

**Федеральное государственное автономное образовательное
учреждение высшего образования
«Московский физико-технический институт
(национальный исследовательский университет)»**

УТВЕРЖДЕНО
Директор физтех-школы
аэрокосмических технологий
С.С. Негодяев

	Рабочая программа дисциплины (модуля)
по дисциплине:	Введение в машинное обучение
по направлению:	Техническая физика
профиль подготовки:	Техническая физика космических летательных аппаратов Физтех-школа Аэрокосмических Технологий центр образовательных программ ФАКТ
курс:	4
квалификация:	бакалавр

Семестр, формы промежуточной аттестации: 7 (осенний) - Экзамен

Аудиторных часов: 30 всего, в том числе:

лекции: 0 час.

семинары: 0 час.

лабораторные занятия: 30 час.

Самостоятельная работа: 30 час.

Подготовка к экзамену: 30 час.

Всего часов: 90, всего зач. ед.: 2

Программу составили:

В.В. Стрижов, д-р физ.-мат. наук

В.Ю. Семака, преподаватель

Программа обсуждена на заседании центра образовательных программ ФАКТ 02.12.2024

Аннотация

Дисциплина "Введение в машинное обучение" отвечает за формирование базовых знаний по основным задачам обучения по прецедентам: классификация, кластеризация, регрессия, понижение размерности; теории вычислительного обучения (computational learning theory, COLT), исследующей проблему надёжности восстановления зависимостей по эмпирическим данным.

1. Цели и задачи

Цель дисциплины

- рассмотрение основных задач обучения по прецедентам: классификация, кластеризация, регрессия, понижение размерности;
- изучение теории вычислительного обучения (computational learning theory, COLT), исследующей проблему надёжности восстановления зависимостей по эмпирическим данным.

Задачи дисциплины

- приобретение теоретических знаний в области обучения по прецедентам;
- изучение методов их решения, как классических, так и новых, созданных за последние 10–15 лет;
- освоение и глубокое понимание математических основ, взаимосвязей, достоинств и ограничений рассматриваемых методов;
- научить студентов оценивать надёжность алгоритмов обучения;
- использовать оценки обобщающей способности для разработки более надёжных алгоритмов;
- применять их для решения прикладных задач классификации, регрессии, прогнозирования.

2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

Код и наименование компетенции	Индикаторы достижения компетенции
ОПК-2 Способен применять методы математического анализа, математического моделирования и оптимизации для решения задач, возникающих в ходе профессиональной деятельности	ОПК-2.1 Знаком с основными методами математического анализа, математического моделирования и оптимизации
	ОПК-2.2 Способен строить математические модели, производить количественные расчеты и оценки
	ОПК-2.3 Способен определять границы применимости полученных результатов
ОПК-6 Способен работать с распределенными базами данных, с информацией в глобальных компьютерных сетях, применяя современные образовательные и информационные технологии	ОПК-6.2 Использует современные образовательные и информационные технологии и сервисы сети Интернет при решении задач в области профессиональной деятельности
ПК-1 Способен применять эффективные методы исследования физико-технических объектов и процессов, проводить испытания технологических процессов и (или) изделий с использованием современных аналитических средств технической физики	ПК-1.1 Способен использовать исследовательские методы при решении новых задач, применяя знания в различных областях науки (техники)
	ПК-1.2 Владеет аналитическими, вычислительными и экспериментальными методами исследования
	ПК-1.3 Способен разрабатывать и применять наиболее подходящие теоретические и экспериментальные методы исследований к конкретной научной задаче и интерпретировать полученные результаты

ПК-2 Готов изучать научно-техническую информацию, отечественный и зарубежный опыт по тематике профессиональной деятельности	ПК-2.1 Владеет методами научного поиска и интеллектуального анализа информации при решении задач профессиональной деятельности
---	--

3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны

знать:

- фундаментальные понятия задач обучения по прецедентам;
- основные методы и алгоритмы решения задач обучения по прецедентам;
- основные области применения этих методов и алгоритмов.

