Белорусский государственный университет

УТВЕРЖДАЮ
Проректор по учебной работе и
образовательным инновациям
_____ О. И. Чуприс
«10» 12 2018 г.

Регистрационный № УД-6163 /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 12 Прикладной компьютерный анализ данных

2018 г.

Учебная программа составлена на основе образовательного стандарта высшего образования ОСВО 1-31 81 12-2015 и учебного плана G-31-251/уч. от 26.05.2017.

СОСТАВИТЕЛЬ:

С.Н. Сталевская, доцент кафедры математического моделирования и анализа данных факультета прикладной математики и информатики Белорусского государственного университета, кандидат физико-математических наук.

РЕКОМЕНДОВАНА К УТВЕРЖДЕНИЮ:

Кафедрой математического моделирования и анализа данных Белорусского государственного университета (протокол № 6 от 6 ноября 2018 г.);

Научно-методическим Советом Белорусского государственного университета (протокол № 1 от 16 ноября 2018 г.).

Бодени И.А. зав. кафедрой ММАД

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Цель преподавания учебной дисциплины – формирование у студентов II ступени (магистрантов) теоретических знаний и практических навыков для решения научно-исследовательских и прикладных задач связанных с выявлением закономерностей в современных прикладных задачах.

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

1) обучение методам анализа данных;
2) формирование практических умений и навыков работы с данными с помощью языка Python;
3) приобретение навыков для интерпретации полученных результатов.

Учебная дисциплина «Методы машинного обучения» относится к циклу дисциплин специальной подготовки компонента учреждения высшего образования.

Учебная программа составлена с учетом межпредметных связей с учебными дисциплинами. Так, основой для изучения дисциплины «Методы машинного обучения» являются дисциплины первой ступени «Теория вероятностей и математическая статистика», «Введение в компьютерный и интеллектуальный анализ данных», а также дисциплины II ступени высшего образования «Методы статистического анализа многомерных данных» и «Скриптовые языки программирования (Python)». Знания, полученные в результате изучения дисциплины, будут использованы при изучении дисциплины II ступени высшего образования «Методы нахождения и анализа зависимостей в данных», а также способствовать успешному прохождению производственной практики по специальности и подготовки магистерской диссертации.

В результате освоения учебной дисциплины студент магистратуры должен:

знать:

– основные методы классификации с учителем;

– основные методы классификации без учителя;

– регрессионные методы;

– деревья решений;

– машину на опорных векторах.

уметь:

– использовать методы для выявления зависимостей из данных;

– строить классификаторы и регрессионные модели для прогнозирования;

– интерпретировать полученные результаты;

владеть:

– основным понятийным аппаратом, описывающим данные с панельной структурой;

– инструментарием на языке Python для прогнозирования данных.

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.

профессиональные компетенции:

ПК-3. Самостоятельно разрабатывать эффективные численные методы и алгоритмы, а также интегрировать их в компьютерные системы анализа данных.

Структура содержания учебной дисциплины включает такие дидактические единицы, как темы (разделы), в соответствии с которыми разрабатываются и реализуются соответствующие лекционные и лабораторные занятия. Примерная тематика занятий приведена в информационно-методической части.

Дисциплина изучается во 2 семестре (II ступень). Всего на освоение учебной дисциплины «Методы машинного обучения» отведено 168 часов, в том числе 56 аудиторных часов, из них: лекции – 20 часов, лабораторные занятия – 18 часов, практические занятия – 18 часов.

Трудоемкость учебной дисциплины составляет 4 зачетные единицы.

Форма текущей аттестации – зачет, экзамен.

СОДЕРЖАНИЕ УЧЕБНОГО МАТЕРИАЛА

Раздел 1. Прогнозирование дискретных данных

Тема 1.1. Введение. Основные понятия и примеры прикладных задач. Типы задач: классификация, регрессия, прогнозирования. Функция потерь и функционал качества. Методика экспериментального исследования и сравнения алгоритмов на модельных и реальных данных.

Тема 1.2. Байесовская классификация и оценивание плотности. Принцип максимума апостериорной вероятности. 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. 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. Проблемы мультиколлинеарности и переобучения. Регуляризация ковариационной матрицы. Параметрический наивный байесовский классификатор. Смесь распределений. EM-алгоритм как метод простых итераций для решения системы нелинейных уравнений. Выбор числа компонентов смеси. Пошаговая стратегия. Априорное распределение Дирихле.Смесь многомерных нормальных распределений. Сеть радиальных базисных функций (RBF) и применение EM-алгоритма для её настройки.

