

Белорусский государственный университет

УТВЕРЖДАЮ
Проректор по учебной работе и
образовательным инновациям
О.И. Чуприс
«14» июня 2019 г.
Регистрационный № УД- 6951 /уч.

МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ

Учебная программа учреждения высшего образования
по учебной дисциплине для специальности:

1-31 80 09 Прикладная математика и информатика

профилизация

Алгоритмы и системы обработки больших данных

2019 г.

Учебная программа составлена на основе ОСВО 1-31 80 09-2019 и учебного плана G31-072/уч. от 11.04.2019 г.

СОСТАВИТЕЛИ:

И.С. Терлиженко – ассистент кафедры дискретной математики и алгоритмики факультета прикладной математики и информатики Белорусского государственного университета;

А.А. Толстиков – старший преподаватель кафедры вычислительной математики факультета прикладной математики и информатики Белорусского государственного университета.

РЕЦЕНЗЕНТЫ:

А.В. Колесов – инженер-программист ООО «ЯндексБел»;

С.В. Боярович – инженер-программист ООО «МэпДэйта».

РЕКОМЕНДОВАНА К УТВЕРЖДЕНИЮ:

Кафедрой дискретной математики и алгоритмики факультета прикладной математики и информатики БГУ
(протокол № 15 от 18 апреля 2019 года);

Научно-методическим Советом БГУ
(протокол № 5 от 28 июня 2019 года).

Заведующий кафедрой
дискретной математики и алгоритмики 13 В.М. Котов

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Цели и задачи учебной дисциплины

Учебная дисциплина «Методы машинного обучения» знакомит студентов магистратуры с основными направлениями в теории обучения машин и закладывает необходимую теоретическую базу для понимания и разработки новых алгоритмов машинного обучения.

Цель учебной дисциплины – создать базу для применения современных методов машинного обучения и сформировать у студентов магистратуры умения анализировать задачи машинного обучения и осуществлять взвешенный выбор того или иного решения.

Задачи учебной дисциплины:

1. Изучение современной постановки задач машинного обучения, различных этапов её решения.
2. Изучение и сравнительный анализ различных моделей машинного обучения: линейные модели, решающие деревья, нейронные сети.
3. Ознакомление с онлайн-задачами машинного обучения и задач машинного обучения без учителя.

Место учебной дисциплины в системе подготовки специалиста с высшим образованием (магистра).

Учебная дисциплина относится к компоненту учреждения высшего образования и входит в модуль «Модели и методы машинного обучения».

Программа составлена с учетом **межпредметных связей** с учебными дисциплинами. Основой для изучения учебной дисциплины являются следующие учебные дисциплины первой ступени высшего образования: «Теория вероятностей и математическая статистика», «Вычислительные методы алгебры», «Методы численного анализа» и «Математический анализ». Знания, полученные в учебной дисциплине, используются при изучении дисциплин государственного компонента «Нейронные сети» модуля «Модели и методы машинного обучения», «Машинное обучение на больших массивах данных» модуля «Избранные главы компьютерных наук».

Освоение учебной дисциплины «Методы машинного обучения» должно обеспечить формирование следующих специализированных компетенций:

специализированные компетенции:

СК-5. Владеть математическими основами теории машинного обучения.

СК-6. Применять методы машинного обучения для решения прикладных задач.

СК-7. Анализировать, выделять особенности и комбинировать методы машинного обучения.

В результате изучения дисциплины студент магистратуры должен:

знать:

- математические основы теории машинного обучения;
- основные классы алгоритмов машинного обучения и их представителей и их взаимосвязь, достоинства и недостатки;

уметь:

- анализировать, выделять особенности и комбинировать методы машинного обучения;
- применять методы машинного обучения для решения прикладных задач;

владеть:

- программными средствами для разработки алгоритмов машинного обучения, алгоритмами построения искусственных нейронных сетей;
- навыками разработки моделей машинного обучения.

Структура учебной дисциплины

Дисциплина изучается в 1-ом и 2-ом семестре. Всего на изучение учебной дисциплины «Методы машинного обучения» отведено:

– для очной формы получения высшего образования – 252 часа, в том числе 80 аудиторных часов, из них: лекции – 40 часов, практические занятия – 40 часов.

Распределение часов по семестрам:

– в 1-ом семестре для очной формы получения высшего образования – 126 часов, в том числе 40 аудиторных часов, из них: лекции – 20 часов, практические занятия – 20 часов;

в 2-ом семестре для очной формы получения высшего образования – 126 часов, в том числе 40 аудиторных часов, из них: лекции – 20 часов, практические занятия – 20 часов;

Трудоемкость учебной дисциплины составляет 6 зачетных единиц.

