Правительство Российской Федерации

Федеральное государственное автономное образовательное учреждение высшего профессионального образования
«Национальный исследовательский университет
«Высшая школа экономики»

Факультет Компьютерных наук
Департамент больших данных и информационного поиска
Базовая кафедра Яндекс

УТВЕРЖДАЮ
Академический руководитель
образовательной программы
«Науки о данных»
по направлению 01.04.02
«Прикладная математика и информатика»
С.О. Кузнецов

_____
«__» _____ 2015 г.

Программа дисциплины «Машинное обучение»
для направления 01.04.02 "Прикладная математика и информатика" подготовки магистра
для магистерской программы "Науки о данных"

Автор программы:
Воронцов К.В., д.т.н. (vokov@forecsys.ru)

Одобрена на заседании базовой кафедры Яндекс « » 2015 г.

Заведующий кафедрой М.А. Бабенко

Рекомендована Академическим советом образовательной программы
«Науки о данных» « » 2015 г.

Менеджер базовой кафедры Яндекс И.И. Алескерова

Москва, 2015

Пояснительная записка

Автор программы

Воронцов К.В., д.т.н.

Требования к студентам

От студентов требуются знания курсов линейной алгебры, математического анализа, теории вероятностей. Знание математической статистики, методов оптимизации и какого-либо языка программирования желательно, но не обязательно.

Аннотация

Дисциплина «Машинное обучение» предназначена для подготовки магистров 01.04.02 – Прикладная математика и информатика.

Теория обучения машин (machine learning, машинное обучение) находится на стыке прикладной статистики, численных методов оптимизации, дискретного анализа, и за последние 50 лет оформилась в самостоятельную математическую дисциплину. Методы машинного обучения составляют основу ещё более молодой дисциплины — интеллектуального анализа данных (data mining).

В курсе рассматриваются основные задачи обучения по прецедентам: классификация, кластеризация, регрессия, понижение размерности. Изучаются методы их решения, как классические, так и новые, созданные за последние 10–15 лет. Упор делается на глубокое понимание математических основ, взаимосвязей, достоинств и ограничений рассматриваемых методов. Отдельные теоремы приводятся с доказательствами.

Все методы излагаются по единой схеме:

- исходные идеи и эвристики;
- их формализация и математическая теория;
- описание алгоритма в виде слабо формализованного псевдокода;
- анализ достоинств, недостатков и границ применимости;
- пути устранения недостатков;
- сравнение с другими методами.
- примеры прикладных задач.

Данный курс расширяет и углубляет набор тем, рекомендованный международным стандартом ACM/IEEE Computing Curricula 2001 по дисциплине «Машинное обучение и нейронные сети» (machine learning and neural networks) в разделе «Интеллектуальные системы» (intelligent systems).

Программа курса предусматривает лекции (40 часов) и практические занятия (34 часа).

Учебные задачи курса

Целью данного курса является изучение основ теории обучения машин, включая дискриминантный, кластерный и регрессионный анализ, овладение навыками практического решения задач интеллектуального анализа данных.

Тематический план дисциплины «Машинное обучение»

№	Название темы	Всего часов по дисциплине	Аудиторные часы		Самостоятельная работа
			Лекции	Сем. и практика	
1	Основные понятия и примеры прикладных задач	38	6	4	24
2	Метрические методы классификации	40	6	6	24
3	Логические методы классификации	46	6	6	36
4	Линейные методы классификации	48	8	6	36
5	Методы регрессионного анализа	48	6	6	36
6	Байесовские методы классификации	46	8	6	36
	Итого	266	40	34	192

I. Источники информации

Список литературы

Основная литература

1. Айвазян С. А., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: основы моделирования и первичная обработка данных. — М.: Финансы и статистика, 1983
2. Айвазян С. А., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: исследование зависимостей. — М.: Финансы и статистика, 1985
3. Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: классификация и снижение размерности. — М.: Финансы и статистика, 1989
4. Вапник В. Н., Червоненкис А. Я. Теория распознавания образов. — М.: Наука, 1974.
5. Вапник В. Н. Восстановление зависимостей по эмпирическим данным. — М.: Наука, 1979
6. Журавлев Ю. И., Рязанов В. В., Сенько О. В. «Распознавание». Математические методы. Программная система. Практические применения. — М.: Фазис, 2006. ISBN 5-7036-0108-8
7. Загоруйко Н. Г. Прикладные методы анализа данных и знаний. — Новосибирск: ИМ СО РАН, 1999. ISBN 5-86134-060-9

Дополнительная литература

1. Шлезингер М., Главач В. Десять лекций по статистическому и структурному распознаванию. — Киев: Наукова думка, 2004. ISBN 966-00-0341-2
2. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. — Springer, 2001. ISBN 0-387-95284-5
3. MacKay D. On-line book: Information Theory, Inference, and Learning Algorithms. — 2005
4. Mitchell T. Machine Learning. — McGraw-Hill Science/Engineering/Math, 1997. ISBN 0-07-042807-7
5. Schölkopf B., Smola A.J. Learning with Kernels. Support Vector Machines, Regularization, Optimization, and Beyond. — MIT Press, Cambridge, MA, 2002 ISBN 13-978-0-262-19475-4
6. Vapnik V.N. Statistical learning theory. — N.Y.: John Wiley & Sons, Inc., 1998.
7. Witten I.H., Frank E. Data Mining: Practical Machine Learning Tools and Techniques (Second Edition). — Morgan Kaufmann, 2005 ISBN 0-12-088407-0

II. Формы контроля и структура итоговой оценки

Текущий контроль: - письменная аудиторная контрольная работа в первом модуле и индивидуальное домашнее задание.

