

Белорусский государственный университет

УТВЕРЖДАЮ

Проректор по учебной работе
и образовательным инновациям


О.И. Чуприс

19.02.2018

Регистрационный № УД-5724/уч.



МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ

**Учебная программа учреждения высшего образования
по учебной дисциплине для специальности второй ступени высшего
образования (магистратуры) с углубленной подготовкой специалиста:**

**1-31 81 09 Алгоритмы и системы обработки больших объемов
информации**

2018 г.

Учебная программа составлена на основе образовательного стандарта высшего образования ОСВО 1-31 81 09-2014 и учебного плана G31-254/уч. от 26.05.2017 г.

Составители:

А.В. Колесов – ассистент кафедры дискретной математики и алгоритмики факультета прикладной математики и информатики Белорусского государственного университета.

Рекомендована к утверждению:

Кафедрой дискретной математики и алгоритмики Белорусского государственного университета (протокол № 8 от 18 января 2018 г.);

Научно-методическим советом Белорусского государственного университета (протокол № 4 от 15 февраля 2018 г.).

Заведующий кафедрой ДМ,
профессор

Комов В.М.

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Учебная программа по учебной дисциплине «Методы машинного обучения» относится к компоненту учреждения высшего образования цикла дисциплин специальной подготовки и разработана с учетом требований следующих нормативных и методических документов:

Образовательный стандарт Республики Беларусь «Высшее образование. II ступень. Специальность 1-31 81 09 «Алгоритмы и системы обработки больших объемов информации», утвержден постановлением Министерства образования Республики Беларусь от 10.09.2014 г. №145;

Учебный план по специальности 1-31 81 09 «Алгоритмы и системы обработки больших объемов информации», утвержденный ректором 26.05.2017 г. (регистрационный № G31-254/уч.).

Программа отражает содержание учебной дисциплины «Методы машинного обучения».

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.

Программа составлена с учетом межпредметных связей с учебными дисциплинами. Так, основой для изучения учебной дисциплины являются следующие учебные дисциплины I ступени высшего образования: «Теория вероятностей и математическая статистика», «Вычислительные методы алгебры», «Методы численного анализа» и «Математический анализ». Знания, полученные в учебной дисциплине «Методы машинного обучения», используются при изучении учебных дисциплин «Нейронные сети» и «Машинное обучение на больших массивах данных».

Цель преподавания учебной дисциплины «Методы машинного обучения»: создание базы для применения современных методов машинного обучения и формирование у студентов магистратуры умения анализировать задачи машинного обучения и осуществлять взвешенный выбор того или иного решения.

При изложении материала учебной дисциплины важно показать многообразие методов машинного обучения и возможность их комбинации для решения поставленной задачи.

В рамках поставленной цели **задачи** учебной дисциплины состоят в следующем:

- изучение современной постановки задач машинного обучения, различных этапов её решения;
- изучение и сравнительный анализ различных моделей машинного обучения: линейные модели, решающие деревья, нейронные сети;

- ознакомление с онлайн-задачами машинного обучения и задач машинного обучения без учителя.

В результате освоения учебной 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.

Освоение учебной дисциплины «Методы машинного обучения» должно обеспечить формирование следующих социально-личностных и профессиональных компетенций:

социально-личностные компетенции:

СЛК-1. Учитывать социальные и нравственно-этические нормы в социально-профессиональной деятельности.

СЛК-2. Быть способным к сотрудничеству и работе в команде.

СЛК-3. Владеть коммуникативными способностями для работы в междисциплинарной и международной среде.

СЛК-4. Совершенствовать и развивать свой интеллектуальный и общекультурный уровень, добиваться нравственного и физического совершенствования своей личности.

СЛК-6. Логично, аргументировано и ясно строить устную и письменную речь, использовать навыки публичной речи, ведения дискуссии и полемики.

СЛК-7. Проявлять инициативу и креативность, в том числе в нестандартных ситуациях);

профессиональных компетенции:

ПК-3. Разрабатывать эффективные численные алгоритмы и интегрировать их в компьютерные системы.

