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

УТВЕРЖДЕНО
Проректор по учебной работе и довузовской подготовке
А.А. Воронов

Рабочая программа дисциплины (модуля)

по дисциплине: Методы анализа данных и распознавания
по направлению: Прикладные математика и физика
профиль подготовки: Общая и прикладная физика
Физтех-школа физики и исследований им. Ландау
кафедра информатики и вычислительной математики
курс: 1
квалификация: магистр

Семестры, формы промежуточной аттестации:
1 (осенний) - Дифференцированный зачет
2 (весенний) - Экзамен

Аудиторных часов: 120 всего, в том числе:
лекции: 60 час.
семинары: 30 час.
лабораторные занятия: 30 час.

Самостоятельная работа: 120 час.

Подготовка к экзамену: 30 час.

Всего часов: 270, всего зач. ед.: 6

Программу составил: В.В. Рязанов, д-р физ.-мат. наук, старший научный сотрудник, профессор

Программа обсуждена на заседании кафедры информатики и вычислительной математики 04.06.2020

Аннотация

В курсе рассматриваются ключевые понятия и современные математические методы и алгоритмы анализа данных и распознавания, а также их приложения в различных областях.

1. Цели и задачи

Цель дисциплины

изучение современных подходов, моделей, алгоритмов анализа данных и решения задач распознавания, классификации, нахождения зависимостей.

Задачи дисциплины

- освоение студентами базовых знаний в области методов анализа данных и распознавания (МАДР);
- приобретение теоретических знаний в области анализа прецедентных данных в условиях их частичной противоречивости и неполноты;
- оказание консультаций и помощи студентам в проведении собственных теоретических и экспериментальных исследований в области МАДР;
- формирование навыков применения МАДР при исследовании экспериментальных, статистических или экспертных данных при выполнении студентами выпускных работ на степень магистра.

2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

Код и наименование компетенции	Индикаторы достижения компетенции
ОПК-3 Способен выбирать и (или) разрабатывать подходы к решению типовых и новых задач в области профессиональной деятельности, учитывая особенности и ограничения различных методов решения	ОПК-3.1 Способен анализировать задачу, планировать пути решения, предлагать и комбинировать способы решения
	ОПК-3.2 Способен использовать исследовательские методы при решении новых задач, применяя знания в различных областях науки (техники)
	ОПК-3.3 Владеет аналитическими и вычислительными методами решения, понимает и учитывает на практике границы применимости получаемых решений
ПК-3 Способен профессионально работать с исследовательским и испытательным оборудованием (приборами и установками, специализированными пакетами прикладных программ) в избранной предметной области	ПК-3.1 Понимает принципы работы используемого оборудования (специализированных пакетов прикладных программ)
	ПК-3.2 Способен проводить эксперимент (моделирование) с использованием исследовательского оборудования (пакетов прикладных программ)
	ПК-3.3 Способен оценивать точность полученных экспериментальных (численных) результатов

3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны

знать:

- фундаментальные понятия и методы теории распознавания по прецедентам и анализа данных;
- современные проблемы анализа данных, теории распознавания, классификации, поиска зависимостей;
- методы и подходы решения практических задач анализа данных и классификации коллективами алгоритмов;
- программные средства решения основных задач анализа данных и классификации.

уметь:

 пользоваться своими знаниями для решения фундаментальных, прикладных и технологических задач в различных предметных областях;
 делать правильные выводы из сопоставления результатов теории и эксперимента, выбирать правильно параметры методов, адекватные размерности обучающих выборок;
 делать качественные и количественные выводы при переходе к предельным условиям в изучаемых проблемах;
 осваивать новые предметные области, теоретические подходы и экспериментальные методики;
 получать оптимальные алгоритмы классификации и правильно оценивать степень их точности и достоверности;
 работать на современном экспериментальном оборудовании;
 планировать оптимальное проведение обучения по прецедентам;
 эффективно использовать информационные технологии и компьютерную технику для достижения необходимых теоретических и прикладных результатов.

владеть:

 навыками анализа большого объема частично противоречивых и неполных признаковых описаний;
 навыками самостоятельной работы в лаборатории с использованием современных компьютерных технологий;
 культурой постановки и планирования последовательности решения задач анализа данных и классификации;
 навыками грамотной обработки статистических многомерных данных, оформления результатов численных расчетов и их сопоставления с теоретическими оценками;
 практикой исследования и решения теоретических и прикладных задач;
 навыками анализа реальных задач из различных предметных областей на уровне отдельных подходов и коллективами алгоритмов.

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. Модели распознавания, основанные на принципе частичной прецедентности.	5		5	13
2	Информативность признаков и эталонов, методы оценки информативности.	5		5	13
3	Логические закономерности классов, их поиск и применение в задачах классификации.	5		5	13
4	Модели распознавания, основанные на построении бинарных решающих деревьев.	5		5	13
5	Алгоритмы распознавания, основанные на построении линейных и кусочно-линейных разделяющих поверхностей	5		5	13
6	Модели распознавания, основанные на построении нелинейных разделяющих поверхностей	5		5	10
7	Нейросетевые модели классификации	4	4		

8	ROC-анализ и AUC- оптимальные классификаторы.	2	2		
9	Статистическая теория распознавания	2	2		
10	Алгебраическая теория распознавания	4	4		15
11	Система анализа данных и классификации РАСПОЗНАВАНИЕ	4	4		15
12	Кластерный анализ	4	4		15
13	Решение задач кластеризации коллективами алгоритмов	4	4		
14	Классификация объектов с неполными признаковыми описаниями, с большим числом классов	4	4		
15	Нахождение функциональных зависимостей по прецедентам	2	2		
Итого часов		60	30	30	120
Подготовка к экзамену		30 час.			
Общая трудоёмкость		270 час., 6 зач.ед.			

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

Семестр: 1 (Осенний)

1. Основные понятия. Модели распознавания, основанные на принципе частичной прецедентности.

