в глобальных компьютерных сетях; - организовывать поиск информации по различным критериям с использованием различных поисковых технологий; - извлекать необходимую информацию из информационных систем и преобразовывать ее к необходимому виду
	Владеет	- методами получения, хранения, переработки и трансляции информации посредством современных компьютерных технологий, в том числе в глобальных компьютерных сетях
ОПК-6 – способность анализировать профессиональную информацию, выделять в ней главное, структурировать, оформлять и представлять в виде аналитических обзоров с обоснованными выводами и рекомендациями	Знает	- современное состояние исследований в области машинного обучения; - принципы построения систем машинного обучения; - модели представления и описания технологий машинного обучения
	Умеет	- проводить анализ предметной области
	Владеет	- навыками структурирования, оформления и представления информации в виде аналитических обзоров с обоснованными выводами и рекомендациями по профилю деятельности; - способностью формулировать обоснованные выводы и рекомендации по предлагаемым техническим решениям
ПК-8 – способность проектировать распределенные информационные системы, их компоненты и протоколы их взаимодействия	Знает	- основные стандарты в области организации доступа к распределенным информационным системам; - основные технологии реализации распределенных систем; - основные технологии поиска информации в распределенных информационных системах; - основные технологии представления и передачи структурированной информации в

		распределенных информационных системах
	Умеет	- проектировать распределенные информационные системы; разрабатывать серверное и клиентское программное обеспечения распределенных информационных систем; пользоваться архивами свободно распространяемого программного обеспечения конструировать программные комплексы для распределенных информационных систем; - организовывать преобразование данных на основе стандартных технологий; создавать пользовательские интерфейсы для доступа к распределенным информационным системам
	Владеет	- навыками программной реализации распределенных информационных систем; конструирования программных комплексов для распределенных информационных систем; создания пользовательских интерфейсов для доступа к распределенным информационным системам
ПК-15 – способностью к созданию программного обеспечения для анализа, распознавания и обработки информации, систем цифровой обработки сигналов	Знает	- математический аппарат, применяемый для анализа, распознавания и обработки информации, систем цифровой обработки сигналов; - базовые алгоритмы цифровой обработки сигналов, распознавания и обработки информации; - возможности современных средств проектирования и моделирования устройств цифровой обработки сигналов
	Умеет	- анализировать поставленную задачу и выбирать методы и средства создания программного обеспечения для анализа, распознавания и обработки информации, систем цифровой обработки сигналов, оптимально подходящие для решения задачи
	Владеет	- навыками создания программного обеспечения для анализа, распознавания и обработки информации, – навыками создания систем цифровой обработки сигналов

Для формирования вышеуказанных компетенций в рамках дисциплины «Машинное обучение и программирование» применяются следующие методы активного/интерактивного обучения: круглый стол (дискуссия, дебаты), семинар, проектная сессия, проектный семинар и др.

I. СТРУКТУРА И СОДЕРЖАНИЕ ТЕОРЕТИЧЕСКОЙ ЧАСТИ КУРСА
I СЕМЕСТР (36 часов)

Тема 1.1 Введение. Методы оптимизации. Градиентный спуск (2 часа)

Задача регрессии, классификации.

Функция потерь. Оптимизация. Перебор по сетке.

Производная, частные производные, градиент.

Градиентный спуск, проблема выбора шага.

Стохастический градиентный спуск. Использование момента.

Adagrad, Adadelta, Adam.

RMSProp*.

Тема 1.2 Линейная регрессия (2 часа)

Постановка задачи линейной регрессии.

Метод наименьших квадратов.

Ковариация, корреляция.

Критерий R^2.

Анализ остатков.

Тема 1.3 Глобальная оптимизация. Генетический алгоритм (4 часа)

Многопараметрическая оптимизация.

Доминация и оптимальность по Парето.

Функция качества (fitness). Аппроксимация качества.

Общая идея генетического алгоритма.

Представление генома.

Методы селекции: пропорционально качеству, универсальная выборка (stochastic universal sampling), с наследием (reward-based), турнир. Стратегия элитизма.

Методы кроссовера. Двух и многоточечный, равномерный (по подмножествам), для перестановок.

Мутация. Влияние на скорость обучения.

Управление популяцией. Сегрегация, старение, распараллеливание.

Генетическое программирование.

Тема 1.4 Метод ближайших соседей (k-NN) (4 часа)

Понятие и свойства метрики. Ослабление требования к неравенству треугольника.

Базовый алгоритм классификации методом 1-NN и k-NN. Преимущества и недостатки.

Метрики L1, L2, Хемминга, Левенштейна, косинусное расстояние.

Потеря точности нормы в высоких размерностях.

Нормализация координат. Предварительная трансформация пространства признаков.

Метрика Махаланобиса.

Кросс-валидация методом "без одного" (leave one out).

Определение границ, показатель пограничности.

Сжатие по данным. Понятия выброса, прототипа, усвоенной точки. Алгоритм Харта (Hart).

Регрессия методом k-NN.

Взвешенные соседи.

Связь с градиентным спуском. Стохастическая формулировка, softmax.

Метод соседних компонент (neighbour component analysis)*.

Связь с выпуклой оптимизацией. Метод большого запаса (Large margin NN)

Оптимизация классификатора, k-d деревья

Хеши чувствительные к локальности, хеши сохраняющие локальность*.

Тема 1.5 Наивный байесов классификатор (2 часа)

Условная вероятность. Байесово решающее правило. Обновление вероятностей.

Наивный классификатор, предположение о независимости признаков.

Оценка плотности распределения для числовых признаков.

Алгоритмические оптимизации.

Алгоритм EM.

Тема 1.6 Логистическая регрессия (2часа)

Сигмоид

Метод наибольшего правдоподобия

Логистическая регрессия для меток −1, 1

Тема 1.7 Деревья решений (4 часа)

Понятие дерева решений.

Борьба с оверфиттингом: bagging, выборки признаков.

Ансамбли, случайный лес (Random Forest).

Понятие энтропии, определение информации по Шеннону.

Метрики: примеси Джини (Gini impurity), добавленная информация (information gain).

Деревья регрессии. Метрика вариации.

Непрерывные признаки. Использование главных компонент вместо признаков.

Сокращение дерева (pruning).

