МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования

«ДАГЕСТАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

Факультет математики и компьютерных наук

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ
Машинное обучение

Кафедра дискретной математики и информатики факультета математики и компьютерных наук

Образовательная программа бакалавриата
01.03.05 Статистика

Направленность(профиль) программы:
Анализ больших данных

Форма обучения
очная

Статус дисциплины: входит в часть ОПОП формируемую участниками образовательных отношений

Махачкала, 2023

Рабочая программа дисциплины “Машинное обучение” составлена в 2023 году в соответствии с требованиями ФГОС ВО – бакалавриат по направлению подготовки 01.03.05 – Статистика от 14 августа 2020 г №1032.

Разработчик(и): кафедра дискретной математики и информатики, преподаватель Ибавов Темирлан Ильмутдинович.

Рабочая программа дисциплины одобрена:

на заседании кафедры дискретной математики и информатики от «18» января 2023г, протокол № 5.

Зав. кафедрой _____ Магомедов А.М.

(подпись)

и

на заседании Методической комиссии ФМиКН от

«25» января 2023г., протокол № 4.

Председатель _____ Ризаев М.К.

(подпись)

Рабочая программа дисциплины согласована с учебно-методическим управлением «20» февраля 2023 г.

Начальник УМУ _____ Гасангаджиева А.Г.

(подпись)

Аннотация рабочей программы дисциплины

Дисциплина «Машинное обучение» входит в часть ОПОП формируемую участниками образовательных отношений бакалавриата по направлению подготовки 01.03.05 – Статистика.

Дисциплина реализуется на факультете математики и компьютерных наук кафедрой дискретной математики и информатики.

Содержание дисциплины охватывает круг вопросов, связанных с основными задачами машинного обучения: кластеризации, классификации, регрессии, понижения размерности, прикладными моделями машинного обучения: нейронные сети, обучение с учителем и без учителя, ранжирование, адаптивное прогнозирование.

Дисциплина нацелена на формирование следующих компетенций выпускника: профессиональных – ПК-1.

Преподавание дисциплины предусматривает проведение следующих видов учебных занятий: лекции и практические занятия.

Рабочая программа дисциплины предусматривает проведение следующих видов контроля успеваемости в форме контрольных работ и коллоквиума, и промежуточный контроль в форме зачёта.

Объем дисциплины 3 зачетные единицы, в том числе в академических часах по видам учебных занятий: 16 лекций, 32 лабораторных , 60 – СРС.

Семестр	Учебные занятия						Форма промежуточной аттестации (зачет, дифференцированный зачет, экзамен)
	в том числе:						
	всего	Контактная работа обучающихся с преподавателем				СРС, в том числе зачет, дифференцированный зачет, экзамен	
		всего	из них				
			Лекции	Лабораторные занятия	Практические занятия		
6	108	48	16	32		60	зачёт

1. Цели освоения дисциплины

Целями освоения дисциплины «Машинное обучение» является изучение математических основ и технологий машинного обучения, получение представления об основных задачах машинного обучения, построения и обучения нейронных сетей.

Цели необходимые для достижения поставленной задачи состоят в следующем:

– Выработать навыки представления задач в пространстве состояний и оптимизации поиска решений.

– Приобрести навыки сведения сложных задач к подзадачам с применением графов «И/ИЛИ».

– Изучить модели представления знаний в интеллектуальных системах.

– Получить представление о принципах организации интерфейса на естественном языке к базе знаний интеллектуальной системы.

– Изучить вопросы организации машинных словарей для решения задач компьютерной обработки текстов естественном языке.

2.Место дисциплины в структуре ОПОП бакалавриата

Дисциплина «Машинное обучение» входит в часть ОПОП формируемую участниками образовательных отношений бакалавриата по направлению подготовки 01.03.05 – Статистика.

Изучение дисциплины «Машинное обучение» основывается на знаниях и умениях студентов, полученных в процессе таких дисциплин как «Основы программирования», «Современные технологии программирования», «Дискретная математика», «Основы математической обработки информации», «Введение в алгебру и геометрию».

Знания, полученные в результате изучения дисциплины «Машинное обучение» будут использоваться в последующем освоении дисциплин, в которых используются информационные технологии, а также при решении практических задач профессиональной деятельности.

Знания, навыки и умения, приобретенные в процессе изучения дисциплины в ходе лекций, практических занятий и самостоятельной работы, должны всесторонне использоваться студентами на завершающем этапе обучения, а также в процессе дальнейшей профессиональной деятельности при решении широкого класса прикладных задач.

3. Компетенции обучающегося, формируемые в результате освоения дисциплины (перечень планируемых результатов обучения и процедура освоения).