Форма текущей аттестации по учебной дисциплине: зачет – в 1-ом семестре, экзамен – во 2-ом семестре.

СОДЕРЖАНИЕ УЧЕБНОГО МАТЕРИАЛА

Раздел 1. Основы машинного обучения

Тема 1.1. Введение

История машинного обучения, основные задачи машинного обучения, краткий обзор курса.

Тема 1.2. Задачи обучения по прецедентам

Объекты, признаки, ответы, функционал качества. Вероятностная постановка задачи обучения. Переобучение, обобщающая способность. Задачи классификации, восстановления регрессии, ранжирования, кластеризации, поиска ассоциаций.

Тема 1.3. Формальная модель машинного обучения

Формальная модель обучения: PAC-learnability. Необходимый размер выборки. Agnostic PAC learning. Оптимальный байесовский классификатор. Обучение через равномерную сходимость. Bias-variance tradeoff. VC-размерность. Другие модели обучения.

Раздел 2. Основные алгоритмы решения задач классификации и восстановления регрессии

Тема 2.1. Метрические методы машинного обучения

Обобщённый метрический классификатор. Метод ближайшего соседа. KNN. Метод потенциальных функций. Метод парзеновского окна. Понятие отступа. Понятие эталонного объекта. Проклятие размерности.

Тема 2.2. Байесовские методы машинного обучения

Оптимальный байесовский классификатор. Задача восстановления плотности распределения. Наивный байесовский классификатор. Непараметрическая оценка плотности. Нормальный дискриминантный анализ. Разделение смеси распределений. EM-алгоритм.

Тема 2.3. Линейные методы машинного обучения и их обобщения

Аппроксимация эмпирического риска. Линейный классификатор и линейная регрессия. Метод стохастического градиентного спуска. Логистическая регрессия. Метод опорных векторов. Kernel trick.

Тема 2.4. Решающие деревья

Решающие деревья. Случайный лес. Бустинг и бэггинг. Примеры решающих деревьев. Информативность. Алгоритмы построения решающих деревьев.

Раздел 3. Избранные главы машинного обучения

Тема 3.1. Визуализация и кластеризация

Качество кластеризации. Статистические, эвристические методы кластеризации. Иерархическая кластеризация. Сети Кохонена. t-SNE. PCA.

Тема 3.2. Искусственные нейронные сети

Модель нейрона. Многослойный перцептрон. Метод обратного распространения ошибки. Свёрточные нейронные сети. Рекуррентные нейронные сети.

УЧЕБНО-МЕТОДИЧЕСКАЯ КАРТА УЧЕБНОЙ ДИСЦИПЛИНЫ

№ п/п	Название раздела, темы	Количество часов				Количество часов УСР	Форма контроля знаний
		Аудиторные					
		Лекции	Практические и семинарские	Лабораторные занятия	Иное		
1	Основы машинного обучения	8	8				
1.1	Введение	2					Устный опрос.
1.2	Задачи обучения по прецедентам	2	4				Отчет по домашнему заданию.
1.3	Формальная модель машинного обучения	4	4				Контрольная работа. Отчет по домашнему заданию с его устной защитой.
2	Основные алгоритмы решения задач классификации и восстановления регрессии	22	22				
2.1	Метрические методы машинного обучения	6	6				Коллоквиум. Отчет по домашнему заданию с его устной защитой.
2.2	Байесовские методы машинного обучения	6	6				Контрольная работа. Отчет по домашнему заданию с его устной защитой.

							заданию с его устной защитой.
2.3	Линейные методы машинного обучения и их обобщения	6	6				Отчет по домашнему заданию с его устной защитой.
2.4	Решающие деревья	4	4				Контрольная работа. Отчет по домашнему заданию с его устной защитой.
3.	Избранные главы машинного обучения	10	10				
3.1	Визуализация и кластеризация	4	4				Контрольная работа. Отчет по домашнему заданию с его устной защитой.
3.2	Искусственные нейронные сети	6	6				Коллоквиум. Отчет по домашнему заданию с его устной защитой.

ИНФОРМАЦИОННО-МЕТОДИЧЕСКАЯ ЧАСТЬ

Перечень основной литературы

1. Shalev-Shwartz Shai, Ben-David Shai. Understanding Machine Learning: From Theory to Algorithms, Cambridge University Press, 2014. — 409 p. — ISBN-13: 978-1107057135.
2. Bishop, C.M. Pattern Recognition and Machine Learning. — Springer, 2006. — 738 p.
3. Вапник, В. Н. Восстановление зависимостей по эмпирическим данным. — М.: Наука, 1979. — 448 с.
4. Hastie, T., Tibshirani, R., Friedman, J. The Elements of Statistical Learning, 2nd edition. — Springer, 2009. — 533 p.