Основные понятия теории распознавания по прецедентам. Признаковые описания, обучающие выборки, компактность, задачи распознавания, кластерного анализа, восстановления регрессий, прогнозирования, поиска закономерностей. Примеры практических применений. Стандартная обучающая информация. Функционал качества распознавания. Тестовый алгоритм, алгоритмы с представительными наборами. Модели алгоритмов вычисления оценок. Эффективные формулы вычисления оценок.

2. Информативность признаков и эталонов, методы оценки информативности.

Различные подходы и методы определения информативности признаков и эталонов. Вычисление оценок информативности. Поиск информативных систем признаков как дискретная оптимизационная задача. Приближенный метод нахождения оптимального признакового подпространства, основанный на применении логических корреляций признаков и методов кластеризации

3. Логические закономерности классов, их поиск и применение в задачах классификации.

Логические закономерности классов, логические описания классов, минимальные и сокращенные описания. Построение решающих функций в моделях голосования по системам логических закономерностей. Нахождение логических закономерностей классов как решение специализированных задач дискретной оптимизации. Поиск логических закономерностей классов с частотным и стандартным критериями качества.

Генетические алгоритмы поиска. Кроссовер, мутация, операторы отбора. Генетический алгоритм поиска логических закономерностей классов.

4. Модели распознавания, основанные на построении бинарных решающих деревьев.

Бинарные решающие деревья. Признаковые предикаты. Представление разбиения дискретного единичного куба в виде бинарного решающего дерева. Алгоритм построения допустимого разбиения. Алгоритмы построения бинарного решающего дерева по прецедентам, практические методы обрезания деревьев.

5. Алгоритмы распознавания, основанные на построении линейных и кусочно-линейных разделяющих поверхностей

Минимизация эмпирического риска. Правило постоянного приращения, теорема Новикова. Поиск максимальной совместной подсистемы системы линейных неравенств. Линейные и кусочно-линейные разделяющие поверхности. Линейная машина. Линейный дискриминант Фишера. Методы построения линейных разделяющих функций (релаксационные методы, псевдообращения, методы линейного программирования). Метод комитетов.

6. Модели распознавания, основанные на построении нелинейных разделяющих поверхностей

Построение полиномиальных разделяющих поверхностей, переход в спрямляющее пространство. Метод потенциальных функций, процедура обучения метода, метод группового учета аргументов. Метод опорных векторов. Сведение задачи построения разделяющей гиперплоскости с максимальным зазором к задаче квадратичного программирования. Случай линейной неразделимости классов. Метод опорных векторов и спрямляющее признаковое пространство. Связь метода опорных векторов и метода потенциальных функций.

Семестр: 2 (Весенний)

7. Нейросетевые модели классификации

Нейросетевые алгоритмы распознавания. Общие понятия. Алгоритм обратного распространения ошибки. Сети Кохонена и Хопфильда, алгоритмы обучения Хэбба, сети встречного распространения, мультипликативные нейронные сети, теорема Колмогорова.

8. ROC-анализ и AUC- оптимальные классификаторы.

Определение ROC-кривых как выбор оптимальных классификаторов. Определение таблицы сопряженности, точки отсечения, ошибки I и II рода, чувствительные и специфичные тесты. Практическое построение и анализ ROC-кривых в моделях классификации.

9. Статистическая теория распознавания

Байесовское решающее правило. Байесовский риск. Классификация с минимальным уровнем ошибок. Классификаторы, разделяющие функции и поверхности решений. Вероятности ошибок, случай нормальной плотности, махаланобисово расстояние, дискретный случай. Параметрические и непараметрические статистические методы распознавания. Функция роста, емкость множества функций. Равномерная сходимость частот ошибок к вероятностям. Примеры моделей распознавания ограниченной и неограниченной емкости.

10. Алгебраическая теория распознавания

Стандартный распознающий алгоритм, распознающий оператор, решающее правило. Основные понятия и определения алгебраического подхода в распознавании. Корректность и полнота моделей. Представление алгоритмов в виде операторных полиномов. Существование корректных алгоритмов. Методы поиска корректных алгоритмов. Операции над распознающими алгоритмами. Логические корректоры, корректор по большинству, байесовский и потенциальный корректоры алгоритмов

11. Система анализа данных и классификации РАСПОЗНАВАНИЕ

Описание графической оболочки. Главные окно и основное меню. Окно проекта. Методы распознавания и классификации. Ввод и предобработка данных, количественные признаки. Обработка номинальных признаков и неизвестных значений. Задание основного признака. Структура программы.

12. Кластерный анализ

Задача кластерного анализа. Меры подобия. Функции критериев для группировки: критерий суммы квадратов ошибок, родственные критерии минимума дисперсии. Матрицы и критерии рассеяния. Критерии кластеризации, основанные на матрицах рассеяния. Некоторые эвристические алгоритмы (метод к-средних, метод размытых к-средних, форель, метод к-эталонов, алгоритм взаимного поглощения). Задача кластеризации в статистической постановке. Восстановление плотностей компонент по плотности смеси. Итеративная оптимизация в кластерном анализе. Минимизация критерия суммы квадратов ошибок. Иерархическая группировка, дендрограммы, агломеративные и делимые процедуры. Алгоритмы "ближайший сосед", "дальний сосед", компромиссы. Пошаговая оптимальная иерархическая группировка. Многомерное масштабирование. Решение задачи кластеризации как поиск минимальных покрытий. Критерии качества кластеризаций, основанные на оценке устойчивости решений. Методы вычисления критериев. Меры концентрации, средняя мера внутриклассового рассеяния. Критерии кластеризации при неизвестном числе кластеров. Решение задач кластеризации при неизвестном числе кластеров

13. Решение задач кластеризации коллективами алгоритмов

Кластеризация коллективами алгоритмов. Комитетный синтез коллективных решений. Размытые и контрастные матрицы оценок. Критерии качества коллективных решений. Методы нахождения оптимальных коллективных решений задач кластерного анализа. Видео - логический метод кластеризации.

