

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

УТВЕРЖДЕНО
Проректор по учебной работе

А.А. Воронов

	Рабочая программа дисциплины (модуля)
по дисциплине:	Анализ данных и методы машинного обучения
по направлению:	Биотехнология
профиль подготовки:	Управление инновациями в бизнесе Физтех-школа бизнеса высоких технологий кафедра информатики и вычислительной математики
курс:	2
квалификация:	бакалавр

Семестр, формы промежуточной аттестации: 4 (весенний) - Дифференцированный зачет

Аудиторных часов: 60 всего, в том числе:

лекции: 20 час.

семинары: 0 час.

лабораторные занятия: 40 час.

Самостоятельная работа: 75 час.

Всего часов: 135, всего зач. ед.: 3

Количество контрольных работ, заданий: 2

Программу составил: Т.Ф. Хирьянов, старший преподаватель

Программа обсуждена на заседании кафедры информатики и вычислительной математики 31.08.2023

Аннотация

В ходе изучения дисциплины студенты освоят современные методы машинного обучения, направленные на решение задачи восстановления зависимостей по эмпирическим данным, кластерный и регрессионный анализ. Курс рассчитан на постепенное погружение в предметную область, начиная от основных понятий и привязки к прикладным задачам. Будут изложены метрические методы, а также логические и линейные методы классификации. В рамках курса студенты приобретут практические навыки построения и работы с нейронными сетями.

1. Цели и задачи

Цель дисциплины

- сформировать теоретические и практические знания в области обучения машин, современных методов восстановления зависимостей по эмпирическим данным, включая дискриминантный, кластерный и регрессионный анализ.

Задачи дисциплины

- правильно формулировать задачу в терминах машинного обучения;
- овладеть навыками практического решения задач интеллектуального анализа данных.

2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

Код и наименование компетенции	Индикаторы достижения компетенции
ОПК-1 Способен изучать, анализировать, использовать биологические объекты и процессы, основываясь на математических, физических, химических, биологических законах, закономерностях и взаимосвязях	ОПК-1.1 Способен анализировать поставленную задачу, намечать пути ее решения
ПК-3 Способен выбирать и применять подходящее оборудование, инструменты и методы исследований для решения задач в избранной предметной области	ПК-3.3 Умеет производить оценку точности численных методов, используемых на ЭВМ, вычислительной сложности используемых алгоритмов и объема требуемых вычислительных ресурсов

3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны

знать:

- основные принципы и проблематику теории обучения машин;
- основные методы и алгоритмы решения задач обучения по прецедентам;
- основные области применения этих методов и алгоритмов;
- классификации, кластеризации и регрессии.

уметь:

- формализовать постановки прикладных задач анализа данных;
- использовать методы обучения по прецедентам для решения практических задач;
- оценивать точность и эффективность полученных решений.

владеть:

- основными понятиями теории машинного обучения;
- навыками самостоятельной работы при решении типовых задач;
- культурой постановки и моделирования практически значимых 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	Основные понятия и примеры прикладных задач	2		4	7
2	Метрические методы	2		4	7
3	Отбор признаков, постоеение	2		4	7
4	Логические методы классификации	2		4	7
5	Линейные методы классификации	2		4	7
6	Методы опорных векторов	2		4	7
7	Многомерная линейная регрессия	2		4	7
8	Байесовская классификация	2		4	7
9	Логическая регрессия	2		4	7
10	Многослойные нейронные сети	1		2	7
11	Методы кластеризации	1		2	5
Итого часов		20		40	75
Подготовка к экзамену		0 час.			
Общая трудоёмкость		135 час., 3 зач.ед.			

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

Семестр: 4 (Весенний)

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.
- Конкурсы по анализу данных kaggle.com. Полигон алгоритмов классификации.
- CRISP-DM — межотраслевой стандарт ведения проектов интеллектуального анализа данных.

