МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение

высшего профессионального образования

«Московский физико-технический институт (государственный университет)»

МФТИ

«УТВЕРЖДАЮ»

Проректор по учебной и методической работе

_____Д.А. Зубцов

« »_____20 г.

Рабочая программа дисциплины (модуля)

по дисциплине: МЕТОДЫ АНАЛИЗА ДАННЫХ И РАСПОЗНАВАНИЯ

по направлению: Прикладные математика и физика (магистратура)

профиль подготовки/

Компьютерные технологии и интеллектуальный анализ данных

факультет: управления и прикладной математики

кафедра: информатики

курс: 5

квалификация: магистр

Семестр, формы промежуточной аттестации: 9(Осенний) - Дифференцированный зачет

Аудиторных часов: 100 всего, в том числе:

лекции: 34 час.

практические (семинарские) занятия: 0 час.

лабораторные занятия: 66 час.

Самостоятельная работа: 34 час., в том числе:

задания, курсовые работы: 0 час.

Подготовка к экзамену: 30 час.

Всего часов: 100, всего зач. ед.: 5

Программу составили:

В.В.Рязанов, доктор физико-математических наук, профессор

Программа обсуждена на заседании кафедры

_____ 2014 г.

СОГЛАСОВАНО:

Заведующий кафедрой И.Б.Петров

Декан факультета управления и прикладной математики А.А. Шананин

Начальник учебного управления И.Р. Гарайшина

1. Цели и задачи

Цель дисциплины

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

Задачи дисциплины

Задачами данного курса являются:

- 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. Место дисциплины (модуля) в структуре образовательной программы магистратуры

Дисциплина «Методы анализа данных и распознавания» включает в себя разделы, которые могут быть отнесены к циклу __М.2__ (шифр цикла).

Дисциплина «Методы анализа данных и распознавания» базируется на дисциплинах и предшествует им:
базовая и вариативная часть кода УЦ ООП Б.2 **(**математический естественнонаучный блок) по дисциплинам «Высшая математика» (математический анализ, высшая алгебра, дифференциальные уравнения и методы математической физики), блока «Общая физика» и региональной составляющей этого блока и относится к профессиональному циклу.

3. Перечень планируемых результатов обучения по дисциплине (модулю), соотнесенных с планируемыми результатами освоения образовательной программы

Освоение дисциплины направлено на формирование следующих общекультурных, общепрофессиональных и профессиональных компетенций магистра:

способность применять теорию и методы математики для построения качественных и количественных моделей объектов и процессов в естественнонаучной сфере деятельности (ОПК-2);
понимать ключевые аспекты и концепции в области их специализации (ОПК-3);
способность выбирать и применять подходящее оборудование, инструменты и методы исследований для решения задач в избранной предметной области (ПК-3);
способность критически оценивать применимость применяемых методик и методов (ПК-4);

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	Виды учебных занятий, включая самостоятельную работу			
№	Тема (раздел) дисциплины	Лекции	Практич. (семинар.) задания	Лаборат. работы	Задания, курсовые работы
1	Основные понятия. Модели распознавания, основанные на принципе частичной прецедентности.	2		6	
2	Информативность признаков и эталонов, методы оценки информативности.	2			
3	Логические закономерности классов, их поиск и применение в задачах классификации.	4		6	
4	Модели распознавания, основанные на построении бинарных решающих деревьев.	3		6	
5	Алгоритмы распознавания, основанные на построении линейных и кусочно-линейных разделяющих поверхностей	4		4	
6	Модели распознавания, основанные на построении нелинейных разделяющих поверхностей	6		4	
7	Нейросетевые модели классификации	2		4	
8	ROC-анализ и AUC- оптимальные классификаторы.			4	
9	Статистическая теория распознавания	3			
10	Алгебраическая теория распознавания	2		8	
11	Система анализа данных и классификации РАСПОЗНАВАНИЕ			8	
12	Кластерный анализ	2		8	
13	Решение задач кластеризации коллективами алгоритмов	2		4	
14	Классификация объектов с неполными признаковыми описаниями, с большим числом классов			4	
15	Нахождение функциональных зависимостей по прецедентам	2			
Итого часов		34		66	
Общая трудоёмкость		164 час., 5 зач.ед.			