Всего на изучение учебной дисциплины «Методы машинного обучения» отведено 180 часов, в том числе 64 аудиторных часа. Примерное распределение аудиторных часов по видам занятий: лекции – 32 часа, практические занятия – 32 часа.

Трудоемкость учебной дисциплины составляет 5 зачётные единицы.

Форма текущей аттестации – экзамен в 3 семестре.

СОДЕРЖАНИЕ УЧЕБНОГО МАТЕРИАЛА

Раздел I. 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 II. 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 III. 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. Многослоный перцептрон. Метод обратного распространения ошибки. Свёрточные нейронные сети. Рекуррентные нейронные сети.

УЧЕБНО-МЕТОДИЧЕСКАЯ КАРТА УЧЕБНОЙ ДИСЦИПЛИНЫ

№ п/п	Название раздела, темы	Распределение аудиторных часов по видам деятельности		Форма контроля знаний
		Лекции	Практические занятия	
1	Основы машинного обучения	6	6	
1.1	Введение	2		
1.2	Задачи обучения по прецедентам	2		Устный опрос
	<i>Основы анализа данных на языке Python</i>		3	Отчёт по домашнему заданию.
1.3	Формальная модель машинного обучения	2		Устный опрос
	<i>Модели машинного обучения</i>		3	Отчёт по домашнему заданию. Контрольная работа №1.
2	Основные алгоритмы решения задач классификации и восстановления регрессии	20	20	
2.1	Метрические методы машинного обучения	6		Устный опрос
	<i>Метод ближайших соседей</i>		4	Отчёт по домашнему заданию.
2.2	Байесовские методы машинного обучения	4		Устный опрос
	<i>Методы восстановления плотности и байесовские классификаторы</i>		6	Контрольная работа №2.
2.3	Линейные методы машинного обучения и их обобщения	6		Устный опрос
	<i>Логистическая регрессия и метод опорных векторов</i>		4	Отчёт по домашнему заданию.
2.4	Решающие деревья	4		Устный опрос

	<i>Бустинг над решающими деревьями</i>		6	Отчёт по домашнему заданию. Контрольная работа №3.
3.	Избранные главы машинного обучения	6	6	
3.1	Визуализация и кластеризация	3		Устный опрос
	<i>t-SNE и PCA</i>		2	Отчёт по домашнему заданию. Контрольная работа №4.
3.2	Искусственные нейронные сети	3		Устный опрос
	<i>Нейронные сети</i>		4	Отчёт по домашнему заданию. Коллоквиум.
ИТОГО		32	32	

ИНФОРМАЦИОННО-МЕТОДИЧЕСКАЯ ЧАСТЬ

Рекомендуемая литература

Основная

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.

Дополнительная

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.

Рекомендуемая тематика контрольных работ

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

Примерный перечень вопросов к экзамену

Основные понятия теории машинного обучения

- 1) Основные понятия теории машинного обучения.
- 2) Задача обучения по прецедентам. Объекты и признаки. Ответы и типы задач.
- 3) Модель алгоритмов и метод обучения. Этап обучения и этап применения.
- 4) Функционал качества. Примеры. Эмпирический риск. Переобучение, его возможные причины и способы его минимизации.
- 5) Сведение задачи обучения к задаче оптимизации. Примеры.