уметь:

- применять методы и алгоритмы к решению задач обучения по прецедентам.

владеть:

- навыками самостоятельной работы при решении типовых задач;
- культурой постановки и моделирования практически значимых задач;
- практикой исследования и решения теоретических и прикладных задач;
- навыками теоретического анализа реальных задач, решаемых с помощью алгоритмов обучения по прецедентам.

4. Содержание дисциплины (модуля), структурированное по темам (разделам) с указанием отведенного на них количества академических часов и видов учебных занятий

4.1. Разделы дисциплины (модуля) и трудоемкости по видам учебных занятий

№	Тема (раздел) дисциплины	Трудоемкость по видам учебных занятий, включая самостоятельную работу, час.			
		Лекции	Семинары	Лаборат. работы	Самост. работа
1	Байесовские методы классификации.			2	2
2	Градиентные линейные методы классификации.			2	2
3	Логистическая регрессия.			2	2
4	Логические методы классификации и решающие деревья.			2	3
5	Метод опорных векторов.			2	3
6	Метрические методы классификации.			2	3
7	Многомерная линейная регрессия.			4	3
8	Нелинейная и непараметрическая регрессия, нестандартные функции потерь.			4	3
9	Основные понятия и примеры прикладных задач.			2	3
10	Генеративные состязательные сети (GAN).			4	3
11	Прогнозирование временных рядов.			4	3
Итого часов				30	30
Подготовка к экзамену		30 час.			
Общая трудоёмкость		90 час., 2 зач.ед.			

4.2. Содержание дисциплины (модуля), структурированное по темам (разделам)

1. Байесовские методы классификации.

Оптимальный байесовский классификатор.

Принцип максимума апостериорной вероятности. Функционал среднего риска. Ошибки I и II рода. Теорема об оптимальности байесовского классификатора. Оценивание плотности распределения: три основных подхода. Наивный байесовский классификатор.

Непараметрическое оценивание плотности.

Ядерная оценка плотности Парзена-Розенблатта. Одномерный и многомерный случаи. Метод парзеновского окна. Выбор функции ядра. Выбор ширины окна, переменная ширина окна. Робастное оценивание плотности. Непараметрический наивный байесовский классификатор.

Параметрическое оценивание плотности.

Нормальный дискриминантный анализ. Многомерное нормальное распределение, геометрическая интерпретация. Выборочные оценки параметров многомерного нормального распределения. Матричное дифференцирование. Вывод оценок параметров многомерного нормального распределения. Квадратичный дискриминант. Вид разделяющей поверхности. Подстановочный алгоритм, его недостатки и способы их устранения. Линейный дискриминант Фишера. Связь с методом наименьших квадратов. Проблемы мультиколлинеарности и переобучения. Регуляризация ковариационной матрицы. Робастное оценивание. Цензурирование выборки (отсев объектов-выбросов). Параметрический наивный байесовский классификатор. Жадное добавление признаков в линейном дискриминанте, метод редукции размерности Шурыгина.

Разделение смеси распределений.

Смесь распределений. EM-алгоритм: основная идея, понятие скрытых переменных. Вывод алгоритма без обоснования сходимости. Псевдокод EM-алгоритма. Критерий останова. Выбор начального приближения. Выбор числа компонентов смеси. Стохастический EM-алгоритм. Смесь многомерных нормальных распределений. Сеть радиальных базисных функций (RBF) и применение EM-алгоритма для её настройки. Сопоставление RBF-сети и SVM с гауссовским ядром.

2. Градиентные линейные методы классификации.