Тема 1.8 Кластеризация (4 часа)

Задача обучения без учителя, применения при эксплораторном анализе

Неметрическая кластеризация: функция схожести, компоненты связности и остовные деревья, иерархическая кластеризация снизу вверх

Метрики, понятие центроида и представителя класса

Центроидные алгоритмы: k-means, k-medoid

Алгоритмы, основанные на плотности: DBSCAN, OPTICS

Алгоритмы, основанные на распределении: сумма гауссиан

Нечёткая кластеризация, алгоритм c-means

Метрики качества: leave-one-out, силуэт, индекс Дэвиса-Болдина (Davies-Bouldin), индекс Данна (Dunn)

Тема 1.9 Работа с текстом (4 часа)

Задачи обработки текста: извлечение, поиск, классификация (тематическая, эмоциональная), перевод

Разбиение на слова, пунктуация, лексический и морфологический анализ

Определение частей речи, имён, основ слов

Частотный анализ, представление bag-of-words, TF-IDF и его варианты

N-грамы, byte-pair encoding.

Векторные представления, семантическая интерпретация алгебраических операций

Унитарный код (One-hot encoding).

Алгоритмы Word2Vec и FastText.

Алгоритм GloVe*.

Тема 1.10 Снижение размерности (4 часа)

Постановка задачи, причины и цели снижения размерности.

Выбор и извлечение признаков.

Подходы к выбору признаков: filtering, wrapping, embedding.

Расстояние между распределениями. Расстояние Кульбака-Лейблера. Взаимная информация.

Алгоритмы выбора признаков: на основе корреляции (CFS), взаимной информации, Relief.

Метод главных компонент (PCA).

Нелинейные обобщения метода главных компонент. Kernel PCA.*

Неотрицательное матричное разложение (NMF).*

Стохастическое вложение соседей с t-распределением (t-SNE).

Тема 1.11 Метод опорных векторов (SVM) (4 часа)

Постановка задачи линейного SVM для линейно разделимой выборки

Линейный SVM в случае линейно неразделимой выборки.

Задача оптимизации с ограничениями. Двойственная задача Лагранжа. Условия Каруша-Куна-Такера

Функция Лагранжа для линейного SVM. Опорный вектор. Типы опорных векторов.

Kernel trick. Полиномиальное ядро. Радиально-базисное ядро (RBF).

SVM для задачи регрессии.

II. СЕМЕСТР
(36 часов)

Тема 2.1 Работа с изображениями (2 часа)

Сверточные фильтры, нпрерывное и дискретное определение свёртки.

Сглаживающие фильтры. Фильтр Гаусса.

Дифференцирующие фильтры: Roberts cross, Sobel, Prewitt, Scharr.

Поиск границ. Алгоритм Кенни (Canny). Адаптивное сглаживание. Определение порога методом Отцу (Otsu).

Преобразование Хафа (Hough). Оптимизация с учётом направления градиента. Обобщения на многопараметрический и многомерный случай.

Извлечение признаков. Признаки Хаара (Haar).

Тема 2.2 Композиция классификаторов. Бустинг (2 часа)

Понятие бустинга.

Градиентный бустинг.

AdaBoost. Алгоритм Виолы-Джонса.

LPBoost, BrownBoost.

Бустинг деревьев решений.

Тема 2.3. Нейронные сети (2часа)

Перцептрон, многослойный перецептрон.

Функции активации, методы оптимизации.

Метод обратного распространения ошибки.

Теория сложности и машинное обучение. Probably Approximately Correct learning.

Тема 2.4. Работа с текстом. Тематическое моделирование (2часа)

Задача тематического моделирования коллекции текстовых документов.

Вероятностный латентный семантический анализ PLSA. Метод максимума правдоподобия. EM-алгоритм. Элементарная интерпретация EM-алгоритма.

Латентное размещение Дирихле LDA. Метод максимума апостериорной вероятности. Сглаженная частотная оценка условной вероятности.

Небайесовская интерпретация LDA и её преимущества. Регуляризаторы разреживания, сглаживания, частичного обучения.

Аддитивная регуляризация тематических моделей. Регуляризованный EM-алгоритм, теорема о стационарной точке (применение условий Каруша–Куна–Таккера).

Рациональный EM-алгоритм. Онлайновый EM-алгоритм и его распараллеливание.

Мультимодальная тематическая модель.

Регуляризаторы классификации и регрессии.

Регуляризаторы декоррелирования и отбора тем.

Внутренние и внешние критерии качества тематических моделей.

Тема 2.5. Обучение с подкреплением (2часа)

Задача о многоруком бандите. Жадные и эпсилон-жадные стратегии. Метод UCB (upper confidence bound). Стратегия Softmax.

Среда для экспериментов.

Адаптивные стратегии на основе скользящих средних. Метод сравнения с подкреплением. Метод преследования.

Постановка задачи в случае, когда агент влияет на среду. Ценность состояния среды. Ценность действия.

Жадные стратегии максимизации ценности. Уравнения оптимальности Беллмана.

Метод временных разностей TD. Метод Q-обучения.

Градиентная оптимизация стратегии (policy gradient). Связь с максимизацией log-правдоподобия.

Постановка задачи при наличии информации о среде в случае выбора действия. Контекстный многорукий бандит.

Линейная регрессионная модель с верхней доверительной оценкой LinUCB.

Оценивание новой стратегии по большим историческим данным.

Тема 2.6. Активное обучение (2часа)

Постановка задачи машинного обучения. Основные стратегии: отбор объектов из выборки и из потока, синтез объектов.

Сэмплирование по неуверенности. Почему активное обучение быстрее пассивного.

Сэмплирование по несогласию в комитете. Сокращение пространства решений.

Сэмплирование по ожидаемому изменению модели.

Сэмплирование по ожидаемому сокращению ошибки.

Синтез объектов по критерию сокращения дисперсии.

Взвешивание по плотности.

Оценивание качества активного обучения.

Введение изучающих действий в стратегию активного обучении. Алгоритмы ε-active и EG-active.

Применение обучения с подкреплением для активного обучения. Активное томпсоновское сэмплирование.

Тема 2.7. Рекомендательные системы (2часа)

Задачи коллаборативной фильтрации, транзакционные данные и матрица субъекты—объекты.

Корреляционные методы user-based, item-based. Задача восстановления пропущенных значений. Меры сходства субъектов и объектов.

Латентные методы на основе би-кластеризации. Алгоритм Брегмана.

Латентные методы на основе матричных разложений. Метод главных компонент для разреженных данных (LFM, Latent Factor Model). Метод стохастического градиента.

Неотрицательные матричные разложения. Метод чередующихся наименьших квадратов ALS.

Модель с учётом неявной информации (implicit feedback).

Рекомендации с учётом дополнительных признаковых данных. Линейная и квадратичная регрессионные модели, libFM.

Измерение качества рекомендаций. Меры разнообразия (diversity), новизны (novelty), покрытия (coverage), догадливости (serendipity).

Тема 2.8. Сверточные нейронные сети. Рекуррентные нейронные сети (2часа)

Свёрточные нейронные сети (CNN). Свёрточный нейрон. Pooling нейрон. Выборка размеченных изображений ImageNet.

Идея обобщения CNN на любые структурированные данные.