- 6) Обобщающая способность. Формализации понятия «обобщающая способность». Проблема обобщающей способности.
- 7) Этапы решения задач машинного обучения.
- 8) Типы задач, причисляемые к машинному обучению.
- 9) Понятие обучения с учителем и без. Примеры алгоритмов. Особенности.
- 10) Постановка задачи классификации. Примеры.
- 11) Постановка задачи регрессии. Примеры.
- 12) Постановка задачи ранжирования. Примеры.
- 13) Вероятностный подход в машинном обучении. Наивный Байесовский классификатор. Оценка его оптимальности и возможности применения на практике.
- 14) Возможность применения одних и тех же алгоритмов машинного обучения для задач разных типов. Примеры.
- 15) Основные алгоритмы, применяемые в задачах классификации.
- 16) Основные алгоритмы, применяемые в задачах регрессии.
- 17) Основные алгоритмы, применяемые в задачах кластеризации.
- 18) Основные алгоритмы, применяемые в задачах коллаборативной фильтрации.
- 19) Основные этапы решения практической задачи.
- 20) Основные проблемы исходных данных и способы их решения (пропуски, выбросы, несовместимые с алгоритмом типы данных).
- 21) Способы визуализации данных.
- 22) Методы минимизации числа признаков. Метод главных компонент (PCA).
- 23) Понятие интерпретируемости алгоритма и принимаемого им решения по отдельному прецеденту. Поясняющие примеры задач.
- 24) Функция потерь. Гладкие аппроксимации пороговой функции потерь. Основные виды и их особенности. Примеры.
- 25) Нейронные сети. Структура сети.
- 26) Сжимающие нейронные сети. Цель. Структура и особенности.
- 27) Применение нейронных сетей для задач генерации признаков.
- 28) Глубокое обучение (Deep learning).
- 29) Два подхода к реализации многоклассового классификатора на основе бинарных классификаторов.
- 30) Статистический анализ данных. Примеры статистик и их интерпретация.
- 31) Сведения задачи регрессии к задаче классификации. Цель. Особенности.
- 32) Задача классификации на N пересекающихся классов. Примеры. Возможные подходы к решению.
- 33) Особенности при решении задач с большим объемом данных (BigData).
- 34) on-line и off-line задачи в машинном обучении. Примеры.

- 35) Внедрение решений практических задач, полученных в off-line конкурсах Kaggle

Метрические методы

- 36) Метрические методы классификации. Гипотеза компактности.
 37) Обобщенный метрический классификатор.
 38) Метод ближайшего соседа и его обобщения. Особенности, преимущества и недостатки.
 39) Метод окна Парзена.
 40) Метод потенциальных функций.
 41) Пример алгоритма настройки весов объектов. Особенности, преимущества и недостатки.
 42) Понятие отступа. Типы объектов в зависимости от отступа.
 43) Понятие эталона. Два подхода к отбору эталонов.
 44) Метрика. Примеры. Взвешенная метрика Минковского, ее анализ.
 45) Полный скользящий контроль. Профиль компактности.

Логические алгоритмы

- 46) Понятия закономерности, информативности и интерпретируемости.
 47) Основные виды закономерностей.
 48) Примеры критериев информативности. Возможные их недостатки.
 49) Пример алгоритма поиска информативных закономерностей.
 50) Бинарное решающее дерево.
 51) Пример алгоритма построения бинарного решающего дерева (ID3). Его достоинства и недостатки.
 52) Обработка пропусков в данных на стадии обучения и на стадии применения алгоритма.
 53) Понятие редукции решающего дерева. Цель. Пример алгоритма.
 54) Небрежные решающие деревья (ODT). Алгоритм обучения ODT.
 55) Бинаризация вещественного признака. Цель. Способы разбиения области значений признака на зоны. Пример алгоритма.

Линейные методы

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

Коллаборативная фильтрация

- 62) Постановка задачи коллаборативной фильтрации. Примеры задач.

63) Определение корреляционных и латентных моделей. Сравнительный анализ возможности их применения к задачам BigData.

64) Корреляционные модели. Особенности. Примеры. Понятие холодного старта.

65) Латентные модели. Типы латентных моделей. Особенности.

Обобщающая способность, методы отбора признаков

66) Внутренний и внешний критерии оценки качества обучения по прецедентам. Преимущества и недостатки.

67) Кросс-проверка (cross-validation). Полная кросс-проверка (CCV). Скользящий контроль (LOO). Кросс-проверка по k блокам, одинарная (k-foldCV) и многократная (t*k-foldCV).

68) Критерий непротиворечивости модели. Цель, преимущества и недостатки.

69) Понятие регуляризации. Цель. Примеры регуляризаторов.

70) Задача отбора признаков в логических закономерностях.

71) Задача отбора признаков. Алгоритм полного перебора. Особенности, преимущества и недостатки.

72) Задача отбора признаков. Алгоритмы Add, Del, Add-Del. Особенности, преимущества и недостатки.

73) Задача отбора признаков. Генетический алгоритм. Особенности, преимущества и недостатки.

Композиция классификаторов

74) Композиция классификаторов. Цель. Определение композиции.

