Программа курса «Анализ и разработка данных»
для образовательной программы «Прикладная математика и информатика»
уровень бакалавр

Утверждена
Академическим советом ООП
Протокол № __от «__»_____20__ г.

Автор	Золотых Н.Ю.
Число кредитов	4
Контактная работа (час.)	60
Самостоятельная работа (час.)	92
Курс	3
Формат изучения дисциплины	Без использования онлайн курса

I. ЦЕЛЬ, РЕЗУЛЬТАТЫ ОСВОЕНИЯ ДИСЦИПЛИНЫ И ПРЕРЕКВИЗИТЫ

Целью освоения дисциплины «Анализ и разработка данных» 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.

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
- оценить трудность решения задачи анализа данных
- разрабатывать программное обеспечение для анализа данных
- оценить параметры жизненного цикла прикладного ПО в области анализа данных.

Владеть

- навыками (приобрести опыт) решения задач, возникающих в прикладных областях

Настоящая дисциплина относится к профессиональному циклу дисциплин, обеспечивающих подготовку бакалавра. Изучение данной дисциплины базируется на знании, полученных при освоении дисциплин: линейная алгебра и геометрия, математический анализ, дискретная математика, математическая статистика, программировании. Полученные знания будут использованы при освоении дисциплин профессионального цикла, подготовке курсовых и выпускных квалификационных работ.

II. СОДЕРЖАНИЕ УЧЕБНОЙ ДИСЦИПЛИНЫ

1. Введение. Примеры практических задач.

Содержательные постановки задач интеллектуального анализа данных и машинного обучения. Связь с другими областями знания и практической деятельности. Основная терми-

нология. Примеры практических задач обучения с учителем и без учителя. Обзор учебных материалов и ресурсов Интернет по тематике дисциплины.

2. Вероятностная постановка задачи обучения с учителем

Регрессионная функция. Байесов классификатор. Принцип максимума апостериорной вероятности. Метод максимального правдоподобия.

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

3. Наивный байесов классификатор. Непараметрические оценки плотности вероятности. Наивный байесов классификатор.

Непараметрические оценки плотности вероятности. Окно Парзена-Розенблатта.

4. Метод опорных векторов

Оптимальная разделяющая гиперплоскость. Сведение метода к задаче квадратичного программирования. Ядра и спрямляющие пространства в методе «машина опорных векторов».

5. Деревья решений

Метод деревьев решений для решения задач машинного обучения. Алгоритм CART. (2 часа)Баггинг. Алгоритм Random Forest.

6. Ансамбли решающих правил

Комбинирование слабых решающих правил. Бустинг. Алгоритм AdaBoost.

7. Нейронные сети

Персептрон Розенблатта. Теорема Новикова о построении разделяющей гиперплоскости. Нейронная сеть. Алгоритм обратного распространения ошибки как градиентный метод. Борьба с переобучением с помощью регуляризации. Представление о глубоком обучении.

8. Глубокое обучение. Эволюция архитектур нейронных сетей. Современные применения глубокого обучения. Обзор state-of-the-art методов глубокого обучения.

9. Обучение с подкреплением

Задача обучения с подкреплением. Марковский процесс принятия решения. Уравнения-Беллмана. Exploration/Exploitation trade-off . Q-learning. SARSA. Deep Q-learning.

III. ОЦЕНИВАНИЕ

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.

На результирующую оценку влияют оценки за выполнение контрольной работы и экзамена ($О_{к.р.}$ и $О_{экзамен}$ соответственно).

$$О_{накопленная} = О_{к.р.}$$

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.

IV. ПРИМЕРЫ ОЦЕНОЧНЫХ СРЕДСТВ

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 с подкреплением. Q-learning.
16. Методы регуляризации в глубоком обучении.

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). Сравните 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.

V. РЕСУРСЫ

5.1 Основная литература

1. Ke-Lin Du Neural Networks and Statistical Learning [Электронный ресурс] / Ke-Lin Du M. N. S. Swamy, Springer 2014. Режим доступа: https://proxylibrary.hse.ru:2184/book/10.1007/978-1-4471-5571-3 - Загл. с экрана.

5.2 Дополнительная литература

1. James G. An Introduction to Statistical Learning with Applications in R. [Электронный ресурс] / – James G., Witten D., Hastie T., Tibshirani R. , Springer, 2013. Режим доступа: https://proxylibrary.hse.ru:2184/book/10.1007/978-1-4614-7138-7 - Загл. с экрана.
2. Hastie T. The elements of statistical learning [Электронный ресурс] / Hastie T., Tibshirani R., Friedman J. Springer Science+Business Media, LLC 2009, Corrected at 12th printing 2017. Режим доступа https://proxylibrary.hse.ru:2184/book/10.1007/978-0-387-84858-7 - Загл. с экрана.

Дополнительная литература для самостоятельного изучения.

1. Воронцов К.В. Машинное обучение. Курс лекций. www.machinelearning.ru
2. Золотых Н.Ю. Машинное обчение. www.uic.unn.ru/~zny/ml
3. 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.
4. Ng A. Machine Learning Course http://ml-class.org
5. Bishop C.M. Pattern recognition and machine learning. Springer, 2006.
6. Ripley B.D. Pattern recognition and neural networks. Cambridge University Press, 1996.
7. Goodfellow I., Bengio Y., Courville A. Deep Learning Book. MIT Press, 2016.
8. Sutton R., Barto A., Reinforcement Learning: An Introduction, MIT Press, 2018.

5.3 Программное обеспечение

№ п/п	Наименование	Условия доступа
	Python Software Foundation Python	свободное лицензионное соглашение

5.4 Профессиональные базы данных, информационные справочные системы, интернет-ресурсы (электронные образовательные ресурсы)

№ п/п	Наименование	Условия доступа
	Профессиональные базы данных, информационно-справочные системы	
1.	Электронно-библиотечная система Юрайт	URL: https://biblio-online.ru/
	Интернет-ресурсы (электронные образовательные ресурсы)	
1.	Открытое образование	URL: https://openedu.ru/

5.3 Материально-техническое обеспечение дисциплины

Учебные аудитории для лекционных занятий по дисциплине обеспечивают использование и демонстрацию тематических иллюстраций, соответствующих программе дисциплины в составе:

– ПЭВМ с доступом в Интернет (операционная система, офисные программы, антивирусные программы);

– мультимедийный проектор с дистанционным управлением.

Учебные аудитории для лабораторных и самостоятельных занятий по дисциплине оснащены ПК, с возможностью подключения к сети Интернет и доступом к электронной информационно-образовательной среде НИУ ВШЭ.