14. Классификация объектов с неполными признаковыми описаниями, с большим числом классов

Существующие методы восстановления значений признаков (marginalisation, imputation, регрессионные и статистические методы). Подходы, основанные на локальном обучении, оптимизации и применении алгоритмов распознавания. Достоинства и недостатки различных методов.

Существующие подходы для решения задач с многими классами. Подходы, основанные на попарном разделении классов, подход «один против всех». Сведение задачи к набору дихотомических классификаций и подходу ECOC.

15. Нахождение функциональных зависимостей по прецедентам

Задачи и методы восстановления регрессий, параметрические и непараметрические подходы (линейная и кусочно-линейная, полиномиальная, логистическая регрессии, ядерное сглаживание).

Восстановление функциональных зависимостей по прецедентам с использованием логических моделей распознавания. Байесовское восстановление, как построение коллективных решений задач распознавания. Восстановление кусочно-постоянных функций по прецедентам.

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

Учебная аудитория, оснащенная мультимедиа-проектором и экраном.

6.Перечень рекомендуемой литературы

Основная литература

1. Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л.Д. Прикладная статистика: классификация и снижение размерности. — М.: Финансы и статистика, 1989.

Дополнительная литература

1. Избранные научные труды [Текст]/Ю. И. Журавлев, -М., Магистр, 1998

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

http://www.machinelearning.org
http://www.machinelearning.ru
http://archive.ics.uci.edu/ml

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

http://www.machinelearning.org
http://www.machinelearning.ru
http://archive.ics.uci.edu/ml

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

Студент, изучающий курс методы анализа данных и распознавания, 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, методы доказательств.

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 учебной и научной литературе), подготовку ответов на вопросы, предназначенных для самостоятельного изучения, доказательство отдельных утверждений, свойств;

– подготовку к дифференцированному зачёту, экзамену.

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

Показателем владения материалом служит знание различных подходов, алгоритмов и методов, а также умение решать задачи. Для формирования умения применять теоретические знания на практике студенту необходимо решать как можно больше задач, в том числе прикладных задач классификации по прецедентам. При решении задач каждое действие необходимо аргументировать, ссылаясь на известные теоретические сведения.

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

ПРИЛОЖЕНИЕ

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

по направлению: Прикладные математика и физика
профиль подготовки: Общая и прикладная физика
Физтех-школа физики и исследований им. Ландау
кафедра информатики и вычислительной математики
курс: 1
квалификация: магистр

Семестры, формы промежуточной аттестации:
1 (осенний) - Дифференцированный зачет
2 (весенний) - Экзамен

Разработчик: В.В. Рязанов, д-р физ.-мат. наук, старший научный сотрудник, профессор

1. Компетенции, формируемые в процессе изучения дисциплины

Код и наименование компетенции	Индикаторы достижения компетенции
ОПК-3 Способен выбирать и (или) разрабатывать подходы к решению типовых и новых задач в области профессиональной деятельности, учитывая особенности и ограничения различных методов решения	ОПК-3.1 Способен анализировать задачу, планировать пути решения, предлагать и комбинировать способы решения
	ОПК-3.2 Способен использовать исследовательские методы при решении новых задач, применяя знания в различных областях науки (техники)
	ОПК-3.3 Владеет аналитическими и вычислительными методами решения, понимает и учитывает на практике границы применимости получаемых решений
ПК-3 Способен профессионально работать с исследовательским и испытательным оборудованием (приборами и установками, специализированными пакетами прикладных программ) в избранной предметной области	ПК-3.1 Понимает принципы работы используемого оборудования (специализированных пакетов прикладных программ)
	ПК-3.2 Способен проводить эксперимент (моделирование) с использованием исследовательского оборудования (пакетов прикладных программ)
	ПК-3.3 Способен оценивать точность полученных экспериментальных (численных) результатов

2. Показатели оценивания компетенций

В результате изучения дисциплины «Методы анализа данных и распознавания» обучающийся должен:

знать:

- фундаментальные понятия и методы теории распознавания по прецедентам и анализа данных;
- современные проблемы анализа данных, теории распознавания, классификации, поиска зависимостей;
- методы и подходы решения практических задач анализа данных и классификации коллективами алгоритмов;
- программные средства решения основных задач анализа данных и классификации.

уметь:

- пользоваться своими знаниями для решения фундаментальных, прикладных и технологических задач в различных предметных областях;
- делать правильные выводы из сопоставления результатов теории и эксперимента, выбирать правильно параметры методов, адекватные размерности обучающих выборок;
- делать качественные и количественные выводы при переходе к предельным условиям в изучаемых проблемах;
- осваивать новые предметные области, теоретические подходы и экспериментальные методики;
- получать оптимальные алгоритмы классификации и правильно оценивать степень их точности и достоверности;
- работать на современном экспериментальном оборудовании;
- планировать оптимальное проведение обучения по прецедентам;
- эффективно использовать информационные технологии и компьютерную технику для достижения необходимых теоретических и прикладных результатов.

владеть:

 навыками анализа большого объема частично противоречивых и неполных признаковых описаний;
 навыками самостоятельной работы в лаборатории с использованием современных компьютерных технологий;
 культурой постановки и планирования последовательности решения задач анализа данных и классификации;
 навыками грамотной обработки статистических многомерных данных, оформления результатов численных расчетов и их сопоставления с теоретическими оценками;
 практикой исследования и решения теоретических и прикладных задач;
 навыками анализа реальных задач из различных предметных областей на уровне отдельных подходов и коллективами алгоритмов.

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

3. Перечень типовых контрольных заданий, используемых для оценки знаний, умений, навыков

Промежуточная аттестация по дисциплине «**Методы анализа данных и распознавания**» осуществляется в форме дифференцированного зачета и экзамена. Зачет проводится в устной форме.