Тема 1.3. Логические методы классификации. Понятие логической закономерности. Параметрические семейства закономерностей: конъюнкции пороговых правил, синдромные правила, шары, гиперплоскости. Переборные алгоритмы синтеза конъюнкций: стохастический локальный поиск, стабилизация, редукция. Двухкритериальный отбор информативных закономерностей, парето-оптимальный фронт в (p,n)-пространстве. Решающее дерево. Жадная нисходящая стратегия «разделяй и властвуй». Алгоритм ID3. Недостатки жадной стратегии и способы их устранения. Проблема переобучения. Вывод критериев ветвления. Мера нечистоты (impurity) распределения. Энтропийный критерий, критерий Джини. Редукция решающих деревьев: предредукция и постредукция. Алгоритм C4.5. Деревья регрессии. Алгоритм CART. Небрежные решающие деревья (oblivious decision tree). Решающий лес. Случайный лес (Random Forest).

Тема 1.4. 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. SVM-регрессия. Регуляризации для отбора признаков: LASSO SVM, Elastic Net SVM, SFM, RFM. Метод релевантных векторов RVM

Раздел 2. Прогнозирование непрерывных данных

Тема 2.1. 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 через сингулярное разложение. Методы отбора признаков: Лассо Тибширани, Elastic Net, сравнение с гребневой регрессией. Метод главных компонент и декоррелирующее преобразование Карунена-Лоэва, его связь с сингулярным разложением. Спектральный подход к решению задачи наименьших квадратов. Задачи и методы низкоранговых матричных разложений.

Тема 2.2. Нелинейная регрессия. Метод Ньютона-Рафсона, метод Ньютона-Гаусса. Обобщённая аддитивная модель (GAM): метод настройки с возвращениями (backfitting) Хасти-Тибширани. Логистическая регрессия. 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. Обобщённая линейная модель (GLM). Экспоненциальное семейство распределений. Неквадратичные функции потерь. Метод наименьших модулей. Квантильная регрессия. Пример прикладной задачи: прогнозирование потребительского спроса. Робастная регрессия, функции потерь с горизонтальными асимптотами.

Тема 2.3. Метрические методы классификации и регрессии. Гипотезы компактности и непрерывности. Обобщённый метрический классификатор. Метод ближайших соседей kNN и его обобщения. Подбор числа k по критерию скользящего контроля. Метод окна Парзена с постоянной и переменной шириной окна. Метод потенциальных функций и его связь с линейной моделью классификации. Непараметрическая регрессия. Локально взвешенный метод наименьших квадратов. Ядерное сглаживание.

Оценка Надарая-Ватсона с постоянной и переменной шириной окна. Выбор функции ядра. Задача отсева выбросов. Робастная непараметрическая регрессия. Алгоритм LOWESS. Задача отбора эталонов. Понятие отступа. Алгоритм СТОЛП. Задача отбора признаков. Жадный алгоритм построения метрики.

Тема 2.4. 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. Локальная адаптация весов с регуляризацией.

УЧЕБНО-МЕТОДИЧЕСКАЯ КАРТА УЧЕБНОЙ ДИСЦИПЛИНЫ

Номер раздела, темы	Название раздела, темы	Количество аудиторных часов			Форма контроля знаний
		Лекции	Семинарские Занятия	Лабораторные занятия	
1	**Прогнозирование дискретных данных**	**12**	**8**	**8**	
1.1	Введение	4			Устный опрос
1.2	Байесовская классификация и оценивание плотности	4	4		Устный опрос
	Лабораторная работа 1.			4	Защита лабораторной работы
1.3.	Логические методы классификации	2	2		Устный опрос
	Лабораторная работа 2.			2	Защита лабораторной работы
1.4.	Метод опорных векторов	2	2		Устный опрос
	Лабораторная работа 3.			2	Защита лабораторной работы
2	**Прогнозирование непрерывных данных**	**8**	**10**	**10**	
2.1	Многомерная линейная регрессия	2	2		Устный опрос
	Лабораторная работа 4.			2	Защита лабораторной работы
2.2	Нелинейная регрессия	2	2		Устный опрос
	Лабораторная работа 5.			2	Защита лабораторной работы
2.3	Метрические методы классификации и регрессии	2	2		Устный опрос
	Лабораторная работа 6.			4	Защита лабораторной работы
2.4.	Прогнозирование временных рядов	2	2		Коллоквиум
	Лабораторная работа 7.			2	Защита лабораторной работы
ИТОГО		**20**	**18**	**18**	