Линейный классификатор, непрерывные аппроксимации пороговой функции потерь. Связь с методом максимума правдоподобия. Метод стохастического градиента и частные случаи: адаптивный линейный элемент ADALINE, персептрон Розенблатта, правило Хэбба. Теорема Новикова о сходимости. Доказательство теоремы Новикова. Эвристики: инициализация весов, порядок предъявления объектов, выбор величины градиентного шага, "выбивание" из локальных минимумов. Метод стохастического среднего градиента SAG. Проблема мультиколлинеарности и переобучения, редукция весов (weight decay). Байесовская регуляризация. Принцип максимума совместного правдоподобия данных и модели. Квадратичный (гауссовский) и лапласовский регуляризаторы. Настройка порога решающего правила по критерию числа ошибок I и II рода. Кривая ошибок (ROC curve). Алгоритм эффективного построения ROC-кривой. Градиентный метод максимизации AUC.

3. Логистическая регрессия.

Гипотеза экспоненциальности функций правдоподобия классов. Теорема о линейности байесовского оптимального классификатора. Оценивание апостериорных вероятностей классов с помощью сигмоидной функции активации. Логистическая регрессия. Принцип максимума правдоподобия и логарифмическая функция потерь. Метод стохастического градиента для логарифмической функции потерь. Сглаженное правило Хэбба. 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.

4. Логические методы классификации и решающие деревья.

Понятия закономерности и информативности.

Понятие логической закономерности. Эвристическое, статистическое, энтропийное определение информативности. Асимптотическая эквивалентность статистического и энтропийного определения. Сравнение областей эвристических и статистических закономерностей. Разновидности закономерностей: конъюнкции пороговых предикатов (гиперпараллелепипеды), синдромные правила, шары, гиперплоскости. Градиентный алгоритм синтеза конъюнкций, частные случаи: жадный алгоритм, стохастический локальный поиск, стабилизация, редукция. Бинаризация признаков. Алгоритм разбиения области значений признака на информативные зоны.

Решающие списки и деревья.

Решающий список. Жадный алгоритм синтеза списка. Решающее дерево. Псевдокод: жадный алгоритм ID3. Недостатки алгоритма и способы их устранения. Проблема переобучения. Редукция решающих деревьев: предредукция и постредукция. Преобразование решающего дерева в решающий список. Алгоритм LISTBB. Небрежные решающие деревья (oblivious decision trees).

5. 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. Обучение SVM методом активных ограничений. Алгоритм INCAS. Алгоритм SMO. Нью-SVM. SVM-регрессия. Метод релевантных векторов RVM. Регуляризации для отбора признаков: LASSO SVM, Elastic Net SVM, SFM, RFM.

6. Метрические методы классификации.

Метод ближайших соседей и его обобщения.

Метод ближайших соседей (kNN) и его обобщения. Подбор числа k по критерию скользящего контроля. Обобщённый метрический классификатор, понятие отступа. Метод потенциальных функций, градиентный алгоритм.

Отбор эталонов и оптимизация метрики.

Отбор эталонных объектов. Псевдокод: алгоритм СТОЛП. Функция конкурентного сходства, алгоритм FRiS-СТОЛП. Функционал полного скользящего контроля, формула быстрого вычисления для метода 1NN. Профиль компактности. Функция вклада объекта. Отбор эталонных объектов на основе минимизации функционала полного скользящего контроля. Эффективные структуры данных для быстрого поиска ближайших объектов в прямых и обратных окрестностях - метрические деревья. Проклятие размерности. Задача настройки весов признаков. Концепция вывода на основе прецедентов (CBR).

7. 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. Лассо Тибширани, сравнение с гребневой регрессией. Метод главных компонент и декоррелирующее преобразование Карунена-Лоэва, его связь с сингулярным разложением.

8. Нелинейная и непараметрическая регрессия, нестандартные функции потерь.

Нелинейная параметрическая регрессия.

Метод Ньютона-Рафсона, метод Ньютона-Гаусса. Обобщённая линейная модель (GLM). Одномерные нелинейные преобразования признаков: метод настройки с возвращениями (backfitting) Хасты-Тибширани.

Непараметрическая регрессия.

Сглаживание. Локально взвешенный метод наименьших квадратов и оценка Надарая-Ватсона. Выбор функции ядра. Выбор ширины окна сглаживания. Сглаживание с переменной шириной окна. Проблема выбросов и робастная непараметрическая регрессия. Алгоритм LOWESS. Доверительный интервал значения регрессии в точке. Проблемы "проклятия размерности" и выбора метрики.