75) Композиция классификаторов. Корректирующие операции: простое голосование, взвешенное голосование, смесь алгоритмов. Примеры.

76) Бустинг. Основная идея. Бустинг для бинарной задачи классификации.

77) Бустинг. Основная идея. Основная теорема бустинга (для AdaBoost).

78) Переобучение при применении бустинга. Особенности в случаях построения комбинации простых и сложных алгоритмов.

79) Алгоритм AdaBoost. Эвристики и рекомендации.

80) Стохастические методы построения композиций: baggingи RSM (random subspace method). Сравнительный анализ: бустинг, baggingи RSM.

81) Случайный лес (random forest). Особенности, преимущества и недостатки.

Методы кластеризации

82) Постановка задачи кластеризации. Цели. Некорректность задачи кластеризации.

83) Основные типы кластерных структур. Особенности и схематичных чертежи.

84) Алгоритм выделения связных компонент.

85) Функции качества кластеризации

86) Агломеративная иерархическая кластеризация. Формула Ланса-Уильямса.

87) Частные случаи формулы Ланса-Уильямса.

- 88) Визуализация кластерных структур: диаграмма вложения, дендрограмма.
- 89) ЕМ-алгоритм. Цель.
- 90) Метод k-средних (k-means).

Методические рекомендации по организации самостоятельной работы обучающихся

Для организации самостоятельной работы магистрантов 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 семинарским 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 т.ч. вопросы для подготовки к экзамену, задания, тесты, вопросы для самоконтроля, список рекомендуемой литературы, информационных ресурсов и др.).

Эффективность самостоятельной работы студентов проверяется в ходе текущего и итогового контроля знаний. Для общей оценки качества усвоения студентами учебного материала рекомендуется использование рейтинговой системы.

Перечень рекомендуемых средств диагностики

Для текущего контроля качества усвоения знаний студентами магистратуры используется следующий диагностический инструментарий:

1. Устная форма: устные опросы, устные защиты домашних заданий, проведение коллоквиума «Методы машинного обучения».
2. Письменная форма: отчеты по домашним заданиям, письменные контрольные работы по отдельным темам учебной дисциплины.
3. Техническая форма: система AnyTask (<https://anytask.org/school/bsu>) – инструменты с эффективной функциональностью контроля, тренинга и самостоятельной работы.

Методика формирования итоговой оценки

Формой текущей аттестации по дисциплине «Методы машинного обучения» учебным планом предусмотрен экзамен.

Рекомендуется использовать рейтинговую оценку знаний студента, дающую возможность проследить и оценить динамику процесса достижения целей обучения.

Рейтинговая оценка предусматривает использование весовых коэффициентов для текущего контроля знаний и текущей аттестации студентов по дисциплине. Примерные весовые коэффициенты, определяющие вклад текущего контроля знаний в рейтинговую оценку:

- 1) работа на практических занятиях – 0.6;
- 2) контрольные работы – 0.2;
- 3) коллоквиумы – 0.2.

Рейтинговая оценка по дисциплине рассчитывается на основе оценки текущей успеваемости и экзаменационной оценки с учетом их весовых коэффициентов:

- 1) оценка по текущей успеваемости – 0.3,
- 2) экзаменационная оценка – 0.7.

Итоговая оценка формируется на основе:

- 1) Правил проведения аттестации студентов (Постановление Министерства образования Республики Беларусь № 53 от 29 мая 2012 г.);
- 2) Положение о рейтинговой системе оценки знаний по дисциплине в БГУ (Приказ ректора БГУ от 18.08.2015 № 382-ОД);
- 3) Критериев оценки знаний студентов (письмо Министерства образования от 22.12.2003).

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 № 8 от 18.01.2018 г
Машинное обучение на больших массивах данных	Дискретной математики и алгоритмики	Нет	Оставить содержание учебной дисциплины без изменения, протокол № 8 от 18.01.2018 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 _____ 201_ г.)

Заведующий кафедрой

(ученая степень, звание)

(подпись)

(И.О. Фамилия)

УТВЕРЖДАЮ
Декан факультета

(ученая степень, звание)

(подпись)

(И.О.Фамилия)