Нижегородский филиал
федерального государственного автономного образовательного
учреждения высшего образования
"Национальный исследовательский университет
"Высшая школа экономики"

Факультет информатики, математики и компьютерных наук
Кафедра прикладной математики и информатики

Рабочая программа дисциплины
«анализ и разработка данных»

для образовательной программы «Программная инженерия»
направлению подготовки
09.03.04 «Программная инженерия»
уровень бакалавр

Разработчик программы:
Золотых Н.Ю. проф. каф. ПМИ

Одобрена на заседании кафедры Прикладной математики и информатики
«___»_____ 2018 г.
Зав. кафедрой В.А. Калягин _____

Утверждена Академическим советом образовательной программы
«___»_____ 2018 г., № протокола_____
Академический руководитель образовательной программы
И.С.Бычков _____

Нижний Новгород, 2018

1 Область применения и нормативные ссылки

Настоящая программа учебной дисциплины устанавливает минимальные требования к знаниям и умениям студента и определяет содержание и виды учебных занятий и отчетности. Программа предназначена для преподавателей, ведущих данную дисциплину, учебных асси стентов и студентов направления подготовки 09.03.04 «Программная инженерия», бакалавр. Программа разработана в соответствии с

- образовательным стандартом НИУ ВШЭ по направлению 09.03.04 «Программная инженерия», степень — бакалавр программной инженерии.
- Образовательной программой 09.03.04 «Программная инженерия»
- Объединенным учебным планом университета по направлению 09.03.04 «Программная инженерия», утвержденным в 2015г.

2. Цели освоения дисциплины

Целью освоения дисциплины «машинное обучение» является получение высшего профессионально профилированного (на уровне бакалавра) образования, позволяющего выпускнику успешно работать в избранной сфере деятельности, обладать универсальными и предметно-специализированными компетенциями, способствующими его социальной мобильности и устойчивости на рынке труда.

3. Компетенции обучающегося, формируемые в результате освоения дисциплины

В результате освоения дисциплины студент должен:

- Знать основные методы интеллектуального анализа данных и машинного обучения
- Уметь применять методы интеллектуального анализа данных и машинного обучения для решения практических задач
- Иметь навыки (приобрести опыт) решения задач, возникающих в некоторых прикладных областях

В результате освоения дисциплины студент осваивает компетенции:

Компетенция	Код по ОС ВШЭ	Уровень формирования компетенции	Дескрипторы – основные признаки освоения (показатели достижения результата)	Формы и методы обучения, способствующие формированию и развитию компетенции	Форма контроля уровня сформированности компетенции
Способен выполнить начальную оценку степени трудности, рисков, затрат и сформировать рабочий график	ПК7	РБ	Студент способен оценить трудность решения задачи анализа данных	Чтение лекций, проведение практических занятий, самостоятельная работа	Экзамен, лабораторные работы
Способен применять основные методы и инструменты разработки программного обеспечения	ПК 17	РБ	Студент способен разрабатывать программное обеспечение для анализа данных	Чтение лекций, проведение практических занятий, самостоятельная работа	Экзамен, лабораторные работы

Компетенция	Код по ОС ВШЭ	Уровень формирования компетенции	Дескрипторы – основные признаки освоения (показатели достижения результата)	Формы и методы обучения, способствующие формированию и развитию компетенции	Форма контроля уровня сформированности компетенции
Способен понимать стандарты и модели жизненного цикла	ПК19	РБ	Студент способен оценить параметры жизненного цикла прикладного ПО в области анализа данных.	Чтение лекций, проведение практических занятий, самостоятельная работа	Экзамен, лабораторные работы

4. Место дисциплины в структуре образовательной программы

Настоящая дисциплина относится к профессиональному циклу дисциплин, обеспечивающих подготовку бакалавра по направлению «Программная инженерия» (вариативная часть). Дисциплина проводится в 1-2 модуле на 3-м курсе.

Изучение данной дисциплины базируется на знании, полученных при освоении дисциплин: линейная алгебра и геометрия, математический анализ, дискретная математика, математическая статистика, программировании.

5. Тематический план учебной дисциплины

№	Название раздела	Всего часов	Аудиторные часы				Самостоятельная работа
			Лекции	Семинары	Практические занятия	Другие виды работы[1]	
1	Введение. Примеры практических задач	15	2		3		10
2	Вероятностная постановка задачи обучения с учителем	15	2		3		10
3	Метод наименьших квадратов	15	2		3		10
4	Статистические методы решения задач классификации	17	2		3		12
5	Нейронные сети	22	4		4		14
6	Метод опорных векторов	15	2		3		10
7	Деревья решений	18	4		4		10
8	Ансамбли решающих правил	20	4		6		10
9	Задача обучения без учителя	15	2		3		10
	ИТОГО	152	24		32		96

[1] Указать другие виды аудиторной работы студентов, если они применяются при изучении данной дисциплины.