ИНФОРМАЦИОННО-МЕТОДИЧЕСКАЯ ЧАСТЬ

Перечень основной литературы

1. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. Springer, 2014. — 739 p.
2. Bishop C. M. Pattern Recognition and Machine Learning. — Springer, 2006. — 738 p.
3. Мерков А. Б. Распознавание образов. Введение в методы статистического обучения. 2011. 256 с.
4. Мерков А. Б. Распознавание образов. Построение и обучение вероятностных моделей. 2014. 238 с.
5. Коэльо Л.П., Ричарт В. Построение систем машинного обучения на языке Python. 2016. 302 с.

Перечень дополнительной литературы

1. Воронцов К.В. Введение в машинное обучение. Интернет ресурс. URL: (http://www.machinelearning.ru/wiki/index.php).
2. Документация по языку Python: scikit-learn URL: https://scikit-learn.org/stable/user_guide.html

Примерный перечень тем для коллоквиумов

1. Прогнозирование временных рядов.

Рекомендуемая тематика контрольных работ

1) Контрольная работа №1. *Выбор наилучшего линейного классификатора.*
2) Контрольная работа №2. *Бинарная классификация в случае несбалансированных групп.*

Методические рекомендации по организации самостоятельной работы обучающихся

Для организации самостоятельной работы студентов магистратуры по учебной дисциплине следует использовать современные информационные технологии: разместить в сетевом доступе комплекс учебных и учебно-методически материалов (учебно-программные материалы, ссылки на учебные издания для теоретического изучения дисциплины, методические указания к лабораторным занятиям, 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 зачету, задания, тесты, вопросы для самоконтроля, тематика рефератов и др., список рекомендуемой литературы, информационных ресурсов и др.). Эффективность самостоятельной работы студентов магистратуры проверяется в ходе текущего и итогового контроля знаний. Для общей оценки качества усвоения студентами магистратуры 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; защиты отчетов по домашним заданиям, при выполнении студентами магистратуры лабораторных работ; проведение коллоквиума; защита подготовленного студентом магистратуры реферата (рефераты используются для обобщения и систематизации учебного материала; в процессе подготовки реферата студент магистратуры мобилизует и актуализирует имеющиеся умения, приобретает самостоятельно новые знания, необходимые для раскрытия темы, сопоставляя разные позиции и точки зрения).

2. Письменная форма: письменные контрольные работы по отдельным темам учебной дисциплины.

Методика формирования итоговой оценки

Формой текущей аттестации по учебной дисциплине «Методы машинного обучения» учебным планом предусмотрены зачет и экзамен.

При оценивании реферата внимание обращается на:

- содержание, корректность и последовательность изложения – 35%;
- релевантность и полноту раскрытие темы – 20 %;
- самостоятельность суждений – 35%;
- оформление – 10%.

Рекомендуется использовать рейтинговую оценку знаний студента магистратуры, 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:

- подготовка реферата – 15 %;
- работа на лабораторных занятиях – 35 %;
- контрольные работы – 25 %;
- коллоквиум – 25 %.

Итоговая оценка формируется на основе:

1) Правил проведения аттестации студентов (Постановление Министерства образования Республики Беларусь № 53 от 29 мая 2012г.);
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 УВО

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)
Методы нахождения и анализа зависимостей в данных	Математического моделирования и анализа данных	нет	Оставить содержание учебной дисциплины без изменения, протокол № 6 от 6 ноября 2018 г.

ДОПОЛНЕНИЯ И ИЗМЕНЕНИЯ К УЧЕБНОЙ ПРОГРАММЕ ПО ИЗУЧАЕМОЙ УЧЕБНОЙ ДИСЦИПЛИНЕ

на _____/_____ учебный год

№ п/п	Дополнения и изменения	Основание

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

Заведующий кафедрой

_____ _____ _____

УТВЕРЖДАЮ
Декан факультета

_____ _____ _____