Перечень дополнительной литературы

1. Cristianini, N. & Shawe-Taylor, J. (2000), An Introduction to Support Vector Machines, Cambridge University Press.
2. Kearns, M. & Vazirani, U. (1994), An Introduction to Computational Learning Theory, MIT Press.
3. Nemirovski, A. & Yudin, D. (1978), Problem complexity and method efficiency in optimization, Nauka Publishers, Moscow.

Перечень рекомендуемых средств диагностики и методика формирования итоговой оценки

Для диагностики компетенции в рамках учебной дисциплины рекомендуется использовать следующие формы:

1. Устная форма: устный опрос, коллоквиум.
2. Письменная форма: контрольные работы.
3. Устно-письменная форма: отчеты по домашним практическим заданиям с их устной защитой, оценивание на основе проектного метода.

Формой текущей аттестации по дисциплине «Методы машинного обучения» учебным планом предусмотрен: в 1-ом семестре – зачет, во 2-ом семестре – экзамен.

При формировании итоговой оценки используется рейтинговая оценка знаний студента, дающая возможность проследить и оценить динамику процесса достижения целей обучения. Рейтинговая оценка предусматривает использование весовых коэффициентов для текущего контроля знаний студентов по дисциплине.

Примерные весовые коэффициенты, определяющие вклад текущего контроля знаний в рейтинговую оценку (формирование оценки за текущую успеваемость):

- отчёты по практическим домашним заданиям с их устной защитой – 30 %;
- контрольные работы – 30 %;
- коллоквиум – 20 %;
- устный опрос – 20%;

Рейтинговая оценка по дисциплине рассчитывается на основе оценки текущей успеваемости и экзаменационной оценки с учетом их весовых коэффициентов Вес оценка по текущей успеваемости составляет 30 %, экзаменационная оценка – 70 %.

Примерная тематика практических занятий

Занятие № 1-2. Основы машинного обучения. ERM-парадигма.

Занятие № 3-4. Модели машинного обучения. Линейные модели.

Занятие № 5-6. VC-размерность. Логистическая регрессия.

Занятие № 7-8. Бустинг. Неравномерная изучаемость.

Занятие № 9-10. Стохастический градиентный спуск. Метод опорных векторов.

Занятие № 11-12. Метрические методы классификации. Кластеризация.

Занятие № 13-14. Понижение размерности. t-SNE и PCA.

Занятие № 15-16. Методы восстановления плотности и байесовские классификаторы.

Занятие № 17-20. Нейронные сети.

Описание инновационных подходов и методов к преподаванию учебной дисциплины (эвристический, проективный, практико-ориентированный)

При организации образовательного процесса большинства практических занятий используется практико-ориентированный подход, который предполагает освоение содержания учебного материала через решение практических задач, а также приобретение навыков эффективного выполнения разных видов профессиональной деятельности.

Кроме этого, при организации образовательного процесса используется комбинация методов группового обучения, проектного обучения и учебной дискуссии. Комбинация методов предполагает: ориентацию на генерирование идей, приобретение навыков для решения исследовательских, творческих и коммуникационных задач, появление нового уровня понимания изучаемой темы, применение знаний (теорий, концепций) при решении проблем, определение способов их решения.

Методические рекомендации по организации самостоятельной работы обучающихся, подготовка к зачету, подготовка к экзамену

Для организации самостоятельной работы студентов магистратуры по учебной дисциплине следует использовать современные информационные технологии: разместить в сетевом доступе комплекс учебных и учебно-методических материалов (учебно-программные материалы, учебное издание для теоретического изучения дисциплины, презентации лекций, методические указания к практическим занятиям, электронные версии домашних заданий, материалы текущего контроля и текущей аттестации, позволяющие определить соответствие учебной деятельности обучающихся требованиям образовательных стандартов высшего образования и учебно-программной документации, в том числе вопросы для подготовки к зачёту, задания, вопросы для самоконтроля, список рекомендуемой литературы, информационных ресурсов и др.).

Примерный перечень вопросов к зачету

Основные понятия теории машинного обучения.