6. Формы контроля знаний студентов

Тип контроля	Форма контроля	4 год								Кафедра/подразделение	Параметры **
		1	2	3	4	1	2	3	4		
Текущий	Контрольная работа										
	Эссе										
	Реферат										
	Коллоквиум										
	Домашнее задание	*	*								Письменная работа, программа (4-5 задач)
	Самостоятельная работа										
	Лабораторная работа										
	Проект										
	Другие формы (указать)										
Итоговый	Экзамен		*								Письменный экзамен, работа оценивается в день проведения экзамена

7. Критерии оценки знаний, навыков

Студент должен быть знаком с методами интеллектуального анализа данных и машинного обучения и приобрести опыт решения практических задач. При выполнении домашних работ, а также экзаменационной работы студент должен продемонстрировать знание теоретического материала соответствующего раздела курса, уметь правильно применять его к решению задач, грамотно формулировать ответ. Оценки по всем формам текущего контроля выставляются по 10-ти балльной шкале.

8. Содержание дисциплины

1. Введение. Примеры практических задач.

Содержательные постановки задач интеллектуального анализа данных и машинного обучения. Связь с другими областями знания и практической деятельности. Основная терминология. Примеры практических задач обучения с учителем и без учителя. Обзор учебных материалов и ресурсов Интернет по тематике дисциплины (3 часа)/ Общий объем самостоятельной работы – 10 часов, для подготовки к семинарам – 10 часов. Формы и методы проведения занятий по разделу: чтение лекций, проведение семинаров. Литература по разделу: [1] (Chapter 1, Sec. 2.1, 2.2), [2,3].

2. Вероятностная постановка задачи обучения с учителем

Регрессионная функция. Байесов классификатор. Принцип максимума апостери-орной вероятности. Метод максимального правдоподобия. (1 час)

Наивный байесовский классификатор. (1 час)

Метод ближайших соседей для задачи классификации и задачи восстановления регрессии. Теорема об оценке риска в методе ближайшего соседа. (1 часа)

Общий объем самостоятельной работы – 10 часов, для подготовки к семинарам – 8 часов, для выполнения домашней работы – 2 часа.

Формы и методы проведения занятий по разделу: чтение лекций, проведение семинаров.

Литература по разделу: [1] (Sec. 2.3, 2.4, 2.6), [2,3].

3. Метод наименьших квадратов

Метод наименьших квадратов для решения задачи восстановления регрессии. Проверка значимости и доверительные интервалы для коэффициентов (регресси-онный анализ). Анализ остатков. (1 часа)

Переобучение в задаче восстановления регрессии. Методы борьбы с переобучени-ем: выбор подмножества признаков; гребневая («ридж») регрессия (регуляриза-ция); метод «лассо». (2 часа)

Общий объем самостоятельной работы – 17 часов, для подготовки к семинарам – 15 часов, для выполнения домашней работы – 2 часа.

Формы и методы проведения занятий по разделу: чтение лекций, проведение семинаров.

Литература по разделу: [1] (Chapter 3), [2,3].

4. Статистические методы решения задач классификации

Дискриминантные и дескриптивные (описательные) методы в задаче классифика-ции. Линейный и квадратичный дискриминантный анализ. (2 часа)

Логистическая регрессия. (2 часа)

Общий объем самостоятельной работы – 17 часов, для подготовки к семинарам – 15 ча-сов, для выполнения домашней работы – 2 часа.

Формы и методы проведения занятий по разделу: чтение лекций, проведение семинаров.

Литература по разделу: [1] (Sec. 4.1–4.4), [2,3].

5. Нейронные сети

Персептрон Розенблатта. Теорема Новикова о построении разделяющей гиперплоскости. (1 часа)

Нейронная сеть. Алгоритм обратного распространения ошибки как градиентный метод. Борьба с переобучением с помощью регуляризации. (1 часа)

Представление о глубоком обучении (1 час).

Общий объем самостоятельной работы – 17 часов, для подготовки к семинарам – 15 часов, для выполнения домашней работы – 2 часа.

Формы и методы проведения занятий по разделу: чтение лекций, проведение семинаров.