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.
- Непараметрическая регрессия. Локально взвешенный метод наименьших квадратов. Ядерное сглаживание.
- Оценка Надарая-Ватсона с постоянной и переменной шириной окна. Выбор функции ядра.

- Задача отсева выбросов. Робастная непараметрическая регрессия. Алгоритм LOWESS.
- Задача отбора эталонов. Понятие отступа. Алгоритм СТОЛП.
- Задача отбора признаков. Жадный алгоритм построения метрики.

2. Метрические методы

Логические методы классификации

- Понятие логической закономерности.
- Параметрические семейства закономерностей: конъюнкции пороговых правил, синдромные правила, шары, гиперплоскости.
- Переборные алгоритмы синтеза конъюнкций: стохастический локальный поиск, стабилизация, редукция.
- Двухкритериальный отбор информативных закономерностей, парето-оптимальный фронт в (p,n) -пространстве.
- Решающее дерево. Жадная нисходящая стратегия «разделяй и властвуй». Алгоритм ID3. Недостатки жадной стратегии и способы их устранения. Проблема переобучения.
- Вывод критериев ветвления. Мера нечистоты (impurity) распределения. Энтропийный критерий, критерий Джини.
- Редукция решающих деревьев: предредукция и постредукция. Алгоритм C4.5.
- Деревья регрессии. Алгоритм CART.
- Небрежные решающие деревья (oblivious decision tree).
- Решающий лес. Случайный лес (Random Forest).

Факультатив

- 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.
- Решающий пень. Бинаризация признаков. Алгоритм разбиения области значений признака на информативные зоны.
- Решающий список. Жадный алгоритм синтеза списка.
- Преобразование решающего дерева в решающий список.

Градиентные методы обучения

- Линейный классификатор, модель МакКаллока-Питтса, непрерывные аппроксимации пороговой функции потерь.
- Метод стохастического градиента SG.
- Метод стохастического среднего градиента SAG.
- 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.
- 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. Сглаженное правило Хэбба. Многоклассовая логистическая регрессия. Регуляризованная логистическая регрессия. Калибровка Платта.

3. Отбор признаков, построение

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

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.

4. Логические методы классификации

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.

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.
- Локальная адаптация весов с регуляризацией.

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.
- Агрегированные и многоступенчатые критерии.
- 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).

/

5. Линейные методы классификации

Байесовская классификация и оценивание плотности

- Принцип максимума апостериорной вероятности. 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.
- Параметрическое оценивание плотности. Нормальный дискриминантный анализ.
- Многомерное нормальное распределение, геометрическая интерпретация. Выборочные оценки параметров многомерного нормального распределения.
- Квадратичный дискриминант. Вид разделяющей поверхности. Подстановочный алгоритм, его недостатки и способы их устранения.
- Линейный дискриминант Фишера.
- Проблемы мультиколлинеарности и переобучения. Регуляризация ковариационной матрицы.
- Параметрический наивный байесовский классификатор.
- Смесь распределений.
- ЕМ-алгоритм как метод простых итераций для решения системы нелинейных уравнений.
- Выбор числа компонентов смеси. Пошаговая стратегия. Априорное распределение Дирихле.
- Смесь многомерных нормальных распределений. Сеть радиальных базисных функций (RBF) и применение ЕМ-алгоритма для её настройки.
- Сравнение RBF-сети и SVM с гауссовским ядром.

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 ЕМ-алгоритм для разделения гауссовской смеси.
- Графовые алгоритмы кластеризации. Выделение связанных 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 редуکتивности. Псевдокод редуکتивной 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.

Поиск ассоциативных правил

- Понятие ассоциативного правила и его связь с понятием логической закономерности.
- Примеры прикладных задач: анализ рыночных корзин, выделение терминов и тематики текстов.
- Алгоритм APriori. Два этапа: поиск частых наборов и рекурсивное порождение ассоциативных правил. Недостатки и пути совершенствования алгоритма APriori.
- Алгоритм FP-growth. Понятия FP-дерева и условного FP-дерева. Два этапа поиска частых наборов в FP-growth: построение FP-дерева и рекурсивное порождение частых наборов.