Перечень контрольных вопросов для сдачи дифференцированного зачета в 9-ом семестре.

1. Найти минимальный тест для таблицы обучения:

а)

0	0	0	1	
0	1	0	1	
1	0	0	0	K_1
1	1	1	0	
0	1	0	0	
1	0	0	1	K_2
1	1	0	1	
0	1	1	0	

б)

0	1	1	0	0	1	
0	1	1	0	1	1	K_1
1	0	1	0	0	1	
0	1	1	0	0	0	
1	1	1	0	1	1	K_2
0	1	1	1	0	1	

2. Найти минимальный представительный набор каждого класса для таблицы обучения:

а)

0	1	1	
1	0	0	K_1
1	1	0	
1	0	1	
1	1	1	K_2
0	0	1	

б)

1	1	1	
0	0	0	K_1
0	1	0	
1	0	0	
1	1	0	K_2
0	1	1	

3. Вычислить веса признаков x_1, x_3 (определяемые через тупиковые тесты) таблиц задачи 1.

4. Привести пример таблицы обучения с бинарными признаками, для которой вес признака x_1 равен нулю. Привести пример таблицы обучения с бинарными признаками, для которой вес признака x_1 равен единице.
5. Длина минимального теста меньше длины минимального представительного набора. Верно ли это утверждение?
6. Дана система линейных уравнений $Z: \sum_{j=1}^{n} a_{ij} x_j + b_i \le 0, i = 1,2,..., m.$ Сопоставимому неравенству булеву переменную $y_i \in \{0,1\}$, а всей системе - булеву функцию $f(y_1, y_2, ..., y_m)$ следующим образом. Пусть $Z(\bar{y})$ - подсистема системы Z, состоящая из всех неравенств, соответствующих единичным компонентам $\bar{y}$.

$$f(\bar{y}) = \begin{cases} 0, & \text{подсистема } Z(\bar{y}) \text{ совместна,} \\ 1, & \text{подсистема } Z(\bar{y}) \text{ несовместна.} \end{cases}$$

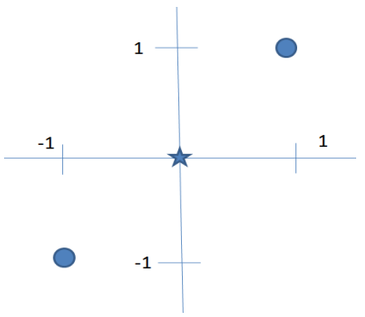
К какому классу булевых функций относится $f(\bar{y})$ и как она связана с задачей поиска максимальной совместной подсистемы системы Z?

7. Постройте решающее дерево для таблицы обучения а), б) задачи 1). Постройте решающее дерево для данных задачи 12.

8. Найти максимальную совместную подсистему системы линейных неравенств:

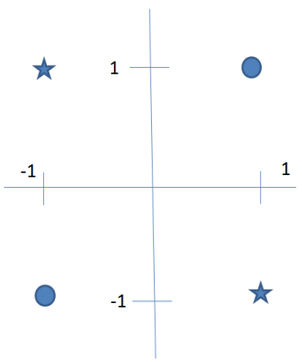
а) $\begin{cases} 2x+3y \geq 6 \\ 5x-2y \geq 10 \\ -x+y \geq -5 \\ -3x-y \geq 3 \end{cases}$ б) $\begin{cases} -x+y \geq 2 \\ x+2y \geq -2 \\ -3x-y \geq 6 \\ 2x-4y \geq 4 \end{cases}$

9. Построить нейронную сеть с линейными функциями состояний, разделяющую

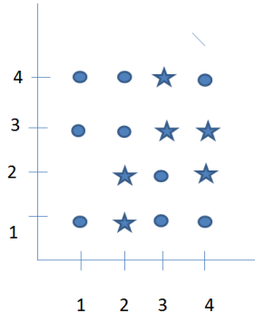


следующие два класса:

10. Построить нейронную сетьс линейными функциями состояний, разделяющую следующие два класса:



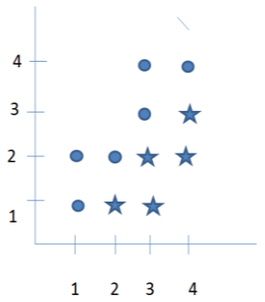
11. На рисунке приведены двумерные объекты обучения в случае двух классов. Привести список логических закономерностей классов со стандартным критерием



качества.

12. На рисунке приведены двумерные объекты обучения в случае двух классов.

Привести список логических закономерностей классов со стандартным критерием



качества.

13. а) Покажите, что расстояние между гиперплоскостью $g(\bar{x}) = \bar{w}^t \bar{x} + w_0 = 0$ и точкой $\bar{x}$ при наличии ограничения $g(\bar{x}_q) = 0$ может быть сделано равным $|g(\bar{x})| / \|\bar{w}\|$ путем минимизации $\|\bar{x} - \bar{x}_q\|^2$.

b) Покажите, что проекция вектора $\bar{x}$ на гиперплоскость задается выражением

$$\bar{x}_p = \bar{x} - \frac{g(\bar{x})}{\|\bar{w}\|^2} \bar{w}.$$

14. Рассмотрите линейную машину с разделяющими функциями $g_i(\bar{x}) = \bar{w}_i^t \bar{x} + w_{i0}, i = 1,2,...,l$. Покажите, что области решения являются выпуклыми.

15. Предложите обобщение на случай многих классов метода потенциальных функций, включающего l разделяющих функций; предложите итеративную процедуру коррекции ошибок для определения разделяющих функций.