Код и наименование компетенции из ОПОП	**Код и наименование индикатора достижения компетенций**	**Планируемые результаты обучения**	**Процедура освоения**
ПК-1. Способен собирать, обрабатывать и интерпретировать данные современных научных исследований, необходимые для формирования выводов по соответствующим научным исследованиям	ПК-1.1. Знает методы сбора и обработки данных, полученными в области математических и естественных наук, программирования и информационных технологий для формирования выводов по соответствующим научным исследованиям	Знает: стандартные методы и технические средства для статистических наблюдений. Умеет: применить стандартные методы и технические средства при статистических наблюдениях. Владеет: методами и техническими средствами для статистических наблюдений.	Реферат, письменный опрос, устный опрос
	ПК-1.2. Умеет сбора и обработки данных, полученными в области математических и естественных наук, программирова-ния и информационных технологий для формирования выводов по соответствующим научным исследованиям.	Знает: собирать данные об объекте исследования и выбрать соответствующий инструментарий для обработки информации. Умеет: собирать исходные данные об объекте исследования и выбрать соответствующий инструментарий для обработки информации. Владеет: методами сбора данных об объекте исследования и выбора соответствующий инструментарий для обработки информации.	

	ПК-1.3. Владеет навыками сбора и обработки данных, полученными в области математических и (или) естественных наук, программирования и информационных технологий для формирования выводов по соответствующим научным исследованиям.	Знает: статистические методы обработки информации, в том числе с применением информационно-коммуникационных технологий. Умеет: применять статистические меты для обработки информации, в том числе с применением информационно-коммуникационных технологий. Владеет: статистическими методами обработки информации, в том числе с применением информационно-коммуникационных технологий	

4. Объем, структура и содержание дисциплины.

4.1. Объем дисциплины составляет 3 зачетные единицы, 108 академических часов.

4.2. Структура дисциплины.

4.2.1. Структура дисциплины в очной форме

№ п/п	Разделы и темы дисциплины по модулям	Семестр	Виды учебной работы, включая самостоятельную работу студентов (в часах)				Формы текущего контроля успеваемости и промежуточной аттестации
			Лекции	Практические занятия	Лабораторные занятия	Самостоятельная работа в т.ч. зачет, экзамен	
	Модуль 1. ***Математические основы машинного обучения***						
1	Основные понятия и примеры прикладных задач	1	2		4	6	Устный опрос, Письменный опрос.
2	Линейный классификатор и стохастический градиент	1	2		4	6	
3	Нейронные сети. Градиентные методы	1	2		4	6	

	оптимизации						
Итого по модулю 1:		**36**	**6**		**12**	**18**	
	Модуль 2. Методы классификации						
1	Метрические методы классификации и регрессии		2		4	4	Контрольная работа
2	Метод опорных векторов		2		4	4	
3	Многомерная линейная регрессия				4	4	
4	Нелинейная регрессия					8	
Итого по модулю 2:		**36**	**4**		**12**	**20**	
Модуль 3. Методы отбора признаков							
1	Критерии выбора моделей и методы отбора признаков		2		2	5	
2	Линейные ансамбли				2	6	
3	Байесовская теория классификации		2		2	6	
4	Кластеризация и частичное обучение		2		2	5	
Итого по модулю 3:		**36**	**6**		**8**	**22**	
ИТОГО:		**72**	**16**		**32**	**60**	**Зачёт**

4.3. Содержание дисциплины, структурированное по темам (разделам).

4.3.1. Содержание лекционных занятий по дисциплине.

Модуль 1. Математические основы машинного обучения

Тема 1. Основные понятия и примеры прикладных задач

Постановка задач обучения по прецедентам. Объекты и признаки. Типы шкал: бинарные, номинальные, порядковые, количественные. Типы задач: классификация, регрессия, прогнозирование, ранжирование. Основные понятия: модель алгоритмов, метод обучения, функция потерь и функционал качества, принцип минимизации эмпирического риска, обобщающая способность, скользящий контроль. Линейные модели регрессии и классификации. Метод наименьших квадратов. Полиномиальная регрессия. Примеры прикладных задач. Методика экспериментального исследования и сравнения алгоритмов на модельных и реальных данных. Полигон алгоритмов классификации. CRISP-DM — межотраслевой стандарт ведения проектов интеллектуального анализа данных.

Тема 2. Линейный классификатор и стохастический градиент

Линейный классификатор, модель МакКаллока-Питтса, непрерывные аппроксимации пороговой функции потерь. Метод стохастического градиента SG. Метод стохастического среднего градиента SAG. Эвристики: инициализация весов, порядок предъявления объектов, выбор величины градиентного шага, «выбивание» из локальных минимумов. Проблема мультиколлинеарности и переобучения, регуляризация или редукция весов (weight decay). Вероятностная постановка задачи классификации. Принцип максимума правдоподобия. Вероятностная интерпретация регуляризации, совместное правдоподобие данных и модели. Принцип максимума апостериорной вероятности. Гауссовский и лапласовский регуляризаторы. Логистическая регрессия. Принцип максимума правдоподобия и логарифмическая функция потерь. Метод стохастического градиента для логарифмической функции потерь. Многоклассовая логистическая регрессия. Регуляризованная логистическая регрессия. Калибровка Платта.

Тема 3. Нейронные сети. Градиентные методы оптимизации

Биологический нейрон, модель МакКаллока-Питтса как линейный классификатор. Функции активации. Проблема полноты. Задача исключающего или. Полнота двухслойных сетей в пространстве булевых функций. Алгоритм обратного распространения ошибок. Быстрые методы стохастического градиента: Поляка, Нестерова, AdaGrad, RMSProp, AdaDelta, Adam, Nadam, диагональный метод Левенберга-Марквардта. Проблема взрыва градиента и эвристика gradient clipping. Метод случайных отключений нейронов (Dropout). Интерпретации Dropout. Обратный Dropout и L2-регуляризация. Функции активации ReLU и PReLU. Проблема «паралича» сети. Эвристики для формирования начального приближения. Метод послойной настройки сети. Подбор структуры сети: методы постепенного усложнения сети, оптимальное прореживание нейронных сетей (optimal brain damage).