- 1) Задача обучения по прецедентам. Объекты и признаки. Ответы и типы задач.
- 2) Модель алгоритмов и метод обучения. Этап обучения и этап применения.
- 3) Функционал качества. Примеры. Эмпирический риск. Переобучение, его возможные причины и способы его минимизации.
- 4) Сведение задачи обучения к задаче оптимизации. Примеры.
- 5) Обобщающая способность. Формализации понятия «обобщающая способность». Проблема обобщающей способности.
- 6) Этапы решения задач машинного обучения.
- 7) Типы задач, причисляемые к машинному обучению.
- 8) Понятие обучения с учителем и без. Примеры алгоритмов. Особенности.
- 9) Постановка задачи классификации. Примеры.
- 10) Постановка задачи регрессии. Примеры.
- 11) Постановка задачи ранжирования. Примеры.
- 12) Вероятностный подход в машинном обучении. Наивный Байесовский классификатор. Оценка его оптимальности и возможности применения на практике.
- 13) Возможность применения одних и тех же алгоритмов машинного обучения для задач разных типов. Примеры.
- 14) Основные алгоритмы, применяемые в задачах классификации.
- 15) Основные алгоритмы, применяемые в задачах регрессии.
- 16) Основные алгоритмы, применяемые в задачах кластеризации.

- 17) Основные алгоритмы, применяемые в задачах коллаборативной фильтрации.
- 18) Основные этапы решения практической задачи.
- 19) Основные проблемы исходных данных и способы их решения (пропуски, выбросы, несовместимые с алгоритмом типы данных).
- 20) Способы визуализации данных.
- 21) Методы минимизации числа признаков. Метод главных компонент (РСА).
- 22) Понятие интерпретируемости алгоритма и принимаемого им решения по отдельному прецеденту. Поясняющие примеры задач.
- 23) Функция потерь. Гладкие аппроксимации пороговой функции потерь. Основные виды и их особенности. Примеры.

Линейные методы

- 24) Понятие оптимальной разделяющей гиперплоскости. Линейно разделимые и линейно неразделимые выборки. Обоснование кусочно-линейной функции потерь.
- 25) Типы объектов: периферийные, опорные (граничные и нарушители).
- 26) Линейный метод опорных векторов (linearSVM). Нелинейное обобщение SVM (kernelSVM). Ядра. Особенности метода проверки, является функция ядром или нет. Конструктивные методы синтеза ядер.
- 27) Примеры ядер.
- 28) Метод опорных векторов. Преимущества и недостатки.
- 29) Определение ROC-кривой и AUC. Характеристики ROC-кривой. Эффективный алгоритм вычисления AUC и его асимптотическая сложность.

Коллаборативная фильтрация

- 30) Бустинг. Основная идея. Бустинг для бинарной задачи классификации.
- 31) Бустинг. Основная идея. Основная теорема бустинга (для AdaBoost).
- 32) Переобучение при применении бустинга. Особенности в случаях построения комбинации простых и сложных алгоритмов.
- 33) Алгоритм AdaBoost. Эвристики и рекомендации.
- 34) Стохастические методы построения композиций: bagging и RSM (random subspace method). Сравнительный анализ: бустинг, bagging и RSM.
- 35) Случайный лес (random forest). Особенности, преимущества и недостатки.

Методы кластеризации

- 36) Постановка задачи кластеризации. Цели. Некорректность задачи кластеризации.
- 37) Основные типы кластерных структур. Особенности и схематичных чертежи.
- 38) Алгоритм выделения связных компонент.
- 39) Функции качества кластеризации
- 40) Агломеративная иерархическая кластеризация. Формула Ланса-Уильямса.

- 41) Частные случаи формулы Ланса-Уильямса.
- 42) Визуализация кластерных структур: диаграмма вложения, дендрограмма.
- 43) EM-алгоритм. Цель.
- 44) Метод k-средних (k-means).

Примерный перечень вопросов к экзамену

Основные понятия теории машинного обучения

- 1) Нейронные сети. Структура сети.
- 2) Сжимающие нейронные сети. Цель. Структура и особенности.
- 3) Применение нейронных сетей для задач генерации признаков.
- 4) Глубокое обучение (Deep learning).
- 5) Два подхода к реализации многоклассового классификатора на основе бинарных классификаторов.
- 6) Статистический анализ данных. Примеры статистик и их интерпретация.
- 7) Сведения задачи регрессии к задаче классификации. Цель. Особенности.
- 8) Задача классификации на N пересекающихся классов. Примеры. Возможные подходы к решению.
- 9) Особенности при решении задач с большим объемом данных (Big Data).
- 10) on-line и off-line задачи в машинном обучении. Примеры.
- 11) Внедрение решений практических задач, полученных в off-line конкурсах Kaggle

Метрические методы

- 12) Метрические методы классификации. Гипотеза компактности.
- 13) Обобщенный метрический классификатор.
- 14) Метод ближайшего соседа и его обобщения. Особенности, преимущества и недостатки.
- 15) Метод окна Парзена.
- 16) Метод потенциальных функций.
- 17) Пример алгоритма настройки весов объектов. Особенности, преимущества и недостатки.
- 18) Понятие отступа. Типы объектов в зависимости от отступа.
- 19) Понятие эталона. Два подхода к отбору эталонов.
- 20) Метрика. Примеры. Взвешенная метрика Минковского, ее анализ.
- 21) Полный скользящий контроль. Профиль компактности.
- 22) Логические алгоритмы
- 23) Понятия закономерности, информативности и интерпретируемости.
- 24) Основные виды закономерностей.
- 25) Примеры критериев информативности. Возможные их недостатки.
- 26) Пример алгоритма поиска информативных закономерностей.

- 27) Бинарное решающее дерево.
- 28) Пример алгоритма построения бинарного решающего дерева (ID3). Его достоинства и недостатки.
- 29) Обработка пропусков в данных на стадии обучения и на стадии применения алгоритма.
- 30) Понятие редукции решающего дерева. Цель. Пример алгоритма.
- 31) Небрежные решающие деревья (ODT). Алгоритм обучения ODT.
- 32) Бинаризация вещественного признака. Цель. Способы разбиения области значений признака на зоны. Пример алгоритма.

Коллаборативная фильтрация

- 33) Постановка задачи коллаборативной фильтрации. Примеры задач.
- 34) Определение корреляционных и латентных моделей. Сравнительный анализ возможности их применения к задачам BigData.
- 35) Корреляционные модели. Особенности. Примеры. Понятие холодного старта.
- 36) Латентные модели. Типы латентных моделей. Особенности.
- 37) Обобщающая способность, методы отбора признаков
- 38) Внутренний и внешний критерии оценки качества обучения по прецедентам. Преимущества и недостатки.
- 39) Кросс-проверка (cross-validation). Полная кросс-проверка (CCV). Скользящий контроль (LOO). Кросс-проверка по k блокам, одинарная (k-foldCV) и многократная (t*k-foldCV).
- 40) Критерий непротиворечивости модели. Цель, преимущества и недостатки.
- 41) Понятие регуляризации. Цель. Примеры регуляризаторов.
- 42) Задача отбора признаков в логических закономерностях.
- 43) Задача отбора признаков. Алгоритм полного перебора. Особенности, преимущества и недостатки.
- 44) Задача отбора признаков. Алгоритмы Add, Del, Add-Del. Особенности, преимущества и недостатки.
- 45) Задача отбора признаков. Генетический алгоритм. Особенности, преимущества и недостатки.
- 46) Композиция классификаторов
- 47) Композиция классификаторов. Цель. Определение композиции.
- 48) Композиция классификаторов. Корректирующие операции: простое голосование, взвешенное голосование, смесь алгоритмов. Примеры.

Рекомендуемая тематика контрольных работ

- 1) Контрольная работа №1. *Модели машинного обучения*
- 2) Контрольная работа №2. *Бустинг над решающими деревьям*
- 3) Контрольная работа №3. *t-SNE и PCA*
- 4) Контрольная работа №4. *Методы восстановления плотности и байесовские классификаторы.*

Текущий контроль знаний проводится в соответствии с учебно-методической картой дисциплины.

ПРОТОКОЛ СОГЛАСОВАНИЯ УЧЕБНОЙ ПРОГРАММЫ

Название учебной дисциплины, с которой требуется согласование	Название кафедры	Предложения об изменениях в содержании учебной программы учреждения высшего образования по учебной дисциплине	Решение, принятое кафедрой, разработавшей учебную программу (с указанием даты и номера протокола)
Нейронные сети	Дискретной математики и алгоритмики	Нет	Оставить содержание учебной дисциплины без изменения, протокол № 15 от 18 апреля 2019 г.
Машинное обучение на больших массивах данных	Дискретной математики и алгоритмики	Нет	Оставить содержание учебной дисциплины без изменения, протокол № 15 от 18 апреля 2019 г.

ДОПОЛНЕНИЯ И ИЗМЕНЕНИЯ К УЧЕБНОЙ ПРОГРАММЕ

на ____/____ учебный год

№№ Пп	Дополнения и изменения	Основание

Учебная программа пересмотрена и одобрена на заседании кафедры дискретной математики и алгоритмики (протокол № ____ от _____ 20__ г.)

Заведующий кафедрой

(ученая степень, звание)

(подпись)

(И.О. Фамилия)

УТВЕРЖДАЮ

Декан факультета

(ученая степень, звание)

(подпись)

(И.О.Фамилия)