16. Рассмотрим множество из семи двумерных векторов $\bar{x}_1^t = (1,0), \bar{x}_2^t = (0,1), \bar{x}_3^t = (0,-1), \bar{x}_4^t = (0,0), \bar{x}_5^t = (0,2), \bar{x}_6^t = (0,-2), \bar{x}_7^t = (-2,0)$. Допустим, что первые четыре имеют метку K_1, а другие три – метку K_2. Постройте полиномиальную разделяющую функцию минимальной степени.

17. Постройте для множества векторов предыдущей задачи гиперплоскость, разделяющую максимальное число объектов классов K_1 и K_2.

18. Пусть условные плотности для одномерной задачи и двух классов заданы распределением Коши: $p(x \mid K_i) = \frac{1}{\pi b} \cdot \frac{1}{1+\left(\frac{x-a_i}{b}\right)^2}, i = 1,2$.

Считая, что $P(K_1) = P(K_2)$, покажите, что $P(K_1 \mid x) = P(K_2 \mid x)$ при $x = (1/2)(a_1 + a_2)$.

Набросайте график $P(K_1 \mid x)$ для случая $a_1 = 3, a_2 = 5, b = 1$. Как ведет себя $P(K_1 \mid x)$ при $x \to -\infty ? + \infty ?$

19. Пусть $\bar{x}$ есть бинарный (0,1) вектор с многомерным распределением Бернулли $P(\bar{x} \mid \bar{\theta}) = \prod_{i=1}^{n} \theta_i^{x_i} (1-\theta_i)^{1-x_i}$, где $\bar{\theta} = (\theta_1, \theta_2, ..., \theta_n)^t$ - неизвестный параметрический вектор, а θ_i - есть вероятность того, что $x_i = 1$. Покажите, что оценка по максимуму правдоподобия для $\bar{\theta}$ есть $\hat{\theta} = \frac{1}{m} \sum_{i=1}^{m} \bar{x}_i$.

20. Пусть выборочное среднее $\bar{m}_n$ и выборочная ковариационная матрица C_n для множества n выборок $\bar{x}_1, \bar{x}_2, ..., \bar{x}_n$ определяются выражениями $\bar{m}_n = \frac{1}{n} \sum_{k=1}^{n} \bar{x}_k$, и $C_n = \frac{1}{n-1} \sum_{k=1}^{n} (\bar{x}_k - \bar{m}_n)(\bar{x}_k - \bar{m}_n)^t$.

Покажите, что влияние на эти величины добавления новой выборки $\bar{x}_{n+1}$ можно выразить рекуррентными формулами $\bar{m}_{n+1} = \bar{m}_n + \frac{1}{n+1}(\bar{x}_{n+1} - \bar{m}_n)$ и

$$C_{n+1} = \frac{n-1}{n} C_n + \frac{1}{n+1} (\bar{x}_{n+1} - \bar{m}_n)(\bar{x}_{n+1} - \bar{m}_n)^t .$$

21. Выражения из задачи 20 позволяют вводить поправки и оценки ковариационных матриц. Однако нередко представляет интерес обратная ковариационная матрица, а ее обращение отнимает много времени. Доказав матричное тождество $(A + \bar{x}\bar{x}^t)^{-1} = A^{-1} - \frac{A^{-1}\bar{x}\bar{x}^t A^{-1}}{1 + \bar{x}^t A^{-1} x}$ покажите, что

$$C_{n+1}^{-1} = \frac{n}{n-1} \left[C_n^{-1} - \frac{C_n^{-1}(\bar{x}_{n+1} - \bar{m}_n)(\bar{x}_{n+1} - \bar{m}_n)^t C_n^{-1}}{\frac{n^2-1}{n} + (\bar{x}_{n+1} - \bar{m}_n)^t C_n^{-1} (\bar{x}_{n+1} - \bar{m}_n)} \right].$$

22. Рассмотрим множество из семи двумерных векторов $\bar{x}_1^t = (1,0), \bar{x}_2^t = (0,1), \bar{x}_3^t = (0,-1), \bar{x}_4^t = (0,0), \bar{x}_5^t = (0,2), \bar{x}_6^t = (0,-2), \bar{x}_7^t = (-2,0)$. Допустим, что первые три имеют метку K_1, а другие четыре – метку K_2.

а) Нарисуйте границу областей решений, полученную в результате применения правила ближайшего соседа. (Она должна состоять из девяти отрезков прямых.)

b) Найдите средние значения выборок $\overline{m}_1$ и $\overline{m}_2$ и нарисуйте границу решения, соответствующую классификации $\bar{x}$ при присвоении ему класса среднего значения ближайшей выборки.

23. Пользуясь определением матрицы разброса между группами, данным для случая многих классов: $S_B = \sum_{i=1}^{l} n_i(\overline{m}_i - \overline{m})(\overline{m}_i - \overline{m})^t$, покажите, что $S_B = [(n_1 n_2)/n](\overline{m}_1 - \overline{m}_2)(\overline{m}_1 - \overline{m}_2)^t$, если $l = 2$.
24. Постановка задачи распознавания по прецедентам. Признаковые описания и виды признаков. Задачи классификации с учителем (распознавания), классификации без учителя (кластерного анализа) и восстановления регрессий как специальные задачи интерполяции.
25. Задачи анализа данных и поиска закономерностей. Примеры практических применений.
26. Тупиковые тесты к-значных таблиц обучения. Сведение задачи поиска тупиковых тестов к поиску неприводимых покрытий бинарной матрицы. Существование тупиковых тестов.
27. Информационные веса признаков, основанные на вычислении тупиковых тестов. Тестовый алгоритм распознавания.
28. Алгоритмы распознавания, основанные на нахождении и голосовании по системам представительных наборов классов. Алгоритм распознавания «Кора».
29. Стохастический аналог тестового алгоритма. Модификация алгоритмов частичнойпрецедентностидля случаеввещественнозначных и смешанных систем признаков.
30. Алгоритмы распознавания, основанные на вычислении оценок (АВО). Опорные множества алгоритмов, основные этапы вычисления оценок, примеры моделей АВО.
31. Эффективные формулы вычисления оценок для систем опорных множеств фиксированной мощности и всех опорных множеств.
32. Оптимизация многопараметрических моделей распознавания.
33. Релаксационный алгоритм решения систем линейных неравенств и методы его ускорения. Применение релаксационного алгоритма для приближенного поиска максимальной совместной подсистемы системы линейных неравенств.
34. Комбинаторный алгоритм поиска максимальной совместной подсистемы системы линейных неравенств.
35. Информативность признаков и эталонов. Методы оценки информативности в моделях частичной прецедентности. Статистические, информационные, эвристические критерии информативности.
36. Поиск информативных систем признаков как задача дискретной оптимизации. Приближенный алгоритм поиска минимального признакового пространства, основанный на вычислении логических корреляций признаков, оценке информативности признаков и кластеризации.
37. Логические закономерности классов. Определения частичной логической закономерности, интервала логической закономерности. Построение минимальной логической закономерности минимальной сложности эквивалентной заданной минимальной логической закономерности.
38. Логические описания классов, минимальные и сокращенные описания.
39. Методы обработки множеств логических закономерностей классов.
40. Построение решающих функций в моделях голосования по системам логических закономерностей. Сглаживание решающих функций, построение решающих функций с максимальным «зазором» между обучающими объектами разных классов.