Модуль 2. Методы классификации

Тема 1. Метрические методы классификации и регрессии

Гипотезы компактности и непрерывности. Обобщённый метрический классификатор. Метод ближайших соседей kNN и его обобщения. Подбор числа k по критерию скользящего контроля. Метод окна Парзена с постоянной и переменной шириной окна. Метод потенциальных функций и его связь с линейной моделью классификации. Задача отбора эталонов. Полный скользящий контроль (CCV), формула быстрого вычисления для метода 1NN. Профиль компактности. Отбор эталонных объектов на основе минимизации функционала полного скользящего контроля. Непараметрическая регрессия. Локально взвешенный метод наименьших квадратов. Ядерное сглаживание. Оценка Надарая-Ватсона с постоянной и переменной шириной окна. Выбор функции ядра и ширины окна сглаживания. Задача отсева выбросов. Робастная непараметрическая регрессия. Алгоритм LOWESS.

Тема 2. Метод опорных векторов

Оптимальная разделяющая гиперплоскость. Понятие зазора между классами (margin). Случаи линейной разделимости и отсутствия линейной разделимости. Связь с минимизацией регуляризованного эмпирического риска. Кусочно-линейная функция

потерь. Задача квадратичного программирования и двойственная задача. Понятие опорных векторов. Рекомендации по выбору константы C. Функция ядра (kernel functions), спрямляющее пространство, теорема Мерсера. Способы конструктивного построения ядер. Примеры ядер. SVM-регрессия. Регуляризации для отбора признаков: LASSO SVM, Elastic Net SVM, SFM, RFM. Метод релевантных векторов RVM

Тема 3. Многомерная линейная регрессия

Задача регрессии, многомерная линейная регрессия. Метод наименьших квадратов, его вероятностный смысл и геометрический смысл. Сингулярное разложение. Проблемы мультиколлинеарности и переобучения. Регуляризация. Гребневая регрессия через сингулярное разложение. Методы отбора признаков: Лассо Тибширани, Elastic Net, сравнение с гребневой регрессией. Метод главных компонент и декоррелирующее преобразование Карунена-Лоэва, его связь с сингулярным разложением. Спектральный подход к решению задачи наименьших квадратов. Задачи и методы низкоранговых матричных разложений.

Тема 4. Нелинейная регрессия

Метод Ньютона-Рафсона, метод Ньютона-Гаусса, Обобщённая аддитивная модель (GAM): метод настройки с возвращениями (backfitting) Хасти-Тибширани., Логистическая регрессия. Метод наименьших квадратов с итеративным пересчётом весов (IRLS). Пример прикладной задачи: кредитный скоринг. Бинаризация признаков. Скоринговые карты и оценивание вероятности дефолта. Риск кредитного портфеля банка, Обобщённая линейная модель (GLM). Экспоненциальное семейство распределений, Неквадратичные функции потерь. Метод наименьших модулей. Квантильная регрессия. Пример прикладной задачи: прогнозирование потребительского спроса, Робастная регрессия, функции потерь с горизонтальными асимптотами.

Модуль 3. Методы отбора признаков

Тема 1. Критерии выбора моделей и методы отбора признаков

Критерии качества классификации: чувствительность и специфичность, ROC-кривая и AUC, точность и полнота, AUC-PR. Внутренние и внешние критерии. Эмпирические и аналитические критерии. Скользящий контроль, разновидности эмпирических оценок скользящего контроля. Критерий непротиворечивости. Разновидности аналитических оценок. Регуляризация. Критерий Акаике (AIC). Байесовский информационный критерий (BIC). Оценка Вапника-Червоненкиса. Сложность задачи отбора признаков. Полный перебор. Метод добавления и удаления, шаговая регрессия. Поиск в глубину, метод ветвей и границ. Усечённый поиск в ширину, многорядный итерационный алгоритм МГУА. Генетический алгоритм, его сходство с МГУА. Случайный поиск и Случайный поиск с адаптацией (СПА).

Тема 2. Линейные ансамбли

Основные понятия: базовый алгоритм, корректирующая операция, Простое голосование (комитет большинства), Стохастические методы: бэггинг и метод случайных подпространств, Случайный лес (Random Forest), Взвешенное голосование. Преобразование простого голосования во взвешенное., Алгоритм AdaBoost. Экспоненциальная аппроксимация пороговой функции потерь. Процесс последовательного обучения базовых алгоритмов. Теорема о сходимости бустинга.

Идентификация нетипичных объектов (выбросов), Теоретические обоснования. Обобщающая способность бустинга, Базовые алгоритмы в бустинге. Решающие пни, Сравнение бэггинга и бустинга, Алгоритм ComBoost. Обобщение на большое число классов.