Литература по разделу: [1] (Sec. 4.5.1, 11.3–11.9 раздел), [2,3].

6. Метод опорных векторов

Оптимальная разделяющая гиперплоскость. Сведение метода к задаче квадратич-ного программирования. (1 часа)

Ядра и спрямляющие пространства в методе «машина опорных векторов». (2 часа)

Общий объем самостоятельной работы – 17 часов, для подготовки к семинарам – 15 ча-са, для выполнения домашней работы – 2 часа.

Формы и методы проведения занятий по разделу: чтение лекций, проведение семинаров.

Литература по разделу: [1] (Sec. 4.5.2, 12.1–12.3), [2,3].

7. Деревья решений

Метод деревьев решений для решения задач машинного обучения. Алгоритм CART. Алгоритм C4.5. Отсечения. (3 часа)

Общий объем самостоятельной работы – 10 часов, для подготовки к семинарам – 8 ча-сов, для выполнения домашней работы – 2 часа.

Формы и методы проведения занятий по разделу: чтение лекций, проведение семинаров.

Литература по разделу: [1] (Sec. 9.2), [2,3].

8. Ансамбли решающих правил

Комбинирование слабых решающих правил. Баггинг. Бустинг. (1 час)

Алгоритм AdaBoost. (1 час)

Алгоритм Random Forest (1 час)

Алгоритм градиентного бустинга деревьев решений (gradient boosting trees). (1 час).

Общий объем самостоятельной работы – 17 часов, для подготовки к семинарам – 15 часов, для выполнения домашней работы – 2 часа.

Формы и методы проведения занятий по разделу: чтение лекций, проведение семинаров.

Литература по разделу: [1] (Chapter 10), [2,3].

9. Задача обучения без учителя

Задача обучения без учителя. Методы средних и медиан для решения задачи кластеризации. (2 часа)

Методы решения задач иерархической кластеризации. (1 час)

Алгоритм PageRank. (1 час)

Алгоритм Apriori. (1 час)

Общий объем самостоятельной работы – 17 часов, для подготовки к семинарам – 15 ча-сов, для выполнения домашней работы – 2 часов.

Формы и методы проведения занятий по разделу: чтение лекций, проведение семинаров.

Литература по разделу: [1] (Chapter 14), [2,3].

9. Образовательные технологии

[При реализации учебной работы используются следующие технологии: решение и разбор практических задач, решение практических задач на компьютере: среда R или среда Jupiter Notebook (язык программирования Python, библиотеки SciPy, NumPy, Scikit-Learn).

10. Оценочные средства для текущего контроля и аттестации студента

10.1 Оценочные средства для оценки качества освоения дисциплины в ходе текущего контроля

Примерный перечень вопросов к экзамену по всему курсу.

1. Вероятностная постановка задачи машинного обучения. Регрессионная функция. Байесов классификатор. Принцип максимума апостериорной вероятности. Метод максимального правдоподобия.
2. Наивный байесов классификатор.
3. Метод ближайших соседей для задачи классификации и задачи восстановления регрессии. Теорема об оценке риска в методе ближайшего соседа (без доказательства).
4. Метод наименьших квадратов для решения задачи восстановления регрессии. Проверка значимости и доверительные интервалы для коэффициентов (регрессионный анализ). Анализ остатков.

5. Переобучение в задаче восстановления регрессии. Методы борьбы с переобучением: выбор подмножества признаков; гребневая («ридж») регрессия (регуляризация); метод «лассо».
6. Дискриминантные и дескриптивные (описательные) методы в задаче классификации. Линейный и квадратичный дискриминантный анализ.
7. Логистическая регрессия.
8. Персептрон Розенблатта. Теорема Новикова о построении разделяющей гиперплоскости.
9. Нейронная сеть. Алгоритм обратного распространения ошибки как градиентный метод. Борьба с переобучением с помощью регуляризации.
10. Оптимальная разделяющая гиперплоскость. Сведение метода к задаче квадратичного программирования.
11. Ядра и спрямляющие пространства в методе «машина опорных векторов».
12. Метод деревьев решений для решения задач машинного обучения. Алгоритм CART. Отсечения.
13. Комбинирование слабых решающих правил. Бустинг. Алгоритм AdaBoost.
14. Алгоритм градиентного бустинга деревьев решений (gradient boosting trees).
15. Задача обучения без учителя. Методы средних и медиан для решения задачи кластеризации.
16. Методы решения задач иерархической кластеризации.