- Общее представление о динамических и иерархических методах поиска ассоциативных правил.

6. Методы опорных векторов

Нейронные сети

- Биологический нейрон, модель МакКаллока-Питтса как линейный классификатор. Функции активации.
- Проблема полноты. Задача исключающего или. Полнота двухслойных сетей в пространстве булевых функций.
- Теоремы Колмогорова, Стоуна, Горбаня (без доказательства).
- Алгоритм обратного распространения ошибок.
- Эвристики: формирование начального приближения, ускорение сходимости, диагональный метод Левенберга-Марквардта. Проблема «паралича» сети.
- Метод послойной настройки сети.
- Подбор структуры сети: методы постепенного усложнения сети, оптимальное прореживание нейронных сетей (optimal brain damage).
- Нейронная сеть Кохонена. Конкурентное обучение, стратегии WTA и WTM.
- Самоорганизующаяся карта Кохонена. Применение для визуального анализа данных. Искусство интерпретации карт Кохонена.

Нейронные сети глубокого обучения

- Быстрые методы стохастического градиента: Поляка, Нестерова, AdaGrad, RMSProp, AdaDelta, Adam, Nadam.
- Проблема взрыва градиента и эвристика gradient clipping
- Метод случайных отключений нейронов (Dropout). Интерпретации Dropout. Обратный Dropout и L2-регуляризация.
- Функции активации ReLU и PReLU.
- 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)

7. Многомерная линейная регрессия

- Основные понятия: базовый алгоритм (алгоритмический оператор), корректирующая операция.
- Взвешенное голосование.
- Алгоритм AdaBoost. Экспоненциальная аппроксимация пороговой функции потерь. Процесс последовательного обучения базовых алгоритмов. Теорема о сходимости бустинга.
- Обобщающая способность бустинга.
- Базовые алгоритмы в бустинге. Решающие пни.
- Варианты бустинга: GentleBoost, LogitBoost, BrownBoost, и другие.
- Алгоритм AnyBoost.
- Градиентный бустинг. Стохастический градиентный бустинг.
- Простое голосование (комитет большинства). Алгоритм ComBoost. Идентификация нетипичных объектов (выбросов).
- Преобразование простого голосования во взвешенное.
- Обобщение на большое число классов.
- Решающий список (комитет старшинства). Алгоритм обучения. Стратегия выбора классов для базовых алгоритмов.

8. Байесовская классификация

Эвристические, стохастические, нелинейные композиции

- Стохастические методы: бэггинг и метод случайных подпространств.
- Случайный лес. Анализ смещения и вариации для простого голосования.
- Смесь алгоритмов (квазилинейная композиция), область компетентности, примеры функций компетентности.
- Выпуклые функции потерь. Методы построения смесей: последовательный и иерархический.
- Построение смеси алгоритмов с помощью ЕМ-подобного алгоритма.
- Нелинейная монотонная корректирующая операция. Случай классификации. Случай регрессии. Задача монотонизации выборки, изотонная регрессия.

Ранжирование

- 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, ОС-SVM.
- Парный подход: RankingSVM, RankNet, LambdaRank.

9. Логическая регрессия

Рекомендательные системы

- 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 признаков данных. Линейная и квадратичная регрессионные модели, libFM.
- Измерение качества рекомендаций. Меры разнообразия (diversity), новизны (novelty), покрытия (coverage), догадливости (serendipity).

Тематическое моделирование

- Задача тематического моделирования коллекции текстовых документов.
- Вероятностный латентный семантический анализ PLSA. Метод максимума правдоподобия. ЕМ-алгоритм. Элементарная интерпретация ЕМ-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 ЕМ-алгоритм, теорема о стационарной точке (применение условий Каруша–Куна–Таккера).
- Рациональный ЕМ-алгоритм. Онлайн-алгоритм и его распараллеливание.
- Мультимодальная тематическая модель.
- Регуляризаторы классификации и регрессии.
- Регуляризаторы декоррелирования и отбора тем.
- Внутренние и внешние критерии качества тематических моделей.