41. Нахождение логических закономерностей классов с частотным критерием качества как решение линейных задач дискретной оптимизации со специальными свойствами монотонности матриц коэффициентов ограничений и целевой функции.
42. Поиск логических закономерностей классов со стандартным критерием качества. Сведение данной задачи к специальным задачам поиска максимальных совместных подсистем систем линейных неравенств при линейных ограничениях относительно бинарных параметров.
43. Генетические алгоритмы поиска оптимальных решений. Операторы «кроссовер», мутация, отбора. Функции кодирования/декодирования подмножеств эталонов классов, функции оценки приспосабливаемости.
44. Генетический алгоритм поиска логических закономерностей классов.
45. Бинарные решающие деревья. Признаковые предикаты. Существование бинарного решающего дерева для заданной обучающей выборки. Представление разбиения дискретного единичного куба в виде бинарного решающего дерева. Алгоритм построения допустимого разбиения.
46. Практические методы построения бинарных решающих деревьев, способы обрезания деревьев.
47. Правило постоянного приращения, теорема Новикова. Доказательство конечности правила постоянного приращения для линейно разделимых классов.
48. Линейные и кусочно-линейные разделяющие поверхности. Линейная машина. Методы построения линейных разделяющих функций (релаксационные методы, методы линейного программирования). Метод комитетов.
49. Метод наименьших квадратов и псевдообращения матриц.
50. Линейный дискриминант Фишера. Определение линейного дискриминанта Фишера и вычисление. Случаи неоднозначности решений и вырожденности. Практические подходы по применению дискриминанта Фишера в случаях вырожденности.
51. Алгоритм «k -ближайших соседей», его ограничения и интерпретация.
52. Метод потенциальных функций и обучение алгоритма.
53. Метод группового учета аргументов, общая схема построения признакового пространства в виде полиномов от исходных признаков.
54. Универсальная система «РАСПОЗНАВАНИЕ» для интеллектуального анализа данных, классификации и прогнозирования: назначение, основные характеристики и функции, интерфейс.
55. Оценка точности алгоритмов распознавания по обучающей выборке в режиме скользящего контроля.
56. Метод опорных векторов. Сведение задачи построения разделяющей гиперплоскости с максимальным зазором к задаче квадратичного программирования.
57. Метод опорных векторов в случае линейной неразделимости классов. Модификация основной оптимизационной задачи метода и ее сведение к задаче квадратичного программирования.
58. Метод опорных векторов и спрямляющее признаковое пространство. Связь метода опорных векторов и метода потенциальных функций.

Перечень контрольных вопросов для сдачи экзамена
в 10-ом семестре.

1. Используя условные плотности $p(x \mid K_i) = \frac{1}{\pi b} \frac{1}{1+\left(\frac{x-a_i}{b}\right)^2}, i = 1,2$ для одномерной задачи и двух классов и полагая априорные вероятности равными, покажите, что

минимальная вероятность ошибки определяется выражением

$$P(\textit{ошибка}) = \frac{1}{2} - \frac{1}{\pi} ctg \left| \frac{a_2 - a_1}{2b} \right|.$$

Рассмотрите это выражение как функцию величины $\left| \frac{a_2 - a_1}{b} \right|$.

2. Пусть $p(x \mid K_i) \sim N(\mu_i, \sigma^2)$ для одномерной задачи и двух классов при $P(K_1) = P(K_2) = 1/2$. Покажите, что минимальная вероятность ошибки определяется выражением $P_e = \frac{1}{\sqrt{2\pi}} \int_a^\infty e^{-(1/2)u^2} du$,

   Где $a = |\mu_2 - \mu_1| / 2\sigma$. Рассмотрите неравенство $\frac{1}{\sqrt{2\pi}} \int_a^\infty e^{-(1/2)u^2} du \le \frac{1}{\sqrt{2\pi}a} e^{-(1/2)a^2}$.

   Покажите, что P_e стремится к нулю при $|\mu_2 - \mu_1| / \sigma \to \infty$.

3. Пусть $p(\bar{x} \mid K_i) \sim N(\bar{\mu}_i, \sigma^2 I)$ для n-мерной задачи и двух классов при $P(K_1) = P(K_2) = 1/2$. Покажите, что минимальная вероятность ошибки определяется выражением

   $$P_e = \frac{1}{\sqrt{2\pi}} \int_a^\infty e^{-(1/2)u^2} du,$$

   где $a = \|\bar{\mu}_2 - \bar{\mu}_1\| / 2\sigma$. Пусть $\bar{\mu}_1 = 0$ и $\bar{\mu}_2 = (\mu, \mu, ..., \mu)^t$. Используя неравенство из задачи 2, покажите, что P_e стремится к нулю при $n \to \infty$. Выразите смысл этого результата словами.

