БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ
ИНСТИТУТ ПО МАТЕМАТИКА И ИНФОРМАТИКА

сигнатура:				
4.6	**I**	**S**	**12**	**v1**
професионално направление	код на докт. програма	вид курс (базов/спец.)	номер	версия
попълва се административно след приемане от НС на ИМИ				

Утвърдил:
(проф. дмн П. Бойваленков, Директор на ИМИ-БАН)

Учебна програма
за специализиран докторантски курс

Област на висше образование:	4. Природни науки, математика и информатика
професионално направление:	4.6. Информатика и компютърни науки
докторска програма:	Информатика
тема:	Увод в извличането на знания от данни
лектор:	доц. д-р Красимира Иванова
данни за връзка с лектора (тел., имейл)	0878966411 kivanova@math.bas.bg
хорариум:	30 часа лекции
кредити съгл. кредитната система на ЦО на БАН:	20

1. Анотация

Курсът запознава с основните принципи на функциониране на самообучаващите се интелигентни системи, както и с множество базови подходи в областта на машинното самообучение. Изучават се основните стъпки в процеса на откриване на знания от данни, както и различните базови алгоритми, използвани за решаване на задачи от областта на подготовката и преобразуването на данни; от сърцевидната част на целия процес – извличането на закономерности от данни; както и при оценката и представянето на получените резултати.

2. Необходими предварителни знания

Необходими са базови знания по структури от данни, теория на вероятностите и математическа статистика.

3. Компетентности, придобити в резултат на обучението

Успешното завършване на курса ще даде на обучаемите:

- знания за основните подходи и методи, които се използват в съвременните интелигентни софтуерни системи;
- знания за базовите алгоритми, реализиращи тези подходи, както и умения за тяхното прилагане;
- знания за основните методи за оценяване на алгоритми за решаване на задачи от областта на машинното самообучение и извличането на закономерности от данни;
- умение за провеждане на цялостен процес на откриване на знания при решаването на конкретна задача.

4. Тематично съдържание

	тема	брой часове лекции
1	Основни типове от данни от гледна точка на машинното самообучение. Примери. Множество от данни. Начини за представяне на множествата от данни, използвани от алгоритмите за машинно самообучение.	2
2	Основни стъпки в процеса на извличане на знания. Стандарти за процеси за извличане на знания от данни – CRISP-DM и ASUM-DM.	2
3	Предварителна обработка на данните – част 1. Липсващи данни, шумни данни, излишни данни, несъгласувани данни – характеристики, видове, начини за отстраняване.	2
4	Предварителна обработка на данните – част 2. Интеграция и трансформация на данните. Техники за намаляване на размерността. Техники за намаляване на разнообразието.	2
5	Откриване на знания в данни (Data Mining). Предмет и основни задачи: класификация, регресия, клъстеризация, асоциативни правила, откриване на изключения.	2
6	Машинно самообучение с учител и смесено самообучение – характеристики и основни видове.	2

7	Основни групи класификационни алгоритми – характеристики, специфични особености.	2
8	Ансамблови методи – същност и характеристики. Видове. Приложение.	2
9	Регресия – характеристика. Видове. Приложение.	2
10	Машинно самообучение без учител. Клъстеризация. Основни методи за откриване на клъстери (нейерархични и йерархични). Методи за оценка на откритите клъстери.	2
11	Откриване на асоциативни правила. Основни понятия. Методи за откриване на асоциативни правила.	2
12	Метрики и мерки за близост – дефиниции, видове, приложение.	2
13	Оценка на резултатите – основни понятия. Методи за постигане на стабилност. Методи за сравнение.	2
14	Извличане на знания от текст. Характеристика и основни методи. Основни стъпки.	2
15	Опасности при прилагането на машинно самообучение. Видове отклонения при избор на проби. Етика и изкуствен интелект.	2

5. Конспект

1. Основни типове от данни от гледна точка на машинното самообучение. Примери. Множество от данни. Начини за представяне на множествата от данни, използвани от алгоритмите за машинно самообучение.
2. Основни стъпки в процеса на извличане на знания. Стандарти за процеси за извличане на знания от данни – CRISP-DM и ASUM-DM.
3. Предварителна обработка на данните – част 1. Липсващи данни, шумни данни, излишни данни, несъгласувани данни – характеристики, видове, начини за отстраняване.
4. Предварителна обработка на данните – част 2. Интеграция и трансформация на данните. Техники за намаляване на размерността. Техники за намаляване на разнообразието.
5. Откриване на знания в данни (Data Mining). Предмет и основни задачи: класификация, регресия, клъстеризация, асоциативни правила, откриване на изключения.
6. Машинно самообучение с учител и смесено самообучение – характеристики и основни видове.
7. Основни групи класификационни алгоритми – характеристики, специфични особености.
8. Ансамблови методи – същност и характеристики. Видове. Приложение.
9. Регресия – характеристика. Видове. Приложение.
10. Машинно самообучение без учител. Клъстеризация. Основни методи за откриване на клъстери (нейерархични и йерархични). Методи за оценка на откритите клъстери.
11. Откриване на асоциативни правила. Основни понятия. Методи за откриване на асоциативни правила.