Тема 3. Байесовская теория классификации

Байесовская теория классификации. Оптимальный байесовский классификатор, Генеративные и дискриминативные модели классификации, Наивный байесовский классификатор. Линейный наивный байесовский классификатор в случае экспоненциального семейства распределений, Мультиномиальный наивный байесовский классификатор для классификации текстов, Метод парзеновского окна. Выбор функции ядра. Выбор ширины окна, переменная ширина окна, Нормальный дискриминантный анализ. Квадратичный дискриминант. Вид разделяющей поверхности. Подстановочный алгоритм, его недостатки и способы их устранения. Линейный дискриминант Фишера, Связь линейного дискриминанта Фишера с методом наименьших квадратов, Смесь многомерных нормальных распределений. Сеть радиальных базисных функций (RBF) и применение EM-алгоритма для её настройки. Сравнение RBF-сети и SVM с гауссовским ядром.

Тема 4. Кластеризация и частичное обучение

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, Алгоритм DBSCAN, Агломеративная кластеризация, Алгоритм Ланса-Вильямса и его частные случаи, Алгоритм построения дендрограммы. Определение числа кластеров, Свойства сжатия/растяжения и монотонности, 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.

4.3.2. Содержание лабораторных занятий по дисциплине.

Модуль 1. Математические основы машинного обучения

Тема 1. Основные понятия и примеры прикладных задач

- ***Лабораторное занятие №1:***
  - При помощи sklearn показать пример Ирисов Фишера.
  - Понятие модели алгоритмов, алгоритм обучения, процесс оптимизации для конкретной задачи.
  - Переход от бинарной к многоклассовой.
  - Переобучение. Борьба с переобучениям.

- Немного о типах задач машинного обучения: прикладные и исследовательские.

Тема 2. Линейный классификатор и стохастический градиент

- ***Лабораторное занятие №2:***
  - Разбор самостоятельного задания №1
  - Метод стохастического градиента на практике.
  - Использования torch framework для нахождения градиента сложной функции.
  - Вероятностная постановка задачи машинного обучения. Регуляризация l1, l2.
  - Анализ решения задачи оптимизации от параметра регуляризации.
  - Выбор параметра регуляризации при помощи LOO.

Тема 3. Нейронные сети. Градиентные методы оптимизации

- ***Лабораторное занятие №3:***
  - Разбор самостоятельного задания №2.
  - Построение простой нейросетевой модели: многослойный персептрон.
  - Обучение персептрона на выборке MNIST.
  - Подбор гиперпараметров модели.
  - Прореживание сетей (без кода, только графики).

Модуль 2. Методы классификации

Тема 1. Метрические методы классификации и регрессии

- ***Лабораторное занятие №1:***
  - Разбор домашнего задания.
  - Пример как можно отказаться от признаков в линейном классификаторе.
  - Метод ближайших соседей, анализ разного количества соседей.
  - Ядра в методе ближайших соседей.
  - Метод Парзеновского окна.
  - Метод потенциальных функций (реализация).
  - Отбор эталонных элементов, алгоритм STOLP.
  - Формула Надарая Ватсона.

Тема 2. Метод опорных векторов

Лабораторное занятие№2:

- SVM для классификации.
- Примеры использования ядер для SVM.
- SVM для регрессии.
- Генерация признаков на основе опорных элементов.

Тема 3. Многомерная линейная регрессия

Лабораторное занятие №3:

- Многомерная линейная регрессия.
- Сингулярное разложение.
- Регуляризация для многомерной регрессии: используя SVD.
- Зависимость качества аппроксимации от числа обусловленности.
- Метод главных компонент: визуализация MNIST.
- Метод главных компонент: для изображений.

Модуль 3. Методы отбора признаков

Тема 1. Критерии выбора моделей и методы отбора признаков

- ***Лабораторное занятие №1:***
  - Оценка качества модели: внешний и внутрений критерии.
  - Отбор признаков: полный перебор, алгоритм Add, алгоритм Add-Del.
  - Качество классификации: Precision, Recall.
  - Пример задачи information retrieval.
  - О составлении выборки для постановки задачи ML.

Тема 2. Линейные ансамбли

- **Лабораторное занятие №2**
  - Логический классификатор реализации.
  - Примеры задач для решения логичеким классификатором.
  - Критерии информативности.
  - Решающий список, простая реализация.
  - Решающее дерево.
  - Случайный лес.

Тема 3. Байесовская теория классификации

- ***Лабораторное занятие №3***
  - Постановка задачи ассоциативных правил.
  - Синтетический пример.
  - Пример расчета данных из kaggle.
  - Алгоритм APriory.
  - Алгоритм FP-роста.
  - Обобщение для открытых данных.
  - Обобщенные ассоциативные правила.

Тема 4. Кластеризация и частичное обучение

- **Лабораторное занятие №4**
  - Задача кластеризации.
  - Примеры кластеров.
  - K-значит.

- ДБСКАН.
- Иерархическая кластеризация.
- Частичное обучение.
- Самообучение, 1970.
- Неразмеченные данные в глубоком обучении.

5. **Образовательные технологии**

Образовательные технологии, используемые при реализации различных видов учебной деятельности:

– лекции: лекция с компьютерными презентациями, интерактивные проблемные лекции; – лабораторная работа: метод обучения, при котором студенты под руководством преподавателя по заранее намеченному плану выполняют определенные практические задания и в процессе их усваивают новый учебный материал;

– деловая игра: метод имитации принятия решений в различных производственных ситуациях при наличии информационной неопределённости. В ходе обучения применяются следующие формы учебного процесса: лекции и лабораторные занятия, участие в деловой игре, самостоятельная внеаудиторная работа.