10.2 Примеры заданий промежуточной аттестации

Примерные задания для домашнего задания:

1. Дана обучающая выборка

x_1	0	1	1	0	0	1	1	2	6
x_2	3	3	1	0	1	1	2	3	1
y	0	0	0	0	1	1	1	1	1

Методом линейного дискриминантного анализа для каждого класса построить дискриминантную функцию и записать уравнение разделяющей поверхности.

2. Дана обучающая выборка (см. таблицу выше). Методом квадратичного дискриминантного анализа построить дискриминантные функции.
3. Дана обучающая выборка (см. таблицу выше). С помощью наивного байесова классификатора оценить вероятности $P(Y = 1 \mid x_1 = 1, x_1 = 2)$
4. По обучающей выборке методом наименьших квадратов построить полиномиальную модель заданной степени.
5. По обучающей выборке методом ридж-регрессии построить полиномиальную модель заданной степени с заданным параметром регуляризации.
6. Доказать, что в случае квадратичной функции потерь минимум среднему риску доставляет условное среднее. Чему равен при этом средний риск?
7. Доказать, что если функция потерь равна модулю разности, то минимум среднему риску доставляет условная медиана. Чему равен при этом средний риск?
8. Пусть ответ задается в виде аналитической функции **x XOR ((y XOR z) OR w)**, где w, x, y и z – принимают значение TRUE или FALSE. Постройте дерево решений, предсказывающее ответ с нулевой ошибкой.
9. Загрузите набор данных Spam ((http://www-stat.stanford.edu/~tibs/ElemStatLearn/). Разделите данные на обучающую и тестовую выборку (согласно меткам в файле spam.traintest). Сравните качество обучения с использованием метода опорных векторов и *K* ближайших соседей. Параметры моделей выберите на Ваше усмотрение.
10. Загрузите набор данных Spam ((http://www-stat.stanford.edu/~tibs/ElemStatLearn/). Разделите данные на обучающую и тестовую выборку (согласно меткам в файле spam.traintest). Срав-

нните качество обучения с использованием деревьев решений и метода *K* ближайших соседей. Параметры моделей выберите на Ваше усмотрение.

11. Загрузите набор данных Spam ((http://www-stat.stanford.edu/~tibs/ElemStatLearn/). Разделите данные на обучающую и тестовую выборку (согласно меткам в файле spam.traintest). Сравните качество обучения с использованием деревьев решений и метода опорных векторов. Параметры моделей выберите на Ваше усмотрение.

11. Порядок формирования оценок по дисциплине

На результирующую оценку влияют оценки за выполнение домашних заданий и экзамена ($О_{дом.з.}$ и $О_{экзамен}$ соответственно).

$$О_{накопленная} = О_{дом.з.}$$

В диплом выставляет результирующая оценка по учебной дисциплине, которая формируется по следующей формуле:

$$О_{результ} = 0.5 \cdot О_{накопленная} + 0.5 \cdot О_{экзамен}$$

Способ округления результирующей оценки по учебной дисциплине: в пользу студента.

12. Учебно-методическое и информационное обеспечение дисциплины

1.1 Базовый учебник

[1] Hastie T., Tibshirani R., Friedman J. The elements of statistical learning. Springer, 2006. 2th Ed. 2013. statweb.stanford.edu/~tibs/ElemStatLearn

1.2 Основная литература

[2] Воронцов К.В. Машинное обучение. Курс лекций. www.machinelearning.ru
[3] Золотых Н.Ю. Машинное обчение. www.uic.unn.ru/~zny/ml

1.3 Дополнительная литература

[4] James G., Witten D., Hastie T., Tibshirani R. An Introduction to Statistical Learning with Applications in R. – Springer, 2013. Рус. пер.: Джеймс Г., Уиттон Д., Хасти Т., Тибширани Р. Введение в статистическое обучение с примерами на языке R. – ДМК Пресс, 2016.
[5] Flach P. Machine Learning: The Art and Science of Algorithms That Make Sense of Data. – Campbridge University Press, 2012. Рус. пер.: Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных. – ДМК Пресс, 2016.
[6] Ng A. Machine Learning Course http://ml-class.org
[7] Bishop C.M. Pattern recognition and machine learning. Springer, 2006.
[8] Ripley B.D. Pattern recognition and neural networks. Cambridge University Press, 1996.

13. Материально-техническое обеспечение дисциплины

Для проведения лекций используется презентационное оборудование, среда статистических вычислений R, среда Jupiter Notebook (язык программирования Python, библиотеки SciPy, NumPy, Scikit-Learn).Приложение 1

Автор программы Золотых Н.Ю.