12. Метрики и мерки за близост – дефиниции, видове, приложение.
13. Оценка на резултатите – основни понятия. Методи за постигане на стабилност. Методи за сравнение.
14. Извличане на знания от текст. Характеристика и основни методи. Основни стъпки.
15. Опасности при прилагането на машинно самообучение. Видове отклонения при избор на проби. Етика и изкуствен интелект.

6. Препоръчителна литература:

1. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. Proc. of the ACM SIGMOD Int. Conf. on Management of Data, Washington, DC, 1993, pp. 207-216.
2. Baer, T.: Understand, Manage, and Prevent Algorithmic Bias: A Guide for Business Users and Data Scientists. Apress, 2019.
3. Boriah, S., Chandola, V., Kumar, V.: Similarity Measures for Categorical Data: A Comparative Evaluation, In Proceedings of 2008 SIAM Data Mining Conference, 2008, Atlanta, pp. 243-254.
4. Cultural Analytics, Lev Manovich – several examples of cultural data analysis, http://lab.softwarestudies.com/
5. Demsar, J., Zupan, B., Leban, G., Curk, T.: Orange: from experimental machine learning to interactive data mining. White Paper (www.ailab.si/orange), Faculty of Computer and Information Science, University of Ljubljana, 2004.
6. Deza, M.-M., Deza, E.: Encyclopedia of Distances. Springer, XIV, 2009, 590 p.
7. EUR-Lex: Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206
8. Gheyas, I., Smith, L.: Feature subset selection in large dimensionality domains. Elsevier, Pattern Recognition, 43 (1), 2010, pp. 5-13.
9. Han, J., Kamber, M.: Data Mining: Concepts and Techniques, Second ed., Morgan Kaufmann Publishers, 2006, 772 p.
10. Ivanova Kr., M. Dobreva, P. Stanchev, G. Totkov (Eds): Access to Digital Cultural Heritage: Innovative Applications of Automated Metadata Generation. ISBN: 978-954-423-722-6, http://www.math.bas.bg/infres/book-ADCH/index.htm
11. Maimon, O., Rokach, L.: Decomposition Methodology for Knowledge Discovery and Data Mining. Vol. 61 of Series in Machine Perception and Artificial Intelligence. World Scientific Press, 2005.
12. Molenberghs, G, Fitzmaurice, G., Kenward, M.G., Tsiatis, A., Verbeke, G. (eds.): Handbook of Missing Data Methodology, Chapman & Hall, 2015.
13. Witten I., E. Frank, M. Hall, Ch. Pal. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Series in Data Management Systems, 4th Edition, ISBN-13: 978-0128042915, https://www.cs.waikato.ac.nz/ml/weka/book.html
14. Zaiane, O., Antonie, M.-L.: On pruning and tuning rules for associative classifiers. In Proc. of Int. Conf. on Knowledge-Based Intelligence Information & Engineering Systems, LNCS, Vol. 3683, 2005, pp.966-973.

7. Ресурсно осигуряване на обучението:

Езици и системи за анализ на данни и извличане на знания:

– Python или R Programming Language

– Mathematica или MATLAB

– RapidMiner, WEKA, ORANGE

Хранилища с примерни данни:

– UCI Machine Learning Repository (*https://archive.ics.uci.edu/ml/datasets.php*);

– Kaggle (https://www.kaggle.com/datasets)...

8. Критерии за оценка

Изпитът е с продължителност 4 часа и се състои от две части – писмен и устен.
На писмения изпит докторантът развива своите идеи и концепции по два въпроса от конспекта.
Крайната оценка е от 2 до 6 (с точност до 0.5).
Тя се формира на базата на следното съответствие:

Отличен (6 или 5.50)	Отлично владее материала. Изложението е изчерпателно, последователно, компетентно, логично и хармонично. Правилно обосновава предлаганите решения, знае как да обобщава и излага материала без да прави грешки. Притежава необходимите умения за изпълнение на практически задачи.
Мн. добър (5 или 4.50)	Познава материала. Излага го правилно без да допуска съществени неточности. Може правилно да прилага теоретични принципи и притежава необходимите умения за изпълнение на практически задачи.
Добър (4 или 3.50)	Владее голяма част материала, но допуска неточности при изложението и отговорите на въпросите. Има известни неясноти при опитите за прилагане на материала в практически ситуации.
Среден (3)	Владее само част от материала, но се затруднява в отделните детайли. Допуска неточности във формулировките и нарушава последователността при представянето на материал. Има затруднения при изпълнение на практически задачи.
Слаб (2)	Не познава значителна част от материала, допуска съществени грешки и с големи трудности изпълнява практически задачи.

Учебната програма е обсъдена и одобрена на заседание на секция „Софтуерни технологии и информационни системи“,

на 12.10.2021 г.

Ръководител секция: _____

(доц. д-р Красимира Иванова)

Разгледана от Директорския съвет на ИМИ-БАН на 26.11.2021 г. (протокол № 46).

Приета от Научния съвет на ИМИ-БАН на 17.12.2021 г. (протокол № 20).