10. Многослойные нейронные сети

- 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.

11. Методы кластеризации

- 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 обучения. Алгоритмы ϵ -active и EG-active.
- Применение обучения с подкреплением для активного обучения. Активное томпсоновское сэмплирование.

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

Учебная аудитория, оснащенная компьютером и мультимедийным оборудованием (проектор, звуковая система).

6. Перечень рекомендуемой литературы

Основная литература

1. Методы распознавания [Текст] : учеб. пособие / А. Л. Горелик, В. А. Скрипкин .— 4-е изд., испр. — М. : Высшая школа, 2004 .— 261 с.

Рекомендуется для самостоятельного изучения:

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.
4. Мерков А. Б. Распознавание образов. Построение и обучение вероятностных моделей. 2014. 238 с.
5. Коэльо Л.П., Ричарт В. Построение систем машинного обучения на языке Python. 2016. 302 с.
6. Головкин В. А. Нейронные сети : Обучение, организация и применение : учеб. пособие для вузов .— М. : ИПРЖР, 2001 .256 с.

Дополнительная литература

1. Теория распознавания образов : Статистические методы [Текст] : учеб. пособие для вузов / А. А. Натан ; М-во высш. и средн. спец. образования РСФСР, Моск. физ.-техн. ин-т .— М. : МФТИ, 1988 .— 84 с. : ил. - 400 экз.

Рекомендуется для самостоятельного изучения:

1. Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. Классификация и снижение размерности. — М. Финансы и статистика. 1989.
2. Айвазян С. А., Енюков И. С., Мешалкин Л. Д. Исследование зависимостей. — М. Финансы и статистика. 1985.
3. Айвазян С. А., Мхитарян В. С. Прикладная статистика и основы эконометрики. — М.: Юнити, 1998.
4. Вагин В. Н., Головина Е. Ю., Загорянская А. А., Фомина М. В. Достоверный и правдоподобный вывод в интеллектуальных системах. — М.: Физматлит. 2004.
5. Вапник В. Н. Восстановление зависимостей по эмпирическим данным. — М.: Наука. 1979.
6. Воронцов К. В. Комбинаторный подход к оценке качества обучаемых алгоритмов // Математические вопросы кибернетики. <http://www.ccas.ru/voron>.
7. Головкин В. А. Нейронные сети: обучение, организация и применение. — М.: ИПРЖР. 2001.
8. Загоруйко Н. Г. Прикладные методы анализа данных и знаний. — Новосибирск: ИМ СО РАН, 1999.
9. Загоруйко Н. Г., Ёлкина В. Н., Лбов Г. С. Алгоритмы обнаружения эмпирических закономерностей. — Новосибирск: Наука, 1985.
10. Ивахненко А. Г., Юрачковский Ю. П. Моделирование сложных систем по экспериментальным данным. — М.: Радио и связь, 1987.
11. Журавлёв Ю. И., Рязанов В. В., Сенько О. В. РАСПОЗНАВАНИЕ. Математические методы. Программная система. Применения. — Москва: Фазис, 2006.
12. Журавлев Ю. И. Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики. — 1978. — Т. 33. — С. 5–68.
13. Казанцев В. С. Задачи классификации и их программное обеспечение. — М. Наука. 1990.
14. Лоусон Ч, Хенсон Р. Численное решение задач метода наименьших квадратов. — М. Наука. 1986.
15. Магнус Я. Р., Катышев П. К., Пересецкий А. А. Эконометрика: начальный курс. М.: Дело. 2004.
16. Саттон Р. С., Барто Э. Г. Обучение с подкреплением. — БИНОМ, 2011.
17. Хардл В. Прикладная непараметрическая регрессия. — М.: Мир. 1993.
18. Шурыгин А. М. Прикладная стохастика: робастность, оценивание, прогноз. — М. Финансы и статистика. 2000.
19. Burges C. J. C. A tutorial on support vector machines for pattern recognition // Data Mining and Knowledge Discovery. — 1998. — Vol. 2, no. 2. — Pp. 121–167. <http://citeseer.ist.psu.edu/burges98tutorial.html>.
20. Martin J. K. An exact probability metric for decision tree splitting and stopping // Machine Learning. — 1997. — Vol. 28, no. 2-3. — Pp. 257–291. <http://citeseer.ist.psu.edu/martin97exact.html>.
21. Marchand M., Shawe-Taylor J. Learning with the set covering machine // Proc. 18th International Conf. on Machine Learning. — Morgan Kaufmann, San Francisco, CA, 2001. — Pp. 345–352. <http://citeseer.ist.psu.edu/452556.html>.
22. Schapire R. The boosting approach to machine learning: An overview // MSRI Workshop on Nonlinear Estimation and Classification, Berkeley, CA. — 2001. <http://citeseer.ist.psu.edu/schapire02boosting.html>.