Рекуррентные нейронные сети (RNN). Обучение рекуррентных сетей: Backpropagation Through Time (BPTT).

Сети долгой кратковременной памяти (Long short-term memory, LSTM).

Автокодировщики. Векторные представления дискретных данных.

Тема 2.9. Генеративные модели (2часа)

Моделирование случайных данных

Неслучайные модельные данные

Тема 2.10. Ранжирование (2часа)

Постановка задачи обучения ранжированию. Примеры.

Признаки в задаче ранжирования поисковой выдачи: текстовые, ссылочные, кликовые. TF-IDF. PageRank.

Критерии качества ранжирования: Precision, MAP, AUC, DCG, NDCG, pFound.

Ранговая классификация, OC-SVM.

Попарный подход: RankingSVM, RankNet, LambdaRank.

Тема 2.11. Кластеризация и частичное обучение (2часа)

Постановка задачи кластеризации. Примеры прикладных задач. Типы кластерных структур.

Постановка задачи Semisupervised Learning, примеры приложений.

Оптимизационные постановки задач кластеризации и частичного обучения.

Алгоритм k-средних и EM-алгоритм для разделения гауссовской смеси.

Графовые алгоритмы кластеризации. Выделение связных компонент. Кратчайший незамкнутый путь.

Алгоритм ФОРЭЛ.

Алгоритм DBSCAN.

Агломеративная кластеризация, Алгоритм Ланса-Вильямса и его частные случаи.

Алгоритм построения дендрограммы. Определение числа кластеров.

Свойства сжатия/растяжения, монотонности и редуктивности. Псевдокод редуктивной версии алгоритма.

Простые эвристические методы частичного обучения: self-training, co-training, co-learning.

Трансдуктивный метод опорных векторов TSVM.

Алгоритм Expectation-Regularization на основе многоклассовой регуляризированной логистической регрессии.

Тема 2.12. Основы статистики для машинного обучения (2часа)

Статистический критерий информативности, точный тест Фишера. Сравнение областей эвристических и статистических закономерностей. Асимптотическая эквивалентность статистического и энтропийного критерия информативности. Разнообразие критериев информативности в (p,n)-пространстве.

Тема 2.13. Доверительные интервалы. Критерии согласия (2часа)

Тема 2.14. АБ-тестирование. Двухвыборочные критерии (2часа)

Тема 2.15. Корреляционный анализ (2часа)

Тема 2.16. Методы частичного обучения (2часа)

Тема 2.17. Автоматическое обучение машин (AutoML) (2часа)

Автоматическая подготовка данных и сбор и сохранение данных (из сырых данных и разнообразных форматов)

Автоматическое конструирование признаков

Автоматический выбор модели

Оптимизация гиперпараметров алгоритма обучения и характеризации

Автоматический выбор каналов по времени, памяти и ограничений сложности

Автоматический выбор метрик оценки / процедур валидации

Автоматическая проверка задач

Автоматический анализ полученных результатов

Пользовательские результаты и визуализация для автоматического обучения машин

III. СТРУКТУРА И СОДЕРЖАНИЕ ПРАКТИЧЕСКОЙ ЧАСТИ КУРСА

(нет)

IV. УЧЕБНО-МЕТОДИЧЕСКОЕ ОБЕСПЕЧЕНИЕ САМОСТОЯТЕЛЬНОЙ РАБОТЫ ОБУЧАЮЩИХСЯ

Учебно-методическое обеспечение самостоятельной работы обучающихся по дисциплине «Математические методы машинного обучения» представлено в Приложении 1 и включает в себя:

- план-график выполнения самостоятельной работы по дисциплине, в том числе примерные нормы времени на выполнение по каждому заданию;
- характеристика заданий для самостоятельной работы обучающихся и методические рекомендации по их выполнению;
- требования к представлению и оформлению результатов самостоятельной работы;
- критерии оценки выполнения самостоятельной работы.

КОНТРОЛЬ ДОСТИЖЕНИЯ ЦЕЛЕЙ КУРСА

№ п/п	Контролируемые разделы / темы дисциплины	Коды и этапы формирования компетенций		Оценочные средства	
				текущий контроль	промежуточная аттестация
1	Тема 1.1. Методы оптимизации. Градиентный спуск	ОК-3 ОК-6 ОК-7 ОПК-5 ОПК-6 ПК-8 ПК-15	знает умеет владеет	ПР-2 ПР-11 ТС	УО-1
2	Тема 1.2. Линейная регрессия	ОК-3 ОК-6 ОК-7 ОПК-5 ОПК-6 ПК-8 ПК-15	знает умеет владеет	ПР-2 ПР-11 ТС	УО-1
3	Тема 1.3 Глобальная оптимизация.	ОК-3 ОК-6	знает	ПР-2 ПР-11	УО-1

	Генетический алгоритм	ОК-7 ОПК-5 ОПК-6 ПК-8 ПК-15	умеет владеет	ТС	
4	Тема 1.4 Метод ближайших соседей (k-NN)	ОК-3 ОК-6 ОК-7 ОПК-5 ОПК-6 ПК-8 ПК-15	знает умеет владеет	ПР-2 ПР-11 ТС	УО-2
4	Тема 1.5 Наивный байесов классификатор	ОК-3 ОК-6 ОК-7 ОПК-5 ОПК-6 ПК-8 ПК-15	знает умеет владеет	ПР-2 ПР-11 ТС	УО-1
4	Тема 1.6 Логистическая регрессия	ОК-3 ОК-6 ОК-7 ОПК-5 ОПК-6 ПК-8 ПК-15	знает умеет владеет	ПР-2 ПР-11 ТС	УО-1
4	Тема 1.7 Деревья решений	ОК-3 ОК-6 ОК-7 ОПК-5 ОПК-6 ПК-8 ПК-15	знает умеет владеет	ПР-2 ПР-11 ТС	УО-1
4	Тема 1.8 Кластеризация	ОК-3 ОК-6 ОК-7 ОПК-5 ОПК-6 ПК-8 ПК-15	знает умеет владеет	ПР-2 ПР-11 ТС	УО-2

4	Тема 1.9 Работа с текстом	ОК-3 ОК-6 ОК-7 ОПК-5 ОПК-6 ПК-8 ПК-15	знает умеет владеет	ПР-2 ПР-11 ТС	УО-1
4	Тема 1.10 Снижение размерности	ОК-3 ОК-6 ОК-7 ОПК-5 ОПК-6 ПК-8 ПК-15	знает умеет владеет	ПР-2 ПР-11 ТС	УО-1
4	Тема 1.11 Метод опорных векторов (SVM)	ОК-3 ОК-6 ОК-7 ОПК-5 ОПК-6 ПК-8 ПК-15	знает умеет владеет	ПР-2 ПР-11 ТС	УО-2

2 СЕМЕСТР

№ п/п	Контролируемые разделы / темы дисциплины	Коды и этапы формирования компетенций		Оценочные средства	
				текущий контроль	промежуточная аттестация
1	Тема 2.1 Работа с изображениями	ОК-3 ОК-6 ОК-7 ОПК-5 ОПК-6 ПК-8 ПК-15	знает умеет владеет	ПР-2 ПР-11 ТС	УО-1
2	Тема 2.2 Композиция классификаторов. Бустинг	ОК-3 ОК-6 ОК-7 ОПК-5 ОПК-6 ПК-8 ПК-15	знает умеет владеет	ПР-2 ПР-11 ТС	УО-1

3	Тема 2.3. Нейронные сети	ОК-3 ОК-6 ОК-7 ОПК-5 ОПК-6 ПК-8 ПК-15	знает умеет владеет	ПР-2 ПР-11 ТС	УО-1
4	Тема 2.4. Работа с текстом. Тематическое моделирование	ОК-3 ОК-6 ОК-7 ОПК-5 ОПК-6 ПК-8 ПК-15	знает умеет владеет	ПР-2 ПР-11 ТС	УО-2
4	Тема 2.5. Обучение с подкреплением	ОК-3 ОК-6 ОК-7 ОПК-5 ОПК-6 ПК-8 ПК-15	знает умеет владеет	ПР-2 ПР-11 ТС	УО-1
4	Тема 2.6. Активное обучение	ОК-3 ОК-6 ОК-7 ОПК-5 ОПК-6 ПК-8 ПК-15	знает умеет владеет	ПР-2 ПР-11 ТС	УО-1
4	Тема 2.7. Рекомендательные системы	ОК-3 ОК-6 ОК-7 ОПК-5 ОПК-6 ПК-8 ПК-15	знает умеет владеет	ПР-2 ПР-11 ТС	УО-1
4	Тема 2.8. Сверточные нейронные сети. Рекуррентные нейронные сети	ОК-3 ОК-6 ОК-7 ОПК-5	знает умеет	ПР-2 ПР-11 ТС	УО-2

		ОПК-6 ПК-8 ПК-15	владеет		
4	Тема 2.9. Генеративные модели	ОК-3 ОК-6 ОК-7 ОПК-5 ОПК-6 ПК-8 ПК-15	знает умеет владеет	ПР-2 ПР-11 ТС	УО-1
4	Тема 2.10. Ранжирование	ОК-3 ОК-6 ОК-7 ОПК-5 ОПК-6 ПК-8 ПК-15	знает умеет владеет	ПР-2 ПР-11 ТС	УО-1
4	Тема 2.11. Кластеризация и частичное обучение	ОК-3 ОК-6 ОК-7 ОПК-5 ОПК-6 ПК-8 ПК-15	знает умеет владеет	ПР-2 ПР-11 ТС	УО-1
2	Тема 2.12. Основы статистики для машинного обучения	ОК-3 ОК-6 ОК-7 ОПК-5 ОПК-6 ПК-8 ПК-15	знает умеет владеет	ПР-2 ПР-11 ТС	УО-2
3	Тема 2.13. Доверительные интервалы. Критерии согласия	ОК-3 ОК-6 ОК-7 ОПК-5 ОПК-6 ПК-8 ПК-15	знает умеет владеет	ПР-2 ПР-11 ТС	УО-1
4	Тема 2.14. АБ-тестирование.	ОК-3 ОК-6	знает	ПР-2 ПР-11	УО-1

	Двухвыборочные критерии	ОК-7 ОПК-5 ОПК-6 ПК-8 ПК-15	умеет владеет	ТС	
4	Тема 2.15. Корреляционный анализ	ОК-3 ОК-6 ОК-7 ОПК-5 ОПК-6 ПК-8 ПК-15	знает умеет владеет	ПР-2 ПР-11 ТС	УО-1
4	Тема 2.16. Методы частичного обучения	ОК-3 ОК-6 ОК-7 ОПК-5 ОПК-6 ПК-8 ПК-15	знает умеет владеет	ПР-2 ПР-11 ТС	УО-1
4	Тема 2.17. Автоматическое обучение машин (AutoML)	ОК-3 ОК-6 ОК-7 ОПК-5 ОПК-6 ПК-8 ПК-15	знает	ПР-2 ПР-11 ТС	УО-2

1. устный опрос (УО): собеседование (УО-1), коллоквиум (УО-2); итоговая презентация (УО-3); круглый стол (УО-4);
2. технические средства контроля (ТС);
3. письменные работы (ПР): тесты (ПР-1), контрольные работы (ПР-2), эссе (ПР-3), рефераты (ПР-4), курсовые работы (ПР-5), научно-учебные отчеты по практикам (ПР-6), конспект (ПР-7), проект (ПР-9). Разноуровневые задачи и задания (ПР-11) и т.п.

Типовые индивидуальные задания, методические материалы, определяющие процедуры оценивания знаний, умений и навыков и (или) опыта деятельности, а также критерии и показатели, необходимые для оценки знаний, умений, навыков и характеризующие этапы формирования компетенций в процессе освоения образовательной программы, представлены в Приложении 2.

V. СПИСОК УЧЕБНОЙ ЛИТЕРАТУРЫ И ИНФОРМАЦИОННО-МЕТОДИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

Основная литература
(электронные и печатные издания)

1. Неделько В.М. Основы статистических методов машинного обучения [Электронный ресурс]: учебное пособие/ Неделько В.М.— Электрон. текстовые данные. — Новосибирск: Новосибирский государственный технический университет, 2010. — 72 с.— Режим доступа: http://www.iprbookshop.ru/45418.html. — ЭБС «IPRbooks»
2. Коэльо, Л.П. Построение систем машинного обучения на языке Python [Электронный ресурс] / Л.П. Коэльо, В. Ричарт; пер. с англ. Слинкин А. А.. — Электрон. дан. — Москва: ДМК Пресс, 2016. — 302 с. — Режим доступа: https://e.lanbook.com/book/82818. — Загл. с экрана.
3. Флах, П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных [Электронный ресурс] / П. Флах. — Электрон. дан. — Москва: ДМК Пресс, 2015. — 400 с. — Режим доступа: https://e.lanbook.com/book/69955. — Загл. с экрана.
4. Рашка, С. Python и машинное обучение: крайне необходимое пособие по новейшей предсказательной аналитике, обязательное для более глубокого понимания методологии машинного обучения [Электронный ресурс]: руководство / С. Рашка; пер. с англ. Логунова А.В. — Электрон. дан. — Москва: ДМК Пресс, 2017. — 418 с. — Режим доступа: https://e.lanbook.com/book/100905. — Загл. с экрана.
5. Шарден, Б. Крупномасштабное машинное обучение вместе с Python [Электронный ресурс]: учебное пособие / Б. Шарден, Л. Массарон, А. Боскетти; пер. с анг. А. В. Логунова. — Электрон. дан. — Москва: ДМК Пресс, 2018. — 358 с. — Режим доступа: https://e.lanbook.com/book/105836. — Загл. с экрана.

6. Кук, Д. Машинное обучение с использованием библиотеки H2O [Электронный ресурс] / Д. Кук; пер. с англ. Огурцова А.Б.. — Электрон. дан. — Москва: ДМК Пресс, 2018. — 250 с. — Режим доступа: https://e.lanbook.com/book/97353. — Загл. с экрана.

Дополнительная литература
(печатные и электронные издания)

1. Информационные аналитические системы [Электронный ресурс]: учебник / Т. В. Алексеева, Ю. В. Амириди, В. В. Дик и др.; под ред. В. В. Дика. - М.: МФПУ Синергия, 2013. - 384 с. - (Университетская серия). - ISBN 978-5-4257-0092-6, http://www.znanium.com/bookread.php?book=451186

2. Домингос, П. Верховный алгоритм: как машинное обучение изменит наш мир [Электронный ресурс] Москва: Манн, Иванов и Фербер, 2016. 336 с. https://e.lanbook.com/book/91645.

3. Гаврилова, И.В. Основы искусственного интеллекта [Электронный ресурс]: учеб. пособие / И.В. Гаврилова, О.Е. Масленникова. Москва: ФЛИНТА, 2013. 282 с. https://e.lanbook.com/book/44749. 4. Ясницкий, Л.Н. Интеллектуальные системы [Электронный ресурс]: учеб. пособ./ Москва : Издательство 'Лаборатория знаний', 2016. 224 с. Режим доступа: https://e.lanbook.com/book/90254.

Перечень ресурсов
информационно-телекоммуникационной сети «Интернет»

1. Байесовские_методы_машинного_обучения_(курс_лекций)_/_2017 Д.П. Ветров - http://www.machinelearning.ru/wiki/index.php?title= Байесовские_методы_машинного_обучения_(курс_лекций)_/_2017 Д.П. Ветров

2. Машинное обучение (курс лекций, Н.Ю. Золотых) - http://www.uic.unn.ru/~zny/ml/
3. Машинное_ обучение_(курс_ лекций С.К.Воронцов). - http://www.machinelearning.ru/wiki/index.php?title=Машинное_ обучение_(курс_ лекций%2C_К.В.Воронцов)
4. Курс «Введение в машинное обучение», К.В.Воронцов (ВШЭ и Яндекс).Хабр об этом курсе.
5. Специализация «Машинное обучение и анализ данных» (МФТИ и Яндекс). Хабр об этом курсе.
6. Машинное обучение (семинары,ФУПМ МФТИ)
7. Машинное обучение (семинары, ВМК МГУ)
8. Машинное обучение (курс лекций, Н.Ю.Золотых)
9. Машинное обучение (курс лекций, СГАУ, С.Лисицын)

VI. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ПО ОСВОЕНИЮ ДИСЦИПЛИНЫ

На изучение дисциплины отводится 72 часа аудиторных (лекционных) занятий 72 часа на самостоятельную работу и 72 часа на контроль. Формы работы: лекции, самостоятельная работа с учебной и научной литературой, самостоятельное выполнение индивидуальных заданий, консультации. На занятиях перед выдачей индивидуальных заданий преподаватель объясняет теоретический материал по заданной теме. Вводит основные требования к его выполнению. Приводит примеры.

По ряду тем студентам предлагается работать самостоятельно, выполняя полный обзор по теме. Преподаватель контролирует работу студентов, отвечает на возникающие вопросы, предоставляет список литературных источников для освоения темы, а также перечень вопросов для самопроверки.

После выполнения задания, студент оформляет материал в форме программного кода и отправляет его на проверку преподавателю по электронной почте, либо предъявляет на компьютере во время занятия. Студент отвечает устно во время занятия по заданной теме.

Рекомендации по подготовке к экзамену

Рекомендуется регулярное посещение всех учебных занятий в течение всего семестра: лекций, консультаций и т.п., а также активное изучение рекомендованной литературы, и выполнение в установленные сроки всех индивидуальных заданий.

При ответе на каждый вопрос экзамена студент должен продемонстрировать знание определения указанного понятия, связанных с ним особенностей реализации и применения, умение реализовать указанную операцию, а также навыки иллюстрации теоретических принципов на предложенных простых примерах.

VII. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

Компьютерный класс: Проектор DLP, 3000 ANSI Lm, WXGA 1280x800, 2000:1 EW330U Mitsubishi,; Моноблок HP ProOne 440 G3 23.8" All-in-One, диагональ экрана 23.8", разрешение экрана 1920x1080, Bluetooth, Wi-Fi, операционная система: Windows 10 Enterprise, оптический привод DVD, процессор: Intel Core i5-7500T, размер оперативной памяти: 8 ГБ, видеопроцессор: Intel HD Graphics 630, объем жесткого диска: 1Tb. Беспроводные ЛВС для обучающихся обеспечены системой на базе точек доступа 802.11a/b/g/n 2x2 MIMO(2SS). Специализированное ПО: Matlab, Simulink, Visual Studio 2019	690922, Приморский край, г. Владивосток, о. Русский, п. Аякс, 10, г. Владивосток, о. Русский, п. Аякс , корпус G, ауд. G468

Приложение 1

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение
высшего профессионального образования
«Дальневосточный федеральный университет»
(ДВФУ)

ШКОЛА ЕСТЕСТВЕННЫХ НАУК

УЧЕБНО-МЕТОДИЧЕСКОЕ ОБЕСПЕЧЕНИЕ САМОСТОЯТЕЛЬНОЙ РАБОТЫ ОБУЧАЮЩИХСЯ

по дисциплине «Математические методы машинного обучения**»**

Направление подготовки – 09.04.01 Информатика и вычислительная техника

магистерская программа «Технологии виртуальной и дополненной реальности»

Форма подготовки очная

Владивосток
2018

План-график выполнения самостоятельной работы по дисциплине

№ п/п, название	Дата/сроки выполнения	Вид СРС	Примерные нормы времени на выполнение	Форма контроля
1. Методы оптимизации. Градиентный спуск	2 недели	ИДЗ	2 недели	Проверка программы
2. Линейная регрессия	2 недели	ИДЗ	2 недели	Проверка программы
3. Глобальная оптимизация. Генетический алгоритм	2 недели	ИДЗ	2 недели	Проверка программы
4. Метод ближайших соседей (k-NN)	2 недели	ИДЗ	2 недели	Проверка программы
5. Наивный байесов классификатор	2 недели	ИДЗ	2 недели	Проверка программы
6. Логистическая регрессия	2 недели	ИДЗ	2 недели	Проверка программы
7. Деревья решений	2 недели	ИДЗ	2 недели	Проверка программы
8. Кластеризация	2 недели	ИДЗ	2 недели	Проверка программы
9. Работа с текстом	2 недели	ИДЗ	2 недели	Проверка программы

Критерии оценивания

Действует балльно-рейтинговая система оценки знаний обучающихся. Суммарно по дисциплине (модулю) можно получить максимум 100 баллов за семестр.

В течение семестра студентам последовательно выдается набор из 9-ти лабораторных работ, каждая из которых имеет вес от 5% до 10%.

Посещаемость занятий также учитывается и имеет вес 2%. Для получения экзамена в 1-ом семестре необходимо:

86 баллов и более – "отлично",

71-85 баллов – "хорошо",

56-70 баллов – "удовлетворительно",

55 баллов и менее – "неудовлетворительно".

ТИПОВЫЕ ИНДИВИДУАЛЬНЫЕ ЗАДАНИЯ

Задача A. Градиентный спуск

Входной файл: Стандартный вход
Выходной файл: Стандартный выход
Ограничение времени: 1 сек
Ограничение памяти: 512 Мб

- ***Условие***

Требуется реализовать класс на языке Python, который соответствует следующему интерфейсу.

```
class GradientOptimizer:
    def __init__(self, oracle, x0):
        self.oracle = oracle
        self.x0 = x0

    def optimize(self, iterations, eps, alpha):
        pass
```

В конструктор принимаются два аргумента — оракул, с помощью которого можно получить градиент оптимизируемой функции, а также точку, с которой необходимо начать градиентный спуск.

Метод `optimize` принимает максимальное число итераций для критерия остановки, L2-норму градиента, которую можно считать оптимальной, а также learning rate. Метод возвращает оптимальную точку.

Оракул имеет следующий интерфейс:

```
class Oracle:
    def get_func(self, x)
    def get_grad(self, x)
```

`x` имеет тип `np.array` вещественных чисел.

- ***Формат выходных данных***

Код должен содержать только класс и его реализацию. Он не должен ничего выводить на экран.

Задача B. Линейная регрессия. Основы

Входной файл: Стандартный вход
Выходной файл: Стандартный выход
Ограничение времени: 1 сек
Ограничение памяти: 512 Мб

- ***Условие***

Требуется реализовать следующие функции на языке Python.

```
def linear_func(theta, x)                  # function value
def linear_func_all(theta, X)              # 1-d np.array of function values of
all rows of the matrix X
def mean_squared_error(theta, X, y)        # MSE value of current regression
def grad_mean_squared_error(theta, X, y)   # 1-d array of gradient by theta
```

`theta` — одномерный `np.array`
`x` — одномерный `np.array`

`X` — двумерный `np.array`. Каждая строка соответствует по размерности вектору `theta`
`y` — реальные значения предсказываемой величины
Матрица XX имеет размер M×NM×N. MM строк и NN столбцов.
Используется линейная функция вида: $h_\theta(x)=\theta_1 x_1+\theta_2 x_2+...+\theta_n x_N h_\theta(x)=\theta 1x1+\theta 2x2+...+\theta nxN$
Mean squared error (MSE) как функция от θθ: $J(\theta)=1_M M\sum_{i=1}(y_i-h_\theta(x_{(i)}))_2 J(\theta)=1M\sum i=1M(yi-h\theta(x(i)))2$.
Где $x_{(i)}x(i)$ — ii-я строка матрицы XX
Градиент функции MSE: $\nabla J(\theta)=\{\partial J\partial\theta_1,\partial J\partial\theta_2,...,\partial J\partial\theta_N\}\nabla J(\theta)=\{\partial J\partial\theta 1,\partial J\partial\theta 2,...,\partial J\partial\theta N\}$

- ***Пример***

```
X = np.array([[1,2],[3,4],[4,5]])
theta = np.array([5, 6])
y = np.array([1, 2, 1])
linear_func_all(theta, X) # -> array([17, 39, 50])
mean_squared_error(theta, X, y) # -> 1342.0
grad_mean_squared_error(theta, X, y) # -> array([215.33333333, 283.33333333])
```

- ***Формат выходных данных***

Код должен содержать только реализацию функций.

Задача C. Найти линейную регрессию

Входной файл: Стандартный вход
Выходной файл: Стандартный выход
Ограничение времени: 10 сек
Ограничение памяти: 512 Мб

- ***Условие***

Требуется реализовать функцию на языке Python, которая находит линейную регрессию заданных векторов, используя метрику MSE.

```
def fit_linear_regression(X, y)   # np.array of linear regression coefs
```

`X` — двумерный `np.array`. Каждая строка соответствует отдельному примеру.
`y` — реальные значения предсказываемой величины

- ***Формат выходных данных***

Код должен содержать только реализацию функций.

Задача A. Распределение задач

Входной файл: input.txt
Выходной файл: output.txt
Ограничение времени: 1 сек
Ограничение памяти: 256 Мб

- ***Условие***

Группа разработчиков работает над проектом. Весь проект разбит на задачи, для каждой задачи указывается ее категория сложности (1, 2, 3 или 4), а также оценочное время выполнения задачи в часах. Проект считается выполненным, если выполнены все задачи. Для каждого разработчика и для каждой категории сложности задачи указывается коэффициент, с которым, как ожидается, будет соотноситься реальное время выполнения задачи данным разработчиком к оценочному времени. Считается, что все разработчики начинают работать с проектом в одно и тоже время и выделяют для работы одинаковое время. Необходимо реализовать программу, распределяющую задачи по разработчикам, с целью минимизировать время выполнения проекта (получить готовый проект за минимальный промежуток времени). Поиск решения необходимо реализовать с помощью генетического алгоритма.

- ***Отправка решения и тестирование***

Данная задача будет проверяться на *ОДНОМ* входном файле. Этот файл можно скачать ЗДЕСЬ.

В качестве решения принимается текстовый файл, содержащий ответ к задаче в требуемом формате (при его отправке следует выбрать в тестирующей системе среду разработки "`Answer text`").
Решение набирает количество баллов, вычисляемое по следующей формуле: Score=$10\frac{6T_{max}}{}$Score=106Tmax. T_{max}Tmax — наибольшее среди всех разработчиков время, затраченное на выполнение выданных соответствующему разработчику задач.

- ***Формат входного файла***

Первая строка входного файла содержит целое число NN количество задач.
Вторая строка — NN целых чисел от 1 до 4 категорий сложности задач.
Третья строка — NN вещественных положительных чисел оценочного времени для задач.
Четвертая строка – целое число MM, количество разработчиков.
Следующие MM строк содержат по 4 вещественных положительных числа — коэффициенты каждого разработчика.

- ***Формат выходного файла***

Первая и единственная строка выходного файла содержит NN целых чисел $w_i$wi — номер разработчика, назначенного на ii - ю задачу.

- ***Ограничения***

- ***Примеры тестов***

№	Входной файл (`input.txt`)	Выходной файл (`output.txt`)
1	3 1 1 4 5.2 3.4 4 2 1 1 2 5 0.7 1 1.2 1.5	1 2 2

Задача A. Логистическая регрессия. Основы

Входной файл: Стандартный вход
Выходной файл: Стандартный выход
Ограничение времени: 1 сек
Ограничение памяти: 512 Мб

- ***Условие***

Требуется реализовать следующие функции на языке Python.

```
def logistic_func(theta, x)                  # function value
def logistic_func_all(theta, X)              # 1-d np.array of function values
of all rows of the matrix X
def cross_entropy_loss(theta, X, y)        # cross entropy loss value of
current regression
def grad_cross_entropy_loss(theta, X, y)    # 1-d array of gradient by theta
```

`theta` — одномерный `np.array`
`x` — одномерный `np.array`
`X` — двумерный `np.array`. Каждая строка соответствует по размерности вектору `theta`
`y` — реальные значения предсказываемой величины
Матрица XX имеет размер M×NM×N. MM строк и NN столбцов.
Используется линейная функция вида: $h_\theta(x)=\theta_1x_1+\theta_2x_2+...+\theta_nx_Nh\theta(x)=\theta1x1+\theta2x2+...+\thetanxN$

- ***Формат выходных данных***

Код должен содержать только реализацию функций.

Задача B. Найти логистическую регрессию

Входной файл: Стандартный вход
Выходной файл: Стандартный выход
Ограничение времени: 10 сек
Ограничение памяти: 512 Мб

- ***Условие***

Требуется реализовать функцию на языке Python, которая находит логистическую регрессию заданных векторов, используя метрику cross entropy loss.

```
def fit_logistic_regression(X, y)   # np.array of logistic regression coefs
```

`X` — двумерный `np.array`. Каждая строка соответствует отдельному примеру.
`y` — реальные значения предсказываемой величины

- ***Формат выходных данных***

Код должен содержать только реализацию функций.

Задача A. News category

Входной файл: input.txt
Выходной файл: output.txt
Ограничение времени: 1 сек
Ограничение памяти: 256 Мб

- ***Условие***

Требуется обучить модель определения категории новости. Обучающую выборку можно скачать ЗДЕСЬ. Категория новости в обучающей выборке представлена столбцом `CAT`.

- `HEADER` — заголовок новости
- `MEDIANAME` — название СМИ
- `WEBSITE` — вебсайт СМИ
- `PTIME` — время публикации

Для определения качества модели будет использоваться тестовая выборка, доступная ЗДЕСЬ.

В тестовой выборке требуется предсказать значения столбца `CAT`, соответствующие каждому тестовому примеру. Категории новостей кодируются одним символом, аналогично данным в обучающей выборке.

- ***Отправка решения и тестирование***

Данная задача будет проверяться на *ОДНОМ* входном файле.
В качестве решения принимается текстовый файл, содержащий ответ к задаче в требуемом формате (при его отправке следует выбрать в тестирующей системе среду разработки "`Answer text`").
Решение набирает количество баллов, вычисляемое по следующей формуле: $Score=10^5 \cdot AccuracyScoreScore=105 \cdot AccuracyScore$. AccuracyScoreAccuracyScore — доля верно классифицированных новостей относительно всех новостей в тестовой выборке.

- ***Формат выходного файла***

Каждая строка выходного файла должна содержать единственный символ, задающий категорию соответствующего тестового примера.

Задача A. Качество вина

Входной файл: input.txt
Выходной файл: output.txt
Ограничение времени: 1 сек
Ограничение памяти: 256 Мб

- ***Условие***

Требуется обучить модель определения качества вина. Качество вина определяется по 1010-балльной шкале. В данной задаче будем использовать бинарную модель и предсказывать, "хорошее" вино или "плохое". Хорошим будем считать вино с качеством строго выше 66. Обучающую выборку можно скачать ЗДЕСЬ. Качество вина представлено столбцом `quality`.

Для определения качества модели будет использоваться тестовая выборка, доступная ЗДЕСЬ.

В тестовой выборке требуется предсказать значения 11 или 00, "хорошее" вино или "плохое" соответственно, для каждого примера. Оценку качества по 1010-балльной шкале предсказывать не требуется.

- ## *Отправка решения и тестирование*

Данная задача будет проверяться на *ОДНОМ* входном файле.
В качестве решения принимается текстовый файл, содержащий ответ к задаче в требуемом формате (при его отправке следует выбрать в тестирующей системе среду разработки "`Answer text`").
Решение набирает количество баллов, вычисляемое по следующей формуле: $Score=10_5 \cdot F1 Score=105 \cdot F1$.

- ## *Формат выходного файла*

Каждая строка выходного файла должна содержать целое число 11 или 00. Количество строк должно быть равно количеству элементов контрольной выборки.

Приложение 2

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение
высшего профессионального образования
«Дальневосточный федеральный университет»
(ДВФУ)

ШКОЛА ЕСТЕСТВЕННЫХ НАУК

ФОНД ОЦЕНОЧНЫХ СРЕДСТВ

по дисциплине «Математические методы машинного обучения**»**

Направление подготовки – 09.04.01 Информатика и вычислительная техника

магистерская программа «Технологии виртуальной и дополненной реальности»

Форма подготовки очная

Владивосток
2018

Сформированность каждой компетенции в рамках освоения данной дисциплины оценивается по трех уровневой шкале:

- пороговый уровень является обязательным для всех обучающихся по завершении освоения дисциплины;
- продвинутый уровень характеризуется превышением минимальных характеристик сформированности компетенции по завершении освоения дисциплины;
- эталонный уровень характеризуется максимально возможной выраженностью компетенции и является важным качественным ориентиром для самосовершенствования.

Уровень сформированности каждой компетенции на различных этапах ее формирования в процессе освоения данной дисциплины оценивается в ходе текущего контроля успеваемости представлен различными видами оценочных средств.