Неквадратичные функции потерь.

Метод наименьших модулей. Квантильная регрессия. Пример прикладной задачи: прогнозирование потребительского спроса. Робастная регрессия, функция Мешалкина. SVM-регрессия.

9. 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, кластеризация. Примеры прикладных задач. Основные понятия: модель алгоритмов, метод обучения, функция потерь и функционал качества, принцип минимизации эмпирического риска, обобщающая способность, скользящий контроль. Методика экспериментального исследования и сравнения алгоритмов на модельных и реальных данных.

10. Генеративные состязательные сети (GAN).

Постановка задачи. Общая идея GAN. Генератор. Дискриминатор. Обучение Генеративно-состязательных сетей. Проблемы GAN: коллапс мод, исчезающие градиенты. Wasserstein GAN, Cramer GAN, Boundary Equilibrium GAN, Conditional GAN. Adversarial VAE. Alpha-GAN. Нормализующие потоки. Simple flows: planar, radial. Autoregressive flow: MAF, IAF. Probability distillation (conjugacy of MAF and IAF). FFJORD, PointFlow, GLOW, VQ-VAE-2.

11. Прогнозирование временных рядов.

Задача прогнозирования временных рядов. Примеры приложений. Экспоненциальное скользящее среднее. Модель Хольта. Модель Тейла-Вейджа. Модель Хольта-Уинтерса. Адаптивная авторегрессионная модель. Следящий контрольный сигнал. Модель Тригга-Лича. Адаптивная селективная модель. Адаптивная композиция моделей. Адаптация весов с регуляризацией.

5. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине (модулю)

Необходимое оборудование для занятий: компьютер и мультимедийное оборудование (проектор, звуковая система).

6. Перечень рекомендуемой литературы

Основная литература

1. Машинное обучение [Текст]/Х. Бринк, Дж. Ричардс, М. Феверолф, Real-World Machine Learning, -СПб., Питер, 2017

Дополнительная литература

1. Машинное обучение с подкреплением на Python /Министерство науки и высшего образования Российской Федерации, Московский физико-технический институт (национальный исследовательский университет), Кафедра системных исследований ; составители: А. И. Панов, А. А. Скрынник , Москва, МФТИ, 2019

7. Перечень ресурсов информационно-телекоммуникационной сети "Интернет", необходимых для освоения дисциплины (модуля)

8. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), включая перечень необходимого программного обеспечения и информационных справочных систем (при необходимости)

На занятиях используются мультимедийные технологии, включая демонстрацию презентаций.

9. Методические указания для обучающихся по освоению дисциплины (модуля)

Студент, изучающий дисциплину, должен с одной стороны, овладеть общим понятийным аппаратом, а с другой стороны, должен научиться применять теоретические знания на практике.

В результате изучения дисциплины студент должен знать основные определения, понятия, аксиомы, алгоритмы.

Успешное освоение курса требует напряжённой самостоятельной работы студента. В программе курса приведено минимально необходимое время для работы студента над темой. Самостоятельная работа включает в себя:

- чтение и конспектирование рекомендованной литературы;
- проработку учебного материала (по конспектам, учебной и научной литературе), подготовку ответов на вопросы, предназначенных для самостоятельного изучения;
- доказательство отдельных утверждений, свойств;
- подготовку к экзамену.

Руководство и контроль за самостоятельной работой студента осуществляется в форме индивидуальных консультаций.

Важно добиться понимания изучаемого материала, а не механического его запоминания. При затруднении изучения отдельных тем, вопросов, следует обращаться за консультациями к преподавателю.