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

1. <http://www.machinelearning.ru> – профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных.
2. <http://shad.yandex.ru> – сайт школы анализа данных Яндекса.
3. http://www.machinelearning.ru/wiki/index.php?title=%D0%9C%D0%B0%D1%88%D0%B8%D0%BD%D0%BD%D0%BE%D0%B5_%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5_%28%D0%BA%D1%83%D1%80%D1%81_%D0%BB%D0%B5%D0%BA%D1%86%D0%B8%D0%B9%2C_%D0%9A.%D0%92.%D0%92%D0%BE%D1%80%D0%BE%D0%BD%D1%86%D0%BE%D0%B2%29

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

На лекционных занятиях используются мультимедийные технологии, включая демонстрацию презентаций.

В процессе самостоятельной работы обучающихся предполагается использование таких программных средств, как WEKA, IPython Notebook и др.

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

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:

- проработку учебного материала (по конспектам лекций, учебной и научной литературе);
- выполнение домашних теоретических заданий;
- подготовку к дифференцированному зачету.

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

по направлению: Биотехнология
профиль подготовки: Управление инновациями в бизнесе
Физтех-школа бизнеса высоких технологий
кафедра информатики и вычислительной математики
курс: 2
квалификация: бакалавр

Семестр, формы промежуточной аттестации: 4 (весенний) - Дифференцированный зачет

Разработчик: Т.Ф. Хирьянов, старший преподаватель

1. Компетенции, формируемые в процессе изучения дисциплины

Код и наименование компетенции	Индикаторы достижения компетенции
ОПК-1 Способен изучать, анализировать, использовать биологические объекты и процессы, основываясь на математических, физических, химических, биологических законах, закономерностях и взаимосвязях	ОПК-1.1 Способен анализировать поставленную задачу, намечать пути ее решения
ПК-3 Способен выбирать и применять подходящее оборудование, инструменты и методы исследований для решения задач в избранной предметной области	ПК-3.3 Умеет производить оценку точности численных методов, используемых на ЭВМ, вычислительной сложности используемых алгоритмов и объема требуемых вычислительных ресурсов

2. Показатели оценивания компетенций

В результате изучения дисциплины «Анализ данных и методы машинного обучения» обучающийся должен:

знать:

- основные принципы и проблематику теории обучения машин;
- основные методы и алгоритмы решения задач обучения по прецедентам;
- основные области применения этих методов и алгоритмов;
- классификации, кластеризации и регрессии.

уметь:

- формализовать постановки прикладных задач анализа данных;
- использовать методы обучения по прецедентам для решения практических задач;
- оценивать точность и эффективность полученных решений.

владеть:

- основными понятиями теории машинного обучения;
- навыками самостоятельной работы при решении типовых задач;
- культурой постановки и моделирования практически значимых 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

С целью контроля освоения обучающимися учебного материала проводится устный опрос в начале занятия по теме прошлого занятия.

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

Перечень контрольных вопросов для сдачи дифференцированного зачёта:

Байесовская классификация

1. Записать общую формулу байесовского классификатора (надо помнить формулу).
2. Какие вы знаете три подхода к восстановлению плотности распределения по выборке?
3. Что такое наивный байесовский классификатор?
4. Что такое оценка плотности Парзена-Розенблатта (надо помнить формулу). Выписать формулу алгоритма классификации в методе парзеновского окна.
5. На что влияет ширина окна, а на что вид ядра в методе парзеновского окна?
6. Многомерное нормальное распределение (надо помнить формулу). Вывести формулу квадратичного дискриминанта. При каком условии он становится линейным?
7. На каких предположениях основан линейный дискриминант Фишера?
8. Что такое «проблема мультиколлинеарности», в каких задачах и при использовании каких алгоритмов она возникает? Какие есть подходы к её решению?
9. Что такое «смесь распределений» (надо помнить формулу)?

10. Что такое ЕМ-алгоритм, какова его основная идея? Какая задача решается на Е-шаге, на М-шаге? Каков вероятностный смысл скрытых переменных?
11. Последовательное добавление компонент в ЕМ-алгоритме, основная идея алгоритма.
12. Что такое стохастический ЕМ-алгоритм, какова основная идея? В чём его преимущество (какой недостаток стандартного ЕМ-алгоритма он устраняет)?
13. Что такое сеть радиальных базисных функций?
14. Что такое «выбросы»? Как осуществляется фильтрация выбросов?

Метрическая классификация

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

Линейная классификация

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

Регрессия

1. Что такое ядерное сглаживание?
2. Что есть общего между ядром в непараметрической регрессии и ядром SVM?
3. На что влияет ширина окна, а на что вид ядра в непараметрической регрессии?
4. Что такое окна переменной ширины, и зачем они нужны?

5. Что такое «выбросы»? Как осуществляется фильтрация выбросов в непараметрической регрессии?
6. Постановка задачи многомерной линейной регрессии. Матричная запись.
7. Что такое сингулярное разложение? Как оно используется для решения задачи наименьших квадратов?
8. Что такое «проблема мультиколлинеарности» в задачах многомерной линейной регрессии? Какие есть три подхода к её устранению?
9. Сравнить гребневую регрессию и лассо. В каких задачах предпочтительнее использовать лассо?
10. Какую проблему решает метод главных компонент в многомерной линейной регрессии? Записать матричную постановку задачи для метода главных компонент.
11. Как свести задачу многомерной нелинейной регрессии к последовательности линейных задач?
12. Метод настройки с возвращениями (backfitting): постановка задачи и основная идея метода.
13. Какие методы построения логистической регрессии Вы знаете?
14. Приведите примеры неквадратичных функций потерь в регрессионных задачах. С какой целью они вводятся?

Примеры задач

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

Выбор модели и отбор признаков

1. В чём отличия внутренних и внешних критериев?
2. Разновидности внешних критериев.
3. Разновидности критерия скользящего контроля.
4. Что такое критерий непротиворечивости? В чём его недостатки?
5. Что такое многоступенчатый выбор модели по совокупности критериев?
6. Основная идея отбора признаков методом полного перебора. Действительно ли это полный перебор?
7. Основная идея отбора признаков методом добавлений и исключений.
8. Что такое шаговая регрессия? Можно ли её использовать для классификации, в каком методе?
9. Основная идея отбора признаков методом поиска в глубину.
10. Основная идея отбора признаков методом поиска в ширину.
11. Что такое МГУА?
12. Основная идея отбора признаков с помощью генетического алгоритма.
13. Основная идея отбора признаков с помощью случайного поиска.
14. В чём отличия случайного поиска от случайного поиска с адаптацией?

Нейронные сети

1. 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?