- Промежуточный контроль – письменный экзамен (120 мин.) в конце первого модуля;
- Итоговый контроль –устный экзамен в конце второго модуля

Формирование оценки.

Оценка работы студентов на семинарских и практических занятиях, $O_{аудиторная,}$, формируется по десятибалльной шкале и выставляется в рабочую ведомость перед промежуточным и перед итоговым контролем. При формировании оценки учитывается: активность на семинарских занятиях, правильность решения задач на семинаре, результаты письменных тестовых опросов.

Результирующая оценка за текущий контроль в первом модуле учитывает результаты студента по текущему контролю следующим образом:

$$O_{текущий} = 0.3\ O_{дз} + 0{,}4 \cdot O_{к/р} + 0{,}3 \cdot O_{аудиторная}\ ;$$

Результирующая оценка за промежуточный контроль в первом модуле в форме экзамена выставляется по следующей формуле, где $O_{экзамен1}$ – оценка за работу непосредственно на экзамене:

$$O_{промежуточный} = 0{,}5 \cdot O_{экзамен1} + 0{,}5 \cdot O_{текущий};$$

Результирующая оценка за текущий контроль во втором модуле учитывает результаты студента по текущему контролю следующим образом:

$$O_{текущий} = O_{аудиторная};$$

Результирующая оценка за итоговый контроль в форме экзамена выставляется по следующей формуле, где $O_{экзамен2}$ – оценка за работу непосредственно на экзамене:

$$O_{итоговый} = 0,4 \cdot O_{экзамен2} + 0,2 \cdot O_{текущий} + 0,4 \cdot O_{промежуточный}.$$

В диплом ставится оценка за итоговый контроль, которая является результирующей оценкой по учебной дисциплине.

Таблица соответствия оценок по десятибалльной и системе зачет/незачет

Оценка по 10-балльной шкале	Оценка по 5-балльной шкале
1	Незачет
2	
3	
4	Зачет
5	
6	
7	
8	
9	
10	

Таблица соответствия оценок по десятибалльной и пятибалльной системе

По десятибалльной шкале	По пятибалльной системе
1 – неудовлетворительно 2 – очень плохо 3 – плохо	неудовлетворительно – 2
4 – удовлетворительно 5 – весьма удовлетворительно	удовлетворительно – 3
6 – хорошо 7 – очень хорошо	хорошо – 4
8 – почти отлично 9 – отлично 10 - блестяще	отлично – 5

III. Программа дисциплины «Машинное обучение»

Тема 1. Основные понятия и примеры прикладных задач

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

Основная литература

1. Айвазян С. А., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: основы моделирования и первичная обработка данных. — М.: Финансы и статистика, 1983
2. Журавлев Ю. И., Рязанов В. В., Сенько О. В. «Распознавание». Математические методы. Программная система. Практические применения. — М.: Фазис, 2006. ISBN 5-7036-0108-8

Дополнительная литература

1. MacKay D. On-line book: Information Theory, Inference, and Learning Algorithms. — 2005
2. Mitchell T. Machine Learning. — McGraw-Hill Science/Engineering/Math, 1997. ISBN 0-07-042807-7

Тема 2. Метрические методы классификации

Метод ближайших соседей и его обобщения

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

Отбор эталонов и оптимизация метрики

- Отбор эталонных объектов. Псевдокод: алгоритм СТОЛП.
- Функция конкурентного сходства, алгоритм FRiS-СТОЛП.
- Функционал полного скользящего контроля, формула быстрого вычисления для метода 1NN. Профиль компактности.

Основная литература

1. Айвазян С. А., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: исследование зависимостей. — М.: Финансы и статистика, 1985
2. 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. Шлезингер М., Главач В. Десять лекций по статистическому и структурному распознаванию. — Киев: Наукова думка, 2004. ISBN 966-00-0341-2
2. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. — Springer, 2001. ISBN 0-387-95284-5

Тема 3. Логические методы классификации

Понятия закономерности и информативности

- Понятие логической закономерности. Эвристическое, статистическое, энтропийное определение информативности. Асимптотическая эквивалентность статистического и

энтропийного определения. Сравнение областей эвристических и статистических закономерностей.

- Разновидности закономерностей: конъюнкции пороговых предикатов (гиперпараллелепипеды), синдромные правила, шары, гиперплоскости.
- Бинаризация признаков. Алгоритм разбиения области значений признака на информативные зоны.