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

по направлению:	Техническая физика
профиль подготовки:	Техническая физика космических летательных аппаратов Физтех-школа Аэрокосмических Технологий центр образовательных программ ФАКТ
курс:	<u>4</u>
квалификация:	бакалавр

Семестр, формы промежуточной аттестации: 7 (осенний) - Экзамен

Разработчики:

В.В. Стрижов, д-р физ.-мат. наук

В.Ю. Семака, преподаватель

1. Компетенции, формируемые в процессе изучения дисциплины

Код и наименование компетенции	Индикаторы достижения компетенции
ОПК-2 Способен применять методы математического анализа, математического моделирования и оптимизации для решения задач, возникающих в ходе профессиональной деятельности	ОПК-2.1 Знаком с основными методами математического анализа, математического моделирования и оптимизации
	ОПК-2.2 Способен строить математические модели, производить количественные расчеты и оценки
	ОПК-2.3 Способен определять границы применимости полученных результатов
ОПК-6 Способен работать с распределенными базами данных, с информацией в глобальных компьютерных сетях, применяя современные образовательные и информационные технологии	ОПК-6.2 Использует современные образовательные и информационные технологии и сервисы сети Интернет при решении задач в области профессиональной деятельности
ПК-1 Способен применять эффективные методы исследования физико-технических объектов и процессов, проводить испытания технологических процессов и (или) изделий с использованием современных аналитических средств технической физики	ПК-1.1 Способен использовать исследовательские методы при решении новых задач, применяя знания в различных областях науки (техники)
	ПК-1.2 Владеет аналитическими, вычислительными и экспериментальными методами исследования
	ПК-1.3 Способен разрабатывать и применять наиболее подходящие теоретические и экспериментальные методы исследований к конкретной научной задаче и интерпретировать полученные результаты
ПК-2 Готов изучать научно-техническую информацию, отечественный и зарубежный опыт по тематике профессиональной деятельности	ПК-2.1 Владеет методами научного поиска и интеллектуального анализа информации при решении задач профессиональной деятельности

2. Показатели оценивания компетенций

В результате изучения дисциплины «Введение в машинное обучение» обучающийся должен:

знать:

- фундаментальные понятия задач обучения по прецедентам;
- основные методы и алгоритмы решения задач обучения по прецедентам;
- основные области применения этих методов и алгоритмов.

уметь:

- применять методы и алгоритмы к решению задач обучения по прецедентам.

владеть:

- навыками самостоятельной работы при решении типовых задач;
- культурой постановки и моделирования практически значимых задач;
- практикой исследования и решения теоретических и прикладных задач;
- навыками теоретического анализа реальных задач, решаемых с помощью алгоритмов обучения по прецедентам.

3. Перечень типовых (примерных) вопросов, заданий, тем для подготовки к текущему контролю

Текущий контроль заключается в учете посещения студентами занятий а также в учете тех или иных видов активности студентов на занятиях: выполнения домашних заданий, решения задач, обсуждения возникающих вопросов по текущему материалу и т.п. Данные по текущему контролю учитываются при выставлении оценок на экзамене.

4. Перечень типовых (примерных) вопросов и тем для проведения промежуточной аттестации обучающихся

Промежуточная аттестация по дисциплине осуществляется в форме экзамена, проводимых в устной форме.

Перечень вопросов для сдачи экзамена:

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?

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.

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 исключений.

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 separable?

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?

Композиции алгоритмов классификации:

Дать определение алгоритмической композиции (помнить формулу). Какие типы корректирующих операций вы знаете?

Какие типы голосования вы знаете? Какой из них наиболее общий? (помнить формулу)

Как обнаружить объекты-выбросы при построении композиции классификаторов для голосования по большинству?

Как обеспечивается различность базовых алгоритмов при голосовании по большинству?

Как обеспечивается различность базовых алгоритмов при голосовании по старшинству?

Какие возможны стратегии выбора классов базовых алгоритмов при голосовании по старшинству?

Какие две эвристики лежат в основе алгоритма AdaBoost?

Как обнаружить объекты-выбросы в алгоритме AdaBoost?

Достоинства и недостатки алгоритма AdaBoost.

Основная идея алгоритма AnyBoost.

Основная идея метода bagging.

Основная идея метода случайных подпространств.