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

Семестр: 1 (Осенний)

1. ПОДХОДЫ И МЕТОДЫ ТЕОРИИ РАСПОЗНАВАНИЯ ПО ПРЕЦЕДЕНТАМ

1. Основные понятия теории распознавания по прецедентам. 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.
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

I. ПОДХОДЫ И МЕТОДЫ КЛАСТЕРНОГО АНАЛИЗА

11. 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.

12. 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.

13. 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.

14. 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.

III. ПОИСК ФУНКЦИОНАЛЬНЫХ ЗАВИСИМОСТЕЙ

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 основной и дополнительной литературы, необходимой для освоения дисциплины (модуля)

Основная литература

1. Журавлев Ю.И. Избранные научные труды. – М.: "Магистр", 1998, 420 с.
2. А.С.Бирюков, В.В.Рязанов, А.С.Шмаков. Решение задач кластерного анализа коллективами алгоритмов. Журнал вычислительной математики и математической физики, Т.48, 2008, N 1, стр. 176-192.
3. Журавлев Ю. И., Рязанов В. В., Сенько О. В. Распознавание». Математические методы. Программная система. Практические применения. — М.: Фазис, 2006.
4. Н.В.Ковшов, В.Л.Моисеев, В.В.Рязанов. Алгоритмы поиска логических закономерностей в задачах распознавания. Журнал вычислительной математики и математической физики, Т.48, 2008, N 2, стр. 329-344.
5. В. В. Рязанов, Ю. И. Ткачев, Восстановление зависимостей на основе байесовской коррекции коллективов алгоритмов классификации// Журнал вычислительной математики и математической физики, Vol. 50, No. 9, 2010.
6. Hastie, T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. — 2nd ed. — Springer-Verlag, 2009. — 746 p.

Дополнительная 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.
2. Vapnik, V. :The Nature of Statistical Learning Theory. Springer-Verlag (1995).
3. Ежегодник "Распознавание, классификация, прогноз (математические методы и их применение)", – М.: "Наука", Вып.1 (1988), 2 (1989), 3 (1990).
4. Журавлев Ю.И. Избранные научные труды. – М.: "Магистр", 1998, 420 с.
5. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. Новосибирск, Институт математики им. С.Л.Соболева СО РАН, 1999, 268 с.
6. Duda R.O., Hart P.E., Stork D.G. Pattern classification, 2nd Edition. Wiley-Interscience, 2001. - 738 pages.

7. Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине (модулю)

1. Little,R.J.A.,Rubin D.B.: Statistical Analysis with Missing Data. Wiley, New York (1987)
2. Hardle, W.: Applied nonparametric regression. Cambridge University Press, Cambridge (1990)
3. Донской В.И. , Башта А.И. Дискретные модели принятия решений при неполной информации. –Симферополь: Таврия, 1992. – 166 с.
4. Дюк В., Самойленко А., Data Mining: учебный курс – СПб: Питер, 2001. – 368 с.
5. Уоссермен Ф., Нейрокомпьютерная техника, М.,Мир, 1992.

8. 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

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

На лекционных и лабораторных занятиях используются мультимедийные технологии, включая демонстрацию презентаций.

10. Методические указания для обучающихся по освоению дисциплины

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 конспектам лекций, учебной и научной литературе), подготовку ответов на вопросы, предназначенных для самостоятельного изучения, доказательство отдельных утверждений, свойств;
– решение задач, предлагаемых студентам на практических занятиях,
– подготовку к практическим занятиям, зачёту.
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 лектору или преподавателю, ведущему практические занятия.

11. Фонд оценочных средств для проведения промежуточной аттестации по итогам обучения

Приложение