МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«ДАГЕСТАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»
Факультет математики и компьютерных наук
Кафедра прикладной математики

ФОНД ОЦЕНОЧНЫХ СРЕДСТВ
по дисциплине
«Машинное обучение»

Кафедра дискретной математики и информатики
факультета математики и компьютерных наук

Образовательная программа бакалавриата
01.03.05 – Статистика

Направленность (профиль) программы
Анализ больших данных

Форма обучения
Очная

Статус дисциплины:

Входит в обязательную часть ОПОП

Махачкала, 2023

Фонд оценочных средств по дисциплине «Машинное обучение» составлена в 2023 году в соответствии с требованиями ФГОС ВО бакалавриата по направлению подготовки 01.03.05 - статистика от 14.08.2020 г. № 1032

Разработчики:

кафедра дискретной математики и информатики, преподаватель Ибавов Т.И.

Фонд оценочных средств по дисциплине «Машинное обучение» одобрен:
на заседании кафедры дискретной математики и информатики от
«20» января 2023 г., протокол № 5.

Зав. кафедрой _____ Магомедов А.М.

на заседании Методической комиссии факультета математики и компьютерных наук от
«25» января 2023 г., протокол № 4.

Председатель _____ Ризаев М.К.

Фонд оценочных средств «Машинное обучение» согласован с учебно-методическим управлением

«20» февраля 2023 г. _____

1. ПАСПОРТ
ФОНДА ОЦЕНОЧНЫХ СРЕДСТВ
по дисциплине
«Машинное обучение»

1.1. Основные сведения о дисциплине

Общая трудоемкость дисциплины составляет 3 зачетные единицы (108 академических часа).

Вид работы	Трудоемкость, академических часов		
	2[и] семестр	___семестр	всего
Общая трудоёмкость	108		108
Контактная работа:	30		30
Лекции (Л)	16		16
Лбораторные занятия (ЛЗ)	32		32
Консультации			
Промежуточная аттестация (зачет, экзамен)	зачет		
Самостоятельная работа 1. работа с лекционным материалом, с учебнойлитературой 2. опережающая самостоятельная работа (изучение новогоматериала до его изложения на занятиях) 3. выполнение домашних заданий, домашних контрольных работ 4. подготовка к лабораторным работам, к практическим и семинарским занятиям 5. подготовка к контрольным работам, коллоквиумам	10 5 15 15 15		10 5 15 15 15

1.2. Требования к результатам обучения по дисциплине, формы ихконтроля и виды оценочных средств

ПАСПОРТ ФОНДА ОЦЕНОЧНЫХ СРЕДСТВ
*по дисциплине «*Машинное обучение*»*

№п/п	Контролируемые модули, разделы (темы) дисциплины	Код контролируемой компетенции (или её части)	Оценочные средства		Способ контроля
			наименование	№№ заданий	
1	Понятия статистической совокупности, статистических показателей и	ПК-1	Вопросы для собеседования	1-17	устно
		ПК-1	Контрольные	1-2	письменно

	средних величин		работы		
2	Показатели вариации, корреляционной связи в статистическом ряду	ПК-1	Вопросы для собеседования	18-30	устно
		ПК-1	Тестовые задания	3	письменно
		ПК-1	Контрольные работы	3-5	письменно