4. Пусть компоненты вектора $\bar{x} = (x_1, x_2, ..., x_n)^t$ тернарны (1, 0 или -1) с вероятностями $p_{ij} = \Pr(x_i = 1 \mid K_j)$, $q_{ij} = \Pr(x_i = 0 \mid K_j)$, $r_{ij} = \Pr(x_i = -1 \mid K_j)$, причем компоненты x_i статистически независимы для всех $\bar{x}$ и K_j.
   Покажите, что можно получить решающее правило с минимальной вероятностью ошибки, используя разделяющие функции $g_j(\bar{x})$, представляющие собой квадратичные функции компонент x_i.

5. Если множество векторов $X = \{\bar{x}_1, \bar{x}_2, ..., \bar{x}_m\}$ разделено на l групп $K_1, K_2, ..., K_l$, то среднее $\bar{m}_i$ для K_i не определено, если K_i пустое. В этом случае сумма квадратов ошибок содержит только непустые подмножества $J_e = \sum_{\substack{\text{непустое} \\ K_i}} \sum_{\bar{x} \in K_i} \|\bar{x} - \bar{m}_i\|^2$. Считая, что $m \ge l$, покажите, что в разделении, минимизирующем J_e, нет пустых подмножеств.

6. Рассмотрим иерархическую процедуру группировки, в которой группы объединяются так, чтобы привести к наименьшему увеличению суммы квадратов ошибок на каждом шаге. Если группа K_i содержит n_i объектов со средним $\bar{m}_i$,

покажите, что наименьшее увеличение получается от слияния двух групп, для которых $\frac{n_i n_j}{n_i + n_j}\left\|\overline{m}_i - \overline{m}_j\right\|^2$ минимальна.

7. Постройте и сравните результаты иерархической кластеризации на 4, 3, 2 кластера для выборки из 5 объектов , описываемых единственным признаком $x_1 = 0, x_2 = 1, x_3 = 3, x_4 = 4, x_5 = 7$, и следующих мерах расстояния:

   $d_{\min}(K_i, K_j) = \min_{\overline{x} \in K_i, \overline{x}' \in K_j} \left|\overline{x} - \overline{x}'\right|,$

   $d_{\max}(K_i, K_j) = \max_{\overline{x} \in K_i, \overline{x}' \in K_j} \left|\overline{x} - \overline{x}'\right|$

   $d_{mean}(K_i, K_j) = \left|\overline{m}_i - \overline{m}_j\right|.$

8. Решите задачу иерархической группировки (при указанных в задаче 7 мерах расстояния) на два кластера для выборкиодномерных объектов:
   $x_1 = 0, x_2 = 0.5, x_3 = 0.7, x_4 = 1, x_5 = 1.1, x_6 = 1.5.$
9. Приведите пример совпадающих размытой и контрастной матриц оценок.
10. Показать, что мера концентрации кластеризации $\mathrm{K} = \{K_1, K_2, ...K_l\}$ имеет вид

    $$Z_1(\mathrm{K}) = \frac{1}{m^2}\sum_{i=1}^{l} n_i^2$$

11. Показать, что мера концентрации кластеризации $\mathrm{K} = \{K_1, K_2, ...K_l\}$ имеет вид

    $$Z_{-1}(\mathrm{K}) = \frac{1}{l}$$

12. Показать, что критерий кластеризации $J_T = \sum_{i=1}^{l}\sum_{\overline{x} \in K_i} (\overline{x} - \overline{m}_i)^t S_T^{-1} (\overline{x} - \overline{m}_i)$ (где S_T - общая матрица рассеяния) инвариантен к невырожденным линейным преобразованиям данных.

13. Нейросетевые алгоритмы распознавания. Общие понятия. Активационные функции. Алгоритм обратного распространения ошибки.
14. Мультипликативные нейронные сети и их обучение.
15. Теорема Колмогорова.
16. ROC-анализ и AUC- оптимальные классификаторы.
17. Байесовское решающее правило. Байесовский риск. Классификация с минимальным уровнем ошибок.
18. Классификаторы сминимальной ошибкой, разделяющие функции и поверхности решений.
19. Вероятности ошибок, случай нормальной плотности. Байесовский классификатор с минимальной ошибкой для нормально распределенных классов.
20. Функция роста, емкость множества функций. Равномерная сходимость частот ошибок к вероятностям при конечном числе функций.
21. Равномерная сходимость частот ошибок к вероятностям ошибок при конечной емкости множества функций.
22. Примеры моделей распознавания ограниченной и неограниченной емкости.
23. Стандартный распознающий алгоритм, распознающий оператор, решающее правило. Основные понятия и определения алгебраического подхода в распознавании.