Решающие списки и деревья

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

Основная литература

1. Айвазян С. А., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: исследование зависимостей. — М.: Финансы и статистика, 1985
2. 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. Шлезингер М., Главач В. Десять лекций по статистическому и структурному распознаванию. — Киев: Наукова думка, 2004. ISBN 966-00-0341-2
2. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. — Springer, 2001. ISBN 0-387-95284-5
3. Schölkopf B., Smola A.J. Learning with Kernels. Support Vector Machines, Regularization, Optimization, and Beyond. — MIT Press, Cambridge, MA, 2002 ISBN 13-978-0-262-19475-4 [2]

Тема 4. Линейные методы классификации

Градиентные методы

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

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.
- Метод релевантных векторов RVM
- Регуляризации для отбора признаков: LASSO SVM, Elastic Net SVM, SFM, RFM.

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
2. Вапник В. Н., Червоненкис А. Я. Теория распознавания образов. — М.: Наука, 1974

Дополнительная литература

1. Mitchell T. Machine Learning. — McGraw-Hill Science/Engineering/Math, 1997. ISBN 0-07-042807-7
2. Vapnik V.N. Statistical learning theory. — N.Y.: John Wiley & Sons, Inc., 1998

Тема 5. Методы регрессионного анализа

Многомерная линейная регрессия

- Задача регрессии, многомерная линейная регрессия.
- Метод наименьших квадратов, его вероятностный смысл и геометрический смысл.
- Сингулярное разложение.
- Проблемы мультиколлинеарности и переобучения.
- Регуляризация. Гребневая регрессия. Лассо Тибширани, сравнение с гребневой регрессией.
- Метод главных компонент и декоррелирующее преобразование Карунена-Лоэва, его связь с сингулярным разложением.

Нелинейная параметрическая регрессия

- Метод Ньютона-Рафсона, метод Ньютона-Гаусса.
- Одномерные нелинейные преобразования признаков: метод настройки с возвращениями (backfitting) Хасти-Тибширани.

Непараметрическая регрессия

- Сглаживание. Локально взвешенный метод наименьших квадратов и оценка Надарая-Ватсона.
- Выбор функции ядра. Выбор ширины окна сглаживания. Сглаживание с переменной шириной окна.
- Проблема выбросов и робастная непараметрическая регрессия. Алгоритм LOWESS.

Неквадратичные функции потерь

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

Прогнозирование временных рядов

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

Основная литература

1. Вапник В. Н. Восстановление зависимостей по эмпирическим данным. — М.: Наука, 1979
2. Журавлев Ю. И., Рязанов В. В., Сенько О. В. «Распознавание». Математические методы. Программная система. Практические применения. — М.: Фазис, 2006. ISBN 5-7036-0108-8

Дополнительная литература

1. Schölkopf B., Smola A.J. Learning with Kernels. Support Vector Machines, Regularization, Optimization, and Beyond. — MIT Press, Cambridge, MA, 2002 ISBN 13-978-0-262-19475-4

2. Witten I.H., Frank E. Data Mining: Practical Machine Learning Tools and Techniques (Second Edition). — Morgan Kaufmann, 2005 ISBN 0-12-088407-0

Тема 6. Байесовские методы классификации

Оптимальный байесовский классификатор

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

Непараметрическое оценивание плотности

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

Параметрическое оценивание плотности

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

Разделение смеси распределений

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

Логистическая регрессия

- Гипотеза экспоненциальности функций правдоподобия классов. Теорема о линейности байесовского оптимального классификатора. Оценивание апостериорных вероятностей классов с помощью сигмоидной функции активации.
- Логистическая регрессия. Принцип максимума правдоподобия и логарифмическая функция потерь.

- Метод стохастического градиента для логарифмической функции потерь. Сглаженное правило Хэбба.
- Метод наименьших квадратов с итеративным пересчётом весов (IRLS).
- Пример прикладной задачи: кредитный скоринг. Бинаризация признаков. Скоринговые карты и оценивание вероятности дефолта. Риск кредитного портфеля банка.

Основная литература

1. Журавлев Ю. И., Рязанов В. В., Сенько О. В. «Распознавание». Математические методы. Программная система. Практические применения. — М.: Фазис, 2006. ISBN 5-7036-0108-8
2. Загоруйко Н. Г. Прикладные методы анализа данных и знаний. — Новосибирск: ИМ СО РАН, 1999. ISBN 5-86134-060-9
3. Шлезингер М., Главач В. Десять лекций по статистическому и структурному распознаванию. — Киев: Наукова думка, 2004. ISBN 966-00-0341-2

Дополнительная литература

1. MacKay D. On-line book: Information Theory, Inference, and Learning Algorithms. — 2005
2. Witten I.H., Frank E. Data Mining: Practical Machine Learning Tools and Techniques (Second Edition). — Morgan Kaufmann, 2005 ISBN 0-12-088407-0

IV. Методические указания студентам

Самостоятельная работа студента предусматривает выполнение теоретических заданий, направленных на овладение методами машинного обучения и умениями использовать эти знания при практическом решении соответствующих задач интеллектуального анализа данных.

Автор программы: _ / <Воронцов К.В.> /