Что такое смесь экспертов (помнить формулу)?

Приведите примеры выпуклых функций потерь. Почему свойство выпуклости помогает строить смеси экспертов?

Логические алгоритмы классификации:

Что такое логическая закономерность? Приведите примеры закономерностей в задаче распознавания спама.

Часто используемые типы логических закономерностей.

Дайте определение эпсилон-дельта-логической закономерности (помнить формулы).

Дайте определение статистической закономерности (помнить формулы).

Сравните области статистических и логических закономерностей в (p,n) -плоскости.

С какой целью делается бинаризация?

В чём заключается процедура бинаризации признака?

Как происходит перебор в жадном алгоритме синтеза информативных конъюнкций?

Какие критерии информативности используются в жадном алгоритме синтеза информативных конъюнкций и почему?

Как приспособить жадный алгоритм синтеза конъюнкций для синтеза информативных шаров?

Что такое стохастический локальный поиск?

В чём отличия редукции и стабилизации? В чём их достоинства и недостатки?

Что такое решающий список?

Какие критерии информативности используются при синтезе решающего списка и почему?

Достоинства и недостатки решающих списков.

Что такое решающее дерево?

Какие критерии информативности используются при синтезе решающего дерева и почему?

Достоинства и недостатки решающих деревьев.

Зачем делается редукция решающих деревьев?

Какие есть два основных типа редукции решающих деревьев?

Как преобразовать решающее дерево в решающий список, и зачем это делается?

Что такое ADT (alternating decision tree)? Как происходит построение ADT?

Основная идея алгоритма КОРА.

Почему возникает проблема предпочтения признаков с меньшими номерами в алгоритме КОРА? Как она решается?

Основная идея алгоритма ТЭМП.

Какие критерии информативности используются в алгоритме ТЭМП и почему?

Почему возникает проблема дублирования закономерностей в алгоритме ТЭМП? Как она решается?

Достоинства и недостатки алгоритма ТЭМП.

Как использовать алгоритм AdaBoost для построения взвешенного голосования закономерностей?

Какой критерий информативности используется в алгоритме AdaBoost?

Структура алгоритма вычисления оценок (АВО).

Что такое ассоциативное правило? Приведите пример ассоциативного правила в задаче анализа потребительских корзин.

Основная идея алгоритма поиска ассоциативных правил APriority.

Кластеризация и таксономия:

Каковы основные цели кластеризации?

Основные типы кластерных структур. Приведите для каждой из этих структур пример алгоритма кластеризации, который для неё НЕ подходит.

В чём заключается алгоритм кратчайшего незамкнутого пути? Как его использовать для кластеризации? Как с его помощью определить число кластеров? Всегда ли это возможно?

Основная идея алгоритма ФорЭл.

Как вычисляются центры кластеров в алгоритме ФорЭл, если объекты — элементы метрического (не обязательно линейного векторного) пространства?

Какие существуют функционалы качества кластеризации и для чего они применяются?

Основные отличия алгоритма k-средних и ЕМ-алгоритма. Кто из них лучше и почему?

Основная идея иерархического алгоритма Ланса-Вильямса.

Какие основные типы расстояний между кластерами применяются в алгоритме Ланса-Вильямса?

Какие расстояния между кластерами, применяемые в алгоритме Ланса-Вильямса, лучше и почему?

Что такое дендрограмма? Всегда ли её можно построить?

Какой функционал качества оптимизируется сетью Кохонена? (помнить формулу)

В чем отличия правил мягкой и жёсткой конкуренции? В чём преимущества мягкой конкуренции?

Как устроена самоорганизующаяся карта Кохонена?

Как интерпретируются карты Кохонена?

Почему задачи с частичным обучением выделены в отдельный класс? Приведите примеры, когда методы классификации и кластеризации дают неадекватное решение задачи с частичным обучением.

Как приспособить графовые алгоритмы кластеризации для решения задачи с частичным обучением?

Как приспособить ЕМ-алгоритм для решения задачи с частичным обучением?