24. Корректность и полнота моделей распознавания. Представление алгоритмов в виде операторных полиномов. Существование корректных алгоритмов.
25. Методы поиска корректных и квазикорректныхалгоритмов.
26. Операции над распознающими алгоритмами. Логические корректоры, корректор по большинству, байесовский и потенциальный корректоры алгоритмов.
27. Эвристические корректоры над распознающими алгоритмами (комитетные, области компетенции, шаблоны принятия решений, динамический метод)
28. Задача кластерного анализа. Меры подобия. Функции критериев для группировки: критерий суммы квадратов ошибок, родственные критерии минимума дисперсии.
29. Матрицы и критерии рассеяния. Критерии кластеризации, основанные на матрицах рассеяния.
30. Эвристические алгоритмы (метод k -средних, форель, метод k -эталонов, алгоритм взаимного поглощения).
31. Задача кластеризации в статистической постановке. Восстановление плотностей компонент по плотности смеси.
32. Итеративная оптимизация в кластерном анализе. Минимизация критерия суммы квадратов ошибок.
33. Дисперсионный критерий при различных весах объектов обучения, минимизация критерия.
34. Иерархическая группировка, дендрограммы, агломеративные и делимые процедуры. Алгоритмы "ближайший сосед", "дальний сосед", компромиссы.
35. Пошаговая оптимальная иерархическая группировка. Многомерное масштабирование.
36. Решение задачи кластеризации как поиск минимальных покрытий.
37. Нейросетевые схемы самообучения. Сети Кохонена и Хопфильда, алгоритмы обучения Хэбба, сети встречного распространения.
38. Критерии качества кластеризаций, основанные на оценке устойчивости решений.
39. Методы вычисления критериев качества при минимизации дисперсионного критерия, в иерархической группировке и в алгоритме «k - внутригрупповых средних».
40. Метод нечетких k - внутригрупповых средних.
41. Кластеризация множеств логических закономерностей классов.
42. Меры концентрации, средняя мера внутриклассового рассеяния. Критерии кластеризации при неизвестном числе кластеров. Решение задач кластеризации при неизвестном числе кластеров.
43. Кластеризация коллективами алгоритмов. Комитетный синтез коллективных решений.
44. Размытые и контрастные матрицы оценок. Критерии качества коллективных решений.
45. Эквивалентность задач максимизации расстояния хэмминга от матрицы оценок коллективного решения до средней размытой матрицы и минимизации расстояния хэмминга от матрицы оценок коллективного решения до множества всех контрастных матриц.
46. Комитетный синтез оптимальных коллективных решений задачи кластерного анализа. Эвристические методы нахождения оптимальных коллективных решений задач кластерного анализа. Видео - логический метод кластеризации.
47. Задачи распознавания и кластеризации при неполноте данных. Методы решения задач классификации неполных данных.
48. Алгоритмы восстановления неизвестных значений признаков, основанные на локальном обучении, оптимизации и применении методов распознавания.
49. Задачи и методы восстановления регрессий, параметрические и непараметрические подходы (линейная, полиномиальная, ядерное сглаживание).

50. Нахождение кусочно-линейных регрессий на базе динамического программирования и моделей распознавания.
51. Логистическая регрессия и ее нахождение.
52. Восстановление функциональных зависимостей по прецедентам с использованием моделей распознавания.
53. Байесовское восстановление, как построение коллективных решений задач распознавания. Корректность байесовского алгоритма восстановления зависимости при базовых корректных алгоритмах распознавания.
54. Линейный корректор и его корректность при базовых корректных алгоритмах распознавания.
55. Восстановление кусочно-постоянных функций по прецедентам. Поиск оптимального числа компонент.
56. Эффективное переобучение в режиме скользящего контроля в модели вычисления оценок для смежной по обучающей выборке задачи распознавания.
57. Восстановление кусочно-линейных функций в евклидовом пространстве, основанное на объединении метода наименьших квадратов, динамического программирования и модели распознавания.

Примеры билетов

Билет №1

1. Мультипликативные нейронные сети и их обучение.

2. Восстановление кусочно-постоянных функций по прецедентам.

Билет №2

1. Теорема Колмогорова.

2. Байесовское восстановление, как построение коллективных решений задач распознавания.

4. Критерии оценивания

Оценка «отлично (10)» выставляется студенту, показавшему всесторонние,систематизированные, глубокие знания учебной программы дисциплины и умениеуверенно применять их на практике при решении конкретных задач, свободное иправильное обоснование принятых решений;

Оценка «отлично (9)» выставляется студенту, показавшему систематизированные, глубокие знания учебной программы дисциплины и умениеприменять их на практике при решении конкретных задач, свободное иправильное обоснование принятых решений;

Оценка «отлично (8)» выставляется студенту, показавшему систематизированные, знания учебной программы дисциплины и умениеприменять их на практике при решении конкретных задач, правильное обоснование принятых решений;

Оценка «хорошо (7)» выставляется студенту, если он твердо знает материал, грамотно и посуществу излагает его, умеет применять полученные знания на практике, нодопускает в ответе или в решении задач некоторые неточности;

Оценка «хорошо (6)» выставляется студенту, если он твердо знает материал, грамотно излагает его, умеет применять полученные знания на практике, нодопускает в ответе или в решении задач некоторые неточности;

Оценка «хорошо (5)» выставляется студенту, если он знает материал, грамотно излагает его, умеет применять полученные знания на практике, нодопускает в ответе или в решении задач некоторые неточности;

Оценка «удовлетворительно (4)» выставляется студенту, показавшему фрагментарный,разрозненный характер знаний, недостаточно правильные формулировки базовыхпонятий, нарушения логической последовательности в изложении программногоматериала, но при этом он владеет основными разделами учебной программы,необходимыми для дальнейшего обучения и может применять полученные знания пообразцу в стандартной ситуации;

Оценка «удовлетворительно (3)» выставляется студенту, показавшему фрагментарный характер знаний, недостаточно правильные формулировки базовыхпонятий, но при этом он владеет основными разделами учебной программы,необходимыми для дальнейшего обучения и может применять полученные знания пообразцу в стандартной ситуации;

Оценка «неудовлетворительно (2)» выставляется студенту, который не знает большей частиосновного содержания учебной программы дисциплины, допускает грубые ошибки вформулировках основных понятий дисциплины и не умеет использоватьполученные знания при решении типовых практических задач.

Оценка «неудовлетворительно (1)» выставляется студенту, который не знает основного содержания учебной программы дисциплины, допускает грубые ошибки вформулировках основных понятий дисциплины и не умеет использоватьполученные знания при решении типовых практических задач.

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

Во время проведения зачета и экзамена обучающиеся могут пользоваться программой дисциплины, Интернетом, справочной литературой, вычислительной техникой.

Зачет и экзамен проводятся путем организации специального опроса, проводимого в устной форме.