1.3. Показатели и критерии определения уровня сформированности компетенций

№ п/п	Код компетенции	Уровни сформированности компетенции			
		Недостаточный	Удовлетворительный (достаточный)	Базовый	Повышенный
	ПК-1	Не знает методы сбора и обработки данных, полученными в области математических и естественных наук, программирования и информационных технологий для формирования выводов по соответствующим научным исследованиям Не умеет собирать и обрабатывать данные, полученные в области математических и естественных наук, программирования и информационных технологий для формирования выводов по соответствующим научным исследованиям. Не владеет навыками сбора и обработки данных, полученными в области математических и (или) естественных наук, программирования и информационных технологий для	Знает на достаточном уровне методы сбора и обработки данных, полученными в области математических и естественных наук, программирования и информационных технологий для формирования выводов по соответствующим научным исследованиям Умеет на достаточном уровне собирать и обрабатывать данные, полученные в области математических и естественных наук, программирования и информационных технологий для формирования выводов по соответствующим научным исследованиям. Владеет на достаточном уровне навыками сбора и обработки данных, полученными в области математических и (или) естественных наук, программирования и информационных технологий для формирования выводов по соответствующим научным исследованиям.	Хорошо знает методы сбора и обработки данных, полученными в области математических и естественных наук, программирования и информационных технологий для формирования выводов по соответствующим научным исследованиям Хорошо умеет собирать и обрабатывать данные, полученные в области математических и естественных наук, программирования и информационных технологий для формирования выводов по соответствующим научным исследованиям. Хорошо владеет навыками сбора и обработки данных, полученными в области математических и (или) естественных наук, программирования и информационных технологий для формирования выводов по соответствующим научным исследованиям.	Отлично знает методы сбора и обработки данных, полученными в области математических и естественных наук, программирования и информационных технологий для формирования выводов по соответствующим научным исследованиям Отлично умеет собирать и обрабатывать данные, полученные в области математических и естественных наук, программирования и информационных технологий для формирования выводов по соответствующим научным исследованиям.

		формирования выводов по соответствующим научным исследованиям.			Отлично владеет навыками сбора и обработки данных, полученными в области математических и (или) естественных наук, программирования и информационных технологий для формирования выводов по соответствующим научным исследованиям.

2. КОНТРОЛЬНЫЕ ЗАДАНИЯ И ИНЫЕ МАТЕРИАЛЫ ОЦЕНКИ знаний, умений, навыков и (или) опыта деятельности, характеризующие этапы формирования компетенций в процессе освоения дисциплины

«Машинное обучение»

В рамках данного задания требуется выполнить 5 задач. Каждая задача должна быть оформлена в виде отдельного task{1,2,3,4,5}.ipynb файла и tensorboard{1,2,3,4,5}.zip файла. В каждом файле *.ipynb должно быть:

- построение архитектуры;
- выполнен процесс обучения;
- показан пример работы модели до обучения и после;

Файл .zip должен содержать результаты эксперимента в формате tensorboard для каждой из задач:

- для каждого набора параметров свой график зависимости качества от обучения (если требуется в задаче);
- примеры работы модели в процессе обучения модели.

Для каждой задачи должны быть представлены выводы:

- какой результат ожидали;
- какой не ожидали;
- что было не ясно.

Код и эксперимент должен быть понятным внешнему читателю:

- В коде должны быть коментарии;
- Названия переменных должно быть интерпретируемые.

Рекомендуется все вычисления проводить на google colab в режиме cuda.

Рекомендуется использовать backup моделей при обучении на google drive.

Задача 1. Распознавания именованных сущностей на основе fasttext

Построить модель расспознавания именованных сущностей на русском языке. В качестве данных использовать выборку NERUS (NER).

- В качестве векторного представления токенов использовать fasttext модель;
- В качестве модели использовать модель LSTM;
- Архитектуру LSTM можно выбрать произвольным образом;
- Весь процесс обучения должен быть визуализирован в tensorboard (метрики качества и пример предсказания)

Выборку можно взять из github.

Для экономинии памяти компютера предлагается воспользоваться сжатием модели fasttext с 300-мерного к 100-мерному (на колаб не хватит оперативки на сжатие до 100-мерного вектора, поэтому работайте сразу с 300-мерными в VEC формате). А также использовать выполнить переопределения модели fasttext в VEC модель.

Задача 2. Классификация даты документа

Построить модель для классификации даты (года) публикации новостной заметки из выборки lenta.ru.

- В качестве векторного представления текста рассмотреть тематический вектор.
- В качестве классификатора использовать любой классификатор на ваш выбор.
- Проанализировать качество классификации в зависимости от добавленных модальностей.
- Провести эксперимент по добавлению регуляризаторов.
- Провести анализ классификации модальности(год рассмотреть как модальность) при помощи встроенных методов bigartm.