В качестве метода проверки знаний используется написание рефератов, устный опрос студентов, решение кейс-задач.

6. Учебно-методическое обеспечение самостоятельной работы студентов.

Самостоятельная работа студентов складывается из

- проработки материала практических занятий (настоятельно рекомендуется самостоятельное практическое решение всех разобранных на занятиях упражнений);

- изучения рекомендованной литературы и материалов соответствующих форумов интернет; - подготовки к сдаче текущих и промежуточных форм контроля (практических работ, урока и реферата).

Пакет заданий для самостоятельной работы выдается по истечению месяца с начала семестра, определяются предельные сроки их выполнения и сдачи.

№	Вид самостоятельной работы	Вид контроля	Учебно-методическое обеспечение
1	Проработка теоретического материала	Контрольный фронтальный опрос	См. разделы 7.3, 8, 9 данного документа
2	Изучение рекомендованной литературы	Контрольный фронтальный опрос, приём и представление рефератов	См. разделы 7.3, 8, 9 данного документа

3	Подготовка к отчётам по лабораторным работам	Проверка выполнения работ, опрос по теме	См. разделы 7.3, 8, 9 данного документа
4	Подготовка к сдаче промежуточных форм контроля	Контрольные работы по каждому модулю	См. разделы 7.3, 8, 9 данного документа

Текущий контроль:

1. Проверка программ на языке высокого уровня по заданиям;
2. Проверка выполнения домашних заданий;
3. Промежуточная аттестация в форме письменной работы.

Текущий контроль включает, кроме еженедельного опроса и проверки знаний по текущему материалу, ведение электронного журнала посещаемости, проверку выполнения компьютерных программ. Подразумевается непрерывное общение по электронной почте (общение по скайпу не целесообразно, т.к. не позволяет осуществлять доскональную проверку заданий).

Промежуточный контроль проводится в виде письменной работы, рассчитанной на 20- 30 минут.

Итоговый контроль проводится в виде письменной работы с обязательным устным собеседованием по результатам предварительной проверки.

Критерии выставления оценок «отлично», «хорошо», «удовлетворительно» определяются степенью владения материалом и достигнутым уровнем компетентности в решении задач дискретной математики. В исключительных случаях учитываются успехи на всероссийских олимпиадах и конкурсах по номинации данной дисциплины.

Задания для самостоятельной работы:

- В задаче по переходу от бинарной классификации к многоклассовой добавить константу и скорректировать соответстветсвующие разделяющие гиперплоскости.
- Подсказка: в LogisticRegresion нужно добавить специальный параметр fit_intercept=False, чтобы внутри черного ящика своя константта не добавлялась(влият на результат).
- Используя вероятностную постановку задачи для линейной регрессии с априорным предположением p(w) = N(0, I) получить аналитическое решение на оптимальный вектор параметров w.
- Использовать метод Cross-Validation вместо метода LOO для выбора оптимального параметра регуляризации gamma.
- Проделать то, что было на семинаре для выборки FashionMnist: подбор гиперпараметров модели (выполнить более подробно чем на семинаре), также провести анализ полученных результатов.
- Указать какие минусы вы увидели в подборе гиперпараметров на семинаре (их как минимум 3).

- Выбрать один из метрических классификаторов (классификации или регрессии) и выполнить поиск оптимальных гиперпараметра при помощи кросс валидации.
- Провести эксперимент с полиномиальным ядром: сгенерировать синтетическую выборку, на которой полиномиальное ядро имеет лучшее качество аппроксимации чем rbf и линейное ядро.
- Для синтетически сгенерированной выборки (beta=2, mu=0.01) построить график зависимости качества аппроксимации контрольной вбыорки от коэффициента регуляризации. Сравнить скорость работы в случае использования SVD разложения и без него.
- Использовать модель для векторизации предложений из семинара. На основе полученных векторов решить задачу сентимент анализа для выборки Twitter (задача бинарной классификации). В качестве модели рассмотреть логистическую регрессию. Рекомендуется использовать модель Perceptron с третьего семинара, а также функцию ошибки torch.nn.BCELoss. Ссылка на данные: https://drive.google.com/file/d/1k4JrnVcoePEENCYt5iy17dyV_h133j2X/view?usp=sharing (предложения для классификации это последний столбец, а целевая переменная это второй столбец).
- реализовать метод отбора признаков Add-Del.
- предложенить внешний критерий качества для задачи поиска ошибок в текстах.

7. Фонд оценочных средств для проведения текущего контроля успеваемости, промежуточной аттестации по итогам освоения дисциплины.

7.1. Типовые контрольные задания

Вопросы к зачету:

Байесовская классификация

- Записать общую формулу байесовского классификатора (надо помнить формулу).
- Какие вы знаете три подхода к восстановлению плотности распределения по выборке?
- Что такое наивный байесовский классификатор?
- Что такое оценка плотности Парзена-Розенблатта (надо помнить формулу). Выписать формулу алгоритма классификации в методе парзеновского окна.
- На что влияет ширина окна, а на что вид ядра в методе парзеновского окна?
- Многомерное нормальное распределение (надо помнить формулу). Вывести формулу квадратичного дискриминанта. При каком условии он становится линейным?
- На каких предположениях осован линейный дискриминант Фишера?
- Что такое «проблема мультиколлинеарности», в каких задачах и при использовании каких алгоритмов она возникает? Какие есть подходы к её решению?
- Что такое «смесь распределений» (надо помнить формулу)?
- Что такое EM-алгоритм, какова его основная идея? Какая задача решается на E-шаге, на M-шаге? Каков вероятностный смысл скрытых переменных?
- Последовательное добавление компонент в EM-алгоритме, основная идея алгоритма.
- Что такое стохастический EM-алгоритм, какова основная идея? В чём его преимущество (какой недостаток стандартного EM-алгоритма он устраняет)?
- Что такое сеть радиальных базисных функций?

- Что такое «выбросы»? Как осуществляется фильтрация выбросов?

Метрическая классификация

- Что такое обобщённый алгоритм классификации (надо помнить формулу)? Какие вы знаете частные случаи?
- Как определяется понятие отступа в метрических алгоритмах классификации?
- Что такое окно переменной ширины, в каких случаях его стоит использовать?
- Что такое метод потенциальных функций? Идея алгоритма настройки. Сравните с методом радиальных базисных функций.
- Зачем нужен отбор опорных объектов в метрических алгоритмах классификации?
- Основная идея алгоритма СТОЛП.
- Что такое функция конкурентного сходства? Основная идея алгоритма FRiS-СТОЛП.
- Приведите пример метрического алгоритма классификации, который одновременно является байесовским классификатором.
- Приведите пример метрического алгоритма классификации, который одновременно является линейным классификатором.

Линейная классификация

- Что такое модель МакКаллока-Питтса (надо помнить формулу)?
- Метод стохастического градиента. Расписать градиентный шаг для квадратичной функции потерь и сигмоидной функции активации.
- Недостатки метода SG и как с ними бороться?
- Что такое линейный адаптивный элемент ADALINE?
- Что такое правило Хэбба?
- Что такое «сокращение весов»?
- Обоснование логистической регрессии (основная теорема), основные посылки (3) и следствия (2). Как выражается апостериорная вероятность классов (надо помнить формулу).
- Как выражается функция потерь в логистической регрессии (надо помнить формулу).
- Две мотивации и постановка задачи метода опорных векторов. Уметь вывести постановку задачи SVM (рекомендуется помнить формулу постановки задачи).
- Какая функция потерь используется в SVM? В логистической регрессии? Какие ещё функции потерь Вы знаете?
- Что такое ядро в SVM? Зачем вводятся ядра? Любая ли функция может быть ядром?
- Какое ядро порождает полимиальные разделяющие поверхности?
- Что такое ROC-кривая, как она определяется? Как она эффективно вычисляется?
- В каких алгоритмах классификации можно узнать не только классовую принадлежность классифицируемого объекта, но и вероятность того, что данный объект принадлежит каждому из классов?
- Каков вероятностный смысл регуляризации? Какие типы регуляризаторов Вы знаете?

- Что такое принцип максимума совместного правдоподобия данных и модели (надо помнить формулу)?

Регрессия

- Что такое ядерное сглаживание?
- Что есть общего между ядром в непараметрической регрессии и ядром SVM?
- На что влияет ширина окна, а на что вид ядра в непараметрической регрессии?
- Что такое окна переменной ширины, и зачем они нужны?
- Что такое «выбросы»? Как осуществляется фильтрация выбросов в непараметрической регрессии?
- Постановка задачи многомерной линейной регрессии. Матричная запись.
- Что такое сингулярное разложение? Как оно используется для решения задачи наименьших квадратов?
- Что такое «проблема мультиколлинеарности» в задачах многомерной линейной регрессии? Какие есть три подхода к её устранению?
- Сравнить гребневую регрессию и лассо. В каких задачах предпочтительнее использовать лассо?
- Какую проблему решает метод главных компонент в многомерной линейной регрессии? Записать матричную постановку задачи для метода главных компонент.
- Как свести задачу многомерной нелинейной регрессии к последовательности линейных задач?
- Метод настройки с возвращениями (backfitting): постановка задачи и основная идея метода.
- Какие методы построения логистической регрессии Вы знаете?
- Приведите примеры неквадратичных функций потерь в регрессионных задачах. С какой целью они вводятся?

Выбор модели и отбор признаков

- В чём отличия внутренних и внешних критериев?
- Разновидности внешних критериев.
- Разновидности критерия скользящего контроля.
- Что такое критерий непротиворечивости? В чём его недостатки?
- Что такое многоступенчатый выбор модели по совокупности критериев?
- Основная идея отбора признаков методом полного перебора. Действительно ли это полный перебор?
- Основная идея отбора признаков методом добавлений и исключнений.
- Что такое шаговая регрессия? Можно ли её использовать для классификации, в каком методе?
- Основная идея отбора признаков методом поиска в глубину.
- Основная идея отбора признаков методом поиска в ширину.
- Что такое МГУА?

- Основная идея отбора признаков с помощью генетического алгоритма.
- Основная идея отбора признаков с помощью случайного поиска.
- В чём отличия случайного поиска от случайного поиска с адаптацией?

Нейронные сети

- Приведите пример выборки, которую невозможно классифицировать без ошибок с помощью линейного алгоритма классификации. Какова минимальная длина выборки, обладающая данным свойством? Какие существуют способы модифицировать линейный алгоритм так, чтобы данная выборка стала линейно разделимой?
- Почему любая булева функция представима в виде нейронной сети? Сколько в ней слоёв?
- Метод обратного распространения ошибок. Основная идея. Основные недостатки и способы их устранения.
- Как можно выбирать начальное приближение в градиентных методах настройки нейронных сетей?
- Как можно ускорить сходимость в градиентных методах настройки нейронных сетей?
- Что такое диагональный метод Левенберга-Марквардта?
- Что такое «паралич» сети, и как его избежать?
- Как выбирать число слоёв в градиентных методах настройки нейронных сетей?
- Как выбирать число нейронов скрытого слоя в градиентных методах настройки нейронных сетей?
- В чём заключается метод оптимального прореживания нейронной сети? Какие недостатки стандартного алгоритма обратного распространения ошибок позволяет устранить метод ODB?

Примеры задач

- Задана цена отказа от классификации. Выписать модифицированную формулу байесовского классификатора.
- Вывести формулу линейного дискриминанта для случая независимых признаков.
- Вывести формулу наивного байесовского классификатора для случая бинарных признаков (доказать, что он линеен).
- Вывести формулу градиентного шага в методе логистической регрессии для задачи классификации с двумя классами. Сравнить с правилом Хэбба.
- Вывести формулу непараметрической регрессии Надарая-Ватсона.
- Вывести формулу регуляризованного решения задачи многомерной линейной регрессии через сингулярное разложение.
- Вывести градиентный метод обучения в логистической регрессии.

Список тем рефератов по дисциплине «Системы искусственного интеллекта»

1. Тема: Системы автоматизации проектных работ (САПР).
2. Тема: Экспертные системы, их применение для решения задач различных предметных областей.
3. Тема: Системы искусственного интеллекта, классификация, особенности.
4. Тема: Роль автоматизированных систем поддержки принятия решений в управлении экономическими объектами.
5. Тема: Области применения нейронных сетей, классы задач, решаемых благодаря их использованию.
6. Тема: Формализация и структурирование знаний при проектировании баз знаний. Модели знаний.
7. Тема: Автоматизированные информационные технологии и системы для интеллектуальной поддержки финансового управления и проведения финансового анализа состояния предприятия.
8. Тема: Назначение и области применения правовых информационно – поисковых справочных систем.
9. Тема: Электронные программы – словари.
10. Тема: Программы перевода текстов с одних языков на другие.
11. Тема: Инструментальные средства и языки программирования, применяемые для разработки систем искусственного интеллекта.
12. Тема: Общая характеристика классов задач, решаемых с помощью систем искусственного интеллекта.
13. Тема: Общая характеристика и основные компоненты автоматизированных систем поддержки принятия решений модельного типа.
14. Тема: Гипертекстовые поисковые Internet – системы.
15. Тема: Интеллектуальные обучающие программы по дисциплинам средней и высшей школы, специальным курсам.
16. Тема: Основные понятия теории предикатов, её использование для представления знаний.
17. Тема: Нечёткие множества, операции над ними. Использование нечётких выводов в экспертных системах.
18. Тема: Определение и методы построения когнитивных карт. Принятие решений с помощью когнитивных карт.
19. Тема: Применение автоматизированных систем поддержки принятия решений модельного типа в управлении предприятиями.
20. Тема: Применение систем искусственного интеллекта для статистического анализа данных и прогнозирования поведения объектов и систем.
21. Тема: OLAP – технологии.
22. Тема: Информационные хранилища: принципы построения, основные компоненты.
23. Тема: CASE – технологии: назначение, примеры.
24. Тема: Классификация систем искусственного интеллекта.
25. Тема: Контекстные системы поиска: назначение, примеры.

7.2. Методические материалы, определяющие процедуру оценивания знаний, умений, навыков и (или) опыта деятельности, характеризующих этапы формирования компетенций.

1. Общий результат по модулю выводится как интегральная оценка, складывающая из текущей работы - 50% и текущего контроля - 50%.
Текущий работа по дисциплине включает:
- посещение занятий - 30 баллов,
- участие на практических занятиях - 70 баллов,

Текущий контроль по дисциплине включает:
- устный опрос - 70 баллов,
- письменная контрольная работа - 20 баллов
- реферат – 10 баллов.

8. Учебно-методическое обеспечение дисциплины.

а) адрес сайта курса

GATE - https://gate.ac.uk/
Protégé - http://protege.stanford.edu/ UIMA - https://uima.apache.org/
Wikipedia
http://ru.wikipedia.org/wiki/%D0%98%D1%81%D0%BA%D1%83%D1%81%D1%81%D1%82%D0%B2%DPPоссийская ассоциация искусственного интеллекта (РАИИ) - http://www.raai.org/

б) основная литература:

1. *Hastie T., Tibshirani R., Friedman J.* The Elements of Statistical Learning. Springer, 2014. — 739 p.
2. *Bishop C. M.* Pattern Recognition and Machine Learning. — Springer, 2006. — 738 p.
3. *Мерков А. Б.* Распознавание образов. Введение в методы статистического обучения. 2011. 256 с.

б) дополнительная литература:

1. *Мерков А. Б.* Распознавание образов. Построение и обучение вероятностных моделей. 2014. 238 с.
2. *Коэльо Л.П., Ричарт В.* Построение систем машинного обучения на языке Python. 2016. 302 с.

9. Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины.

1. Редактор онтологий Protégé – свободно распространяемый инструментарий для построения и использования онтологий.

2. IHMC CmapTool – свободно распространяемый инструментарий для построения концептуальных карт знаний (http://cmap.ihmc.us/).

3. Загорулько Ю.А. Электронный учебник «Инженерия знаний». НГУ, 2011. – url: http://193.124.209.204/default.aspx?db=book_zagorulko&int=VIEW&el=1684&templ= I206

10. Методические указания для обучающихся по освоению дисциплины.

Курс предусматривает занятия в компьютерном классе, подключенном к Интернету с установленным специализированным программных обеспечением. Предусмотрены

лекции, практические занятия в виде выполнения практических заданий по работе со знаниями.

Для эффективного изучения практической части дисциплины настоятельно рекомендуется:

– систематически выполнять подготовку к практическим занятиям по предложенным преподавателем темам;

– своевременно выполнять и защищать практические задания.

Самостоятельная работа студента - один из важнейших этапов в подготовке специалистов. Она приобщает студентов к исследовательской работе, обогащает опытом и знаниями, необходимыми для дальнейшего их становления как специалистов, прививает навыки работы с литературой.

Цель самостоятельной работы - систематизация, закрепление и расширение теоретических и практических знаний с использованием современных информационных технологий и литературных источников. Для развития навыков самостоятельной работы студентами во время самостоятельной работы выполняются: – эссе по проблемам современных тенденций развития информационных технологий управления; – домашние задания по поиску в Интернете информации на заданную научную тему и подготовке доклада. Эссе или доклад готовится студентом самостоятельно, в нём обобщаются теоретические материалы по исследуемой теме с использованием материалов из общетехнической и специальной литературы, нормативно-правовых документов, стандартизирующих рассматриваемую сферу. В содержании доклада должен быть собственный анализ и критический подход к решению проблемы по выбранной теме исследования. Материалы должны быть изложены на высоком теоретическом уровне, с применением практических данных, примеров. Студентам рекомендуется непрерывно проводить научные исследования под руководством преподавателя кафедры по избранной теме и готовить сообщения на научные конференции, статьи в Сборник молодых исследователей и научные журналы.

Обучение студентов с ограниченными возможностями организуется в соответствии с требованиями «Методических рекомендаций по организации образовательного процесса для обучения инвалидов и лиц с ограниченными возможностями здоровья в образовательных организациях высшего профессионального образования» от «8» апреля 2014 г. В освоении дисциплины инвалидами и лицами с ограниченными возможностями здоровья большое значение имеет индивидуальная учебная работа (консультации) – дополнительное разъяснение учебного материала.

Индивидуальные консультации по предмету являются важным фактором, способствующим индивидуализации обучения и установлению воспитательного контакта между преподавателем и обучающимся инвалидом или лицом с ограниченными возможностями здоровья.

11. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине, включая перечень программного обеспечения и информационных справочных систем.

Перечень необходимого программного обеспечения

− Microsoft Word 2010 или более поздний;

− Программный продукт Microsoft Visio;

− Средство чтения PDF-файлов Adobe Acrobat или аналог.

− Среда разработки PyCharm/Intelij Idea

12. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине.

Мультимедийная аудитория состоит из интегрированных инженерных систем с единой системой управления, оснащенная современными средствами воспроизведения и визуализации любой видео и аудио информации, получения и передачи электронных документов. Типовая комплектация мультимедийной аудитории состоит из: мультимедийного проектора, автоматизированного проекционного экрана, акустической системы, а также интерактивной трибуны преподавателя, персональный компьютер (с техническими характеристиками не ниже Intel Core i3-2100, DDR3 4096Mb, 500Gb), блок управления оборудованием, интерфейсы подключения: USB,audio, HDMI. Интерактивная трибуна преподавателя является ключевым элементом управления, объединяющим все устройства в единую систему, и служит полноценным рабочим местом преподавателя. Преподаватель имеет возможность легко управлять всей системой, не отходя от трибуны, что позволяет проводить лекции, практические занятия, презентации, вебинары, конференции и другие виды аудиторной нагрузки обучающихся в удобной и доступной для них форме с применением современных интерактивных средств обучения, в том числе с использованием в процессе обучения всех корпоративных ресурсов. Мультимедийная аудитория также оснащена широкополосным доступом в сеть интернет. Компьютерное оборудованием имеет соответствующее лицензионное программное обеспечение. Компьютерный класс, представляющий собой рабочее место преподавателя и не менее 15 рабочих мест студентов, включающих компьютерный стол, стул, персональный компьютер, лицензионное программное обеспечение. Каждый компьютер имеет широкополосный доступ в сеть Интернет. Учебно-методическая литература для данной дисциплины имеется в наличии в электронно-библиотечной системе.