Какие способы решения задачи с частичным обучением Вы знаете?

Перечень экзаменационных билетов для сдачи экзамена:

Билет1

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.

Билет2

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 основан линейный дискриминант Фишера?

Что такое «проблема мультиколлинеарности», в каких задачах и при использовании каких алгоритмов она возникает? Какие есть подходы к её решению?

Билет3

Что такое «смесь распределений» (надо помнить формулу)?

Что такое ЕМ-алгоритм, какова его основная идея? Какая задача решается на Е-шаге, на М-шаге? Каков вероятностный смысл скрытых переменных?

Последовательное добавление компонент в ЕМ-алгоритме, основная идея алгоритма.

Что такое стохастический ЕМ-алгоритм, какова основная идея? В чём его преимущество (какой недостаток стандартного ЕМ-алгоритма он устраняет)?

Билет4

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?

Билет5

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.

Билет 6

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.

Билет 7

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?

Билет 8

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).

Билет 9

Что такое ядро в SVM?

Что такое окна переменной ширины, и зачем они нужны?

Постановка задачи многомерной линейной регрессии. Матричная запись.

Вывести формулу непараметрической регрессии Надарая-Ватсона.

Билет 10

Каков вероятностный смысл регуляризации? Какие типы регуляризаторов Вы знаете?

Как свести задачу многомерной нелинейной регрессии к последовательности линейных задач?

Основная идея отбора признаков с помощью случайного поиска.

Приведите примеры выпуклых функций потерь. Почему свойство выпуклости помогает строить смеси экспертов?

Критерии оценивания

Оценка отлично 10 баллов - выставляется студенту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины, проявляющему интерес к данной предметной области, продемонстрировавшему умение уверенно и творчески применять их на практике при решении конкретных задач, свободное и правильное обоснование принятых решений.

Оценка отлично 9 баллов - выставляется студенту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины и умение уверенно применять их на практике при решении конкретных задач, свободное и правильное обоснование принятых решений.

Оценка отлично 8 баллов - выставляется студенту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины и умение уверенно применять их на практике при решении конкретных задач, правильное обоснование принятых решений, с некоторыми недочетами.

Оценка хорошо 7 баллов - выставляется студенту, если он твердо знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но недостаточно грамотно обосновывает полученные результаты.

Оценка хорошо 6 баллов - выставляется студенту, если он твердо знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач некоторые неточности.

Оценка хорошо 5 баллов - выставляется студенту, если он в основном знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач достаточно большое количество неточностей.

Оценка удовлетворительно 4 балла - выставляется студенту, показавшему фрагментарный, разрозненный характер знаний, недостаточно правильные формулировки базовых понятий, нарушения логической последовательности в изложении программного материала, но при этом он освоил основные разделы учебной программы, необходимые для дальнейшего обучения, и может применять полученные знания по образцу в стандартной ситуации.

Оценка удовлетворительно 3 балла - выставляется студенту, показавшему фрагментарный, разрозненный характер знаний, допускающему ошибки в формулировках базовых понятий, нарушения логической последовательности в изложении программного материала, слабо владеет основными разделами учебной программы, необходимыми для дальнейшего обучения и с трудом применяет полученные знания даже в стандартной ситуации.

Оценка неудовлетворительно 2 балла - выставляется студенту, который не знает большей части основного содержания учебной программы дисциплины, допускает грубые ошибки в формулировках основных принципов и не умеет использовать полученные знания при решении типовых задач.

Оценка неудовлетворительно 1 балл - выставляется студенту, который не знает основного содержания учебной программы дисциплины, допускает грубейшие ошибки в формулировках базовых понятий дисциплины и вообще не имеет навыков решения типовых практических задач.

5. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности

Во время проведения экзамена обучающиеся могут пользоваться программой дисциплины, а также справочной литературой, вычислительной техникой и конспектами.

Экзамен может проводиться по итогам текущей успеваемости и сдачи заданий, или путем организации специального опроса, проводимого в устной форме.