Результаты обучения (компетенции из ФГОС)	Знает	Умеет	Владеет
ОК-3 ОК-6 ОК-7 ОПК-5 ОПК-6 ПК-8 ПК-15	- современное состояние исследований в области машинного обучения; - принципы построения систем машинного обучения; - модели представления и описания технологий машинного обучения.	- проводить анализ предметной области; - определять назначение, выбирать методы и средства для построения систем машинного обучения;	Навыками: - использования аппарата простейшего анализ данных; - применения методов классификации информации; - реализации алгоритмов машинного обучения.
Эталонный	Основной и дополнительный материал, предусмотренный компетенцией, без ошибок и	**Умеет** в полном объеме …	всеми навыками, демонстрируя их не только в стандартных ситуациях, но и при решении нестандартных задач

	погрешностей		
Продвинутый	основной материал, предусмотренный компетенцией, без ошибок и погрешностей	**Умеет** с незначительными погрешностями …	основными навыками, демонстрируя их в стандартных ситуациях, в том числе при решении дополнительных задач
Пороговый	большинство основных понятий, изучаемых в рамках дисциплины	**Умеет** с погрешностями …	некоторыми основными навыками, демонстрируя их в стандартных ситуациях

МАТЕРИАЛЫ ЭКЗАМЕНА
по дисциплине «Математические методы машинного обучения**»**

Основные вопросы

Тема 1.1 Введение. Методы оптимизации. Градиентный спуск

1. Задача регрессии, классификации.
2. Функция потерь. Оптимизация. Перебор по сетке.
3. Производная, частные производные, градиент.
4. Градиентный спуск, проблема выбора шага.
5. Стохастический градиентный спуск. Использование момента.
6. Adagrad, Adadelta, Adam.
7. RMSProp*.

Тема 1.2 Линейная регрессия

8. Постановка задачи линейной регрессии.
9. Метод наименьших квадратов.
10. Ковариация, корреляция.
11. Критерий R2.
12. Анализ остатков.

Тема 1.3 Глобальная оптимизация. Генетический алгоритм

13. Многопараметрическая оптимизация.
14. Доминация и оптимальность по Парето.
15. Функция качества (fitness). Аппроксимация качества.
16. Общая идея генетического алгоритма.
17. Представление генома.

18. Методы селекции: пропорционально качеству, универсальная выборка (stochastic universal sampling), с наследием (reward-based), турнир. Стратегия элитизма.
19. Методы кроссовера. Двух и многоточечный, равномерный (по подмножествам), для перестановок.
20. Мутация. Влияние на скорость обучения.
21. Управление популяцией. Сегрегация, старение, распараллеливание.
22. Генетическое программирование.

Тема 1.4 Метод ближайших соседей (k-NN)

23. Понятие и свойства метрики. Ослабление требования к неравенству треугольника.
24. Базовый алгоритм классификации методом 1-NN и k-NN. Преимущества и недостатки.
25. Метрики L1, L2, Хемминга, Левенштейна, косинусное расстояние.
26. Потеря точности нормы в высоких размерностях.
27. Нормализация координат. Предварительная трансформация пространства признаков.
28. Метрика Махаланобиса.
29. Кросс-валидация методом "без одного" (leave one out).
30. Определение границ, показатель пограничности.
31. Сжатие по данным. Понятия выброса, прототипа, усвоенной точки. Алгоритм Харта (Hart).
32. Регрессия методом k-NN.
33. Взвешенные соседи.
34. Связь с градиентным спуском. Стохастическая формулировка, softmax.
35. Метод соседних компонент (neighbour component analysis)*.
36. Связь с выпуклой оптимизацией. Метод большого запаса (Large margin NN)
37. Оптимизация классификатора, k-d деревья
38. Хеши чувствительные к локальности, хеши сохраняющие локальность*.

Тема 1.5 Наивный байесов классификатор

39. Условная вероятность. Байесово решающее правило. Обновление вероятностей.
40. Наивный классификатор, предположение о независимости признаков.
41. Оценка плотности распределения для числовых признаков.
42. Алгоритмические оптимизации.
43. Алгоритм EM.

Тема 1.6 Логистическая регрессия

44. Сигмоид
45. Метод наибольшего правдоподобия
46. Логистическая регрессия для меток −1, 1

Тема 1.7 Деревья решений

47. Понятие дерева решений.
48. Борьба с оверфиттингом: bagging, выборки признаков.
49. Ансамбли, случайный лес (Random Forest).
50. Понятие энтропии, определение информации по Шеннону.
51. Метрики: примеси Джини (Gini impurity), добавленная информация (information gain).
52. Деревья регрессии. Метрика вариации.
53. Непрерывные признаки. Использование главных компонент вместо признаков.
54. Сокращение дерева (pruning).

Тема 1.8 Кластеризация

55. Задача обучения без учителя, применения при эксплораторном анализе
56. Неметрическая кластеризация: функция схожести, компоненты связности и остовные деревья, иерархическая кластеризация снизу вверх
57. Метрики, понятие центроида и представителя класса
58. Центроидные алгоритмы: k-means, k-medoid
59. Алгоритмы, основанные на плотности: DBSCAN, OPTICS
60. Алгоритмы, основанные на распределении: сумма гауссиан
61. Нечёткая кластеризация, алгоритм c-means

62. Метрики качества: leave-one-out, силуэт, индекс Дэвиса-Болдина (Davies-Bouldin), индекс Данна (Dunn)

Тема 1.9 Работа с текстом

63. Задачи обработки текста: извлечение, поиск, классификация (тематическая, эмоциональная), перевод
64. Разбиение на слова, пунктуация, лексический и морфологический анализ
65. Определение частей речи, имён, основ слов
66. Частотный анализ, представление bag-of-words, TF-IDF и его варианты
67. N-грамы, byte-pair encoding.
68. Векторные представления, семантическая интерпретация алгебраических операций
69. Унитарный код (One-hot encoding).
70. Алгоритмы Word2Vec и FastText.
71. Алгоритм GloVe*.

Тема 1.10 Снижение размерности

72. Постановка задачи, причины и цели снижения размерности.
73. Выбор и извлечение признаков.
74. Подходы к выбору признаков: filtering, wrapping, embedding.
75. Расстояние между распределениями. Расстояние Кульбака-Лейблера. Взаимная информация.
76. Алгоритмы выбора признаков: на основе корреляции (CFS), взаимной информации, Relief.
77. Метод главных компонент (PCA).
78. Нелинейные обобщения метода главных компонент. Kernel PCA.*
79. Неотрицательное матричное разложение (NMF).*
80. Стохастическое вложение соседей с t-распределением (t-SNE).

Тема 1.11 Метод опорных векторов (SVM)

81. Постановка задачи линейного SVM для линейно разделимой выборки
82. Линейный SVM в случае линейно неразделимой выборки.
83. Задача оптимизации с ограничениями. Двойственная задача Лагранжа. Условия Каруша-Куна-Такера

84. Функция Лагранжа для линейного SVM. Опорный вектор. Типы опорных векторов.
85. Kernel trick. Полиномиальное ядро. Радиально-базисное ядро (RBF).
86. SVM для задачи регрессии.