2. Почему любая булева функция представима в виде нейронной сети? Сколько в ней слоёв?
3. Метод обратного распространения ошибок. Основная идея. Основные недостатки и способы их устранения.
4. Как можно выбирать начальное приближение в градиентных методах настройки нейронных сетей?
5. Как можно ускорить сходимость в градиентных методах настройки нейронных сетей?
6. Что такое диагональный метод Левенберга-Марквардта?
7. Что такое «паралич» сети, и как его избежать?
8. Как выбирать число слоёв в градиентных методах настройки нейронных сетей?
9. Как выбирать число нейронов скрытого слоя в градиентных методах настройки нейронных сетей?
10. В чём заключается метод оптимального прореживания нейронной сети? Какие недостатки стандартного алгоритма обратного распространения ошибок позволяет устранить метод ODB?

Композиции алгоритмов классификации

1. Дать определение алгоритмической композиции (помнить формулу). Какие типы корректирующих операций вы знаете?
2. Какие типы голосования вы знаете? Какой из них наиболее общий? (помнить формулу)
3. Как обнаружить объекты-выбросы при построении композиции классификаторов для голосования по большинству?
4. Как обеспечивается различность базовых алгоритмов при голосовании по большинству?
5. Как обеспечивается различность базовых алгоритмов при голосовании по старшинству?
6. Какие возможны стратегии выбора классов базовых алгоритмов при голосовании по старшинству?
7. Какие две эвристики лежат в основе алгоритма AdaBoost?
8. Как обнаружить объекты-выбросы в алгоритме AdaBoost?
9. Достоинства и недостатки алгоритма AdaBoost.
10. Основная идея алгоритма AnyBoost.
11. Основная идея метода bagging.
12. Основная идея метода случайных подпространств.
13. Что такое смесь экспертов (помнить формулу)?
14. Приведите примеры выпуклых функций потерь. Почему свойство выпуклости помогает строить смеси экспертов?

Логические алгоритмы классификации

1. Что такое логическая закономерность? Приведите примеры закономерностей в задаче распознавания спама.
2. Часто используемые типы логических закономерностей.
3. Дайте определение эпсилон-дельта-логической закономерности (помнить формулы).
4. Дайте определение статистической закономерности (помнить формулы).
5. Сравните области статистических и логических закономерностей в (p, n) -плоскости.
6. С какой целью делается бинаризация?
7. В чём заключается процедура бинаризации признака?
8. Как происходит перебор в жадном алгоритме синтеза информативных конъюнкций?
9. Какие критерии информативности используются в жадном алгоритме синтеза информативных конъюнкций и почему?
10. Как приспособить жадный алгоритм синтеза конъюнкций для синтеза информативных шаров?
11. Что такое стохастический локальный поиск?
12. В чём отличия редукции и стабилизации? В чём их достоинства и недостатки?
13. Что такое решающий список?
14. Какие критерии информативности используются при синтезе решающего списка и почему?
15. Достоинства и недостатки решающих списков.
16. Что такое решающее дерево?

17. Какие критерии информативности используются при синтезе решающего дерева и почему?
18. Достоинства и недостатки решающих деревьев.
19. Зачем делается редукция решающих деревьев?
20. Какие есть два основных типа редукции решающих деревьев?
21. Как преобразовать решающее дерево в решающий список, и зачем это делается?
22. Что такое ADT (alternating decision tree)? Как происходит построение ADT?
23. Основная идея алгоритма КОРА.
24. Почему возникает проблема предпочтения признаков с меньшими номерами в алгоритме КОРА? Как она решается?
25. Основная идея алгоритма ТЭМП.
26. Какие критерии информативности используются в алгоритме ТЭМП и почему?
27. Почему возникает проблема дублирования закономерностей в алгоритме ТЭМП? Как она решается?
28. Достоинства и недостатки алгоритма ТЭМП.
29. Как использовать алгоритм AdaBoost для построения взвешенного голосования закономерностей?
30. Какой критерий информативности используется в алгоритме AdaBoost?
31. Структура алгоритма вычисления оценок (АВО).
32. Что такое ассоциативное правило? Приведите пример ассоциативного правила в задаче анализа потребительских корзин.
33. Основная идея алгоритма поиска ассоциативных правил APriority.

Кластеризация и таксономия

1. Каковы основные цели кластеризации?
2. Основные типы кластерных структур. Приведите для каждой из этих структур пример алгоритма кластеризации, который для неё НЕ подходит.
3. В чём заключается алгоритм кратчайшего незамкнутого пути? Как его использовать для кластеризации? Как с его помощью определить число кластеров? Всегда ли это возможно?
4. Основная идея алгоритма ФорЭл.
5. Как вычисляются центры кластеров в алгоритме ФорЭл, если объекты — элементы метрического (не обязательно линейного векторного) пространства?
6. Какие существуют функционалы качества кластеризации и для чего они применяются?
7. Основные отличия алгоритма k-средних и ЕМ-алгоритма. Кто из них лучше и почему?
8. Основная идея иерархического алгоритма Ланса-Вильямса.
9. Какие основные типы расстояний между кластерами применяются в алгоритме Ланса-Вильямса?
10. Какие расстояния между кластерами, применяемые в алгоритме Ланса-Вильямса, лучше и почему?
11. Что такое дендрограмма? Всегда ли её можно построить?
12. Какой функционал качества оптимизируется сетью Кохонена? (помнить формулу)
13. В чём отличия правил мягкой и жёсткой конкуренции? В чём преимущества мягкой конкуренции?
14. Как устроена самоорганизующаяся карта Кохонена?
15. Как интерпретируются карты Кохонена?
16. Почему задачи с частичным обучением выделены в отдельный класс? Приведите примеры, когда методы классификации и кластеризации дают неадекватное решение задачи с частичным обучением.
17. Как приспособить графовые алгоритмы кластеризации для решения задачи с частичным обучением?
18. Как приспособить ЕМ-алгоритм для решения задачи с частичным обучением?
19. Какие способы решения задачи с частичным обучением Вы знаете?

Билет 1

1. Как устроена самоорганизующаяся карта Кохонена?
2. Как интерпретируются карты Кохонена?

Билет 2

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

Билет 3

1. Что такое дендрограмма? Всегда ли её можно построить?
2. Какой функционал качества оптимизируется сетью Кохонена?

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.

Пример для билета из трех вопросов:

1. Решающие списки: принцип работы, схема алгоритма построения по обучающей выборке, стратегии выбора классов при построении. Примеры задач, не решаемых решающими списками.
2. Метод опорных векторов. Оптимизационная задача с ограничениями в виде неравенств и безусловная. Опорные векторы. Kerneltrick. Оптимизационная задача в S3VM и SVR.

3. Логистическая регрессия. Принцип максимума правдоподобия и логарифмическая функция потерь. Метод стохастического градиента в логистической регрессии.

За первое задание студент получает от 0 до 4 баллов, за второе и третье – от 0 до 3 баллов за каждое в зависимости от полноты представленного ответа (решения). Количество набранных баллов определяет оценку за дифференцированный зачёт:

Оценка	Набранные баллы
отлично (10)	более 9
отлично (9)	от 8 до 9 включительно
хорошо (8)	от 7 до 8 включительно
хорошо (7)	от 6 до 7 включительно
хорошо (6)	от 5 до 6 включительно
удовлетворительно (5)	от 4 до 5 включительно
удовлетворительно (4)	от 3 до 4 включительно
удовлетворительно (3)	от 2 до 3 включительно
неудовлетворительно (2)	от 1 до 2 включительно
неудовлетворительно (1)	не более 1

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

Во время проведения дифференцированного зачёта обучающиеся могут пользоваться программой дисциплины, а также справочной литературой, вычислительной техникой, конспектами лекций.

Дифференцированный зачёт проводится в устной форме, но учитывает и текущую успеваемость на лабораторных работах, итоги сдачи заданий.