Задача 3. Posterior Sampling в задаче RL

Большая задача на разбор статьи. Требуется решить проблемы "Задачи о заплыве" связанные с тем, что алгоритм не доходит до левого края и начинает всегда скатываться по течению.

Реализируйте метод Posterior Sampling из статьи.

Задача 4. Анализ модели CNN

Провести анализ качества аппроксимации выборки EMNIST-letters моделью сверточной нейронной сети в зависимости от:

- размера ядра (можно ввести ограничение, что на каждом слое размер ядра одинаковый);
- числа слоев;
- от пулинга;
- добавления BatchNorm;
- параметра dropout.

Все выводы должны быть представленны в формате tensorboard (каждый набор параметров, свой график, пример)

Выборку можно взять из torchvision.

Если не работает скачиваение EMNIST использовать FahionMnist.

Пояснение: В данном задании важно продемонстрировать умение строить различные структуры модели CNN. Не обязательно выполнять перебор всех вариантов нейросети (проходить по сетке гиперпараметров), но описание экспериментов должны присутвовать.

Задача 5. Анализ модели LSTM

Провести анализ качества аппроксимации выборки NERUS (предсказание POS tag для токенов) моделью LSTM в зависимости от:

- размера слоя;
- числа слоев;
- параметра dropout;
- добавления BatchNorm;
- размера словаря;
- *токенизатора* - дополнительное задание (со звездочкой).

Все выводы должны быть представленны в формате tensorboard (каждый набор параметров, свой график, пример --- второй семинар).

Выборку можно взять из github.

Предлагается использовать разные варианты токенизатора:

- взять все слова из обучающего датасета;
- использовать предобученые BPE токены из LaBSE модели (пока не сильно важно что это, об этом 4й семинар):

```
from transformers import AutoTokenizer, AutoModel

tokenizer = AutoTokenizer.from_pretrained("sentence-transformers/LaBSE")

model = YourLSTMmodel()
tokens = tokenizer(['Hello Mr. Bernz.', 'I am Homer Simpsone'], padding=True,
                   truncation=True,
                   max_length=510,
                   return_tensors='pt')['tokens_ids']
answer = model(tokens)
```

Сначала выборку нужно привести формат согласно Вашему токенизатору, то есть выполнить отображение исходной выборки с токенами в исходном формате в выборку с токенами, которые согласованы с Вашим токенизатором.

Пояснение: В данном задании важно продемонстрировать умение работать с моделью LSTM, а также умение преобразовать данные под разные модели и данные. В качестве базового решения продемостировать аппроксимации "чистой" выборки NERUS без преобразования данных (взять исходные токены из выборки). Более сложным является задание, когда Вам дают другой токенизатор предложения и доступные данные нужно переформатировать в нужный Вам формат данных.

Задача 6. Модель автокодировщика

Провести анализ модели автокодировщика (не вариационного) для выборки Twitter (эмбединги предложений). Требуется сравнить качество востановления предложения в зависимости от:

- размера слоя;
- числа слоев;
- параметра dropout;
- добавления BatchNorm;
- размера словаря;

- *токенизатора* - дополнительное задание (со звездочкой.

Все выводы должны быть представленны в формате tensorboard (каждый набор параметров, свой график, пример --- второй семинар).

Рекомендуется использовать предобученый BPE токенизатор для снижения размерности словаря (см. задачу 5).

Задача 7. Вариационный автокодировщик

Провести синтетический эксперимент с моделью вариационного автокодировщика в случае, если данные не из бернуливского распределения, а из нормального. В качестве данных использовать синтетическую выборку, которая состоит из нескольких кластеров в виде гаусиан (каждый кластер является множеством векторов из нормального распределения с парметрами mu, Sigma). В рамках эксперимента требуется исследовать:

- зависимость качества востановления от размера скрытого представления;
- зависимость качества востановления от размера исходного пространства;
- зависимость качества востановления от отношения размера скрытого представления к исходном;
- зависимость качества востановления от сложности модели нейросети.

Все выводы должны быть представленны в формате tensorboard (каждый набор параметров, свой график, пример --- второй семинар).

P.S. в рамках семинара мы востанавливали параметры бернуливского распределения, так как изображение это числа от 0 до 1 --- вероятности бернуливской случайной величины. Теперь требуется, чтобы модель decoder востанаввливала параметры нормального случайного вектора.

P.S.S. в качестве модели encoder и decoder можно выбирать любую архитектуру нейросети.

Задача 8. Генерация аннотации к изображению

Требуется построить модель генерации описания изображения по изображению. В качестве выборки рассматривается подвыборка вывборки COCO. Требуется в качестве модели encoder использовать предобученую модель resnet152 без последнего слоя, в качестве модели decoder обучить LSTM модель.

Все выводы должны быть представленны в формате tensorboard (показать, как менялись описания одного и того же изображения при обучении модели, а также график качества в зависимости от итерации).

Рекомендуется взять подвыборку общей выборки из сайта COCO.

Критерии оценки:

- оценка «отлично» выставляется студенту, если верно и правильно выполнено 90%-100% заданий;
- оценка «хорошо» выставляется студенту, если верно и правильно выполнено 70%-80% заданий;
- оценка «удовлетворительно» выставляется студенту, если верно и правильно решено 50%-60% заданий, возможны некоторые исправления при решении;
- оценка «неудовлетворительно» выставляется студенту, если верно выполнено менее 50% заданий;

Вопросы для коллоквиумов, собеседования

Выбор модели и отбор признаков

- В чём отличия внутренних и внешних критериев?
- Разновидности внешних критериев.
- Разновидности критерия скользящего контроля.
- Что такое критерий непротиворечивости? В чём его недостатки?
- Что такое многоступенчатый выбор модели по совокупности критериев?
- Основная идея отбора признаков методом полного перебора. Действительно ли это полный перебор?
- Основная идея отбора признаков методом добавлений и исключнений.
- Что такое шаговая регрессия? Можно ли её использовать для классификации, в каком методе?
- Основная идея отбора признаков методом поиска в глубину.
- Основная идея отбора признаков методом поиска в ширину.
- Что такое МГУА?
- Основная идея отбора признаков с помощью генетического алгоритма.
- Основная идея отбора признаков с помощью случайного поиска.
- В чём отличия случайного поиска от случайного поиска с адаптацией?

Нейронные сети

- Приведите пример выборки, которую невозможно классифицировать без ошибок с помощью линейного алгоритма классификации. Какова минимальная длина выборки, обладающая данным свойством? Какие существуют способы модифицировать линейный алгоритм так, чтобы данная выборка стала линейно разделимой?

- Почему любая булева функция представима в виде нейронной сети? Сколько в ней слоёв?
- Метод обратного распространения ошибок. Основная идея. Основные недостатки и способы их устранения.
- Как можно выбирать начальное приближение в градиентных методах настройки нейронных сетей?
- Как можно ускорить сходимость в градиентных методах настройки нейронных сетей?
- Что такое диагональный метод Левенберга-Марквардта?
- Что такое «паралич» сети, и как его избежать?
- Как выбирать число слоёв в градиентных методах настройки нейронных сетей?
- Как выбирать число нейронов скрытого слоя в градиентных методах настройки нейронных сетей?
- В чём заключается метод оптимального прореживания нейронной сети? Какие недостатки стандартного алгоритма обратного распространения ошибок позволяет устранить метод ODB?

Композиции алгоритмов классификации

- Дать определение алгоритмической композиции (помнить формулу). Какие типы корректирующих операций вы знаете?
- Какие типы голосования вы знаете? Какой из них наиболее общий? (помнить формулу)
- Как обнаружить объекты-выбросы при построении композиции классификаторов для голосования по большинству?
- Как обеспечивается различность базовых алгоритмов при голосовании по большинству?
- Как обеспечивается различность базовых алгоритмов при голосовании по старшинству?
- Какие возможны стратегии выбора классов базовых алгоритмов при голосовании по старшинству?
- Какие две эвристики лежат в основе алгоритма AdaBoost?
- Как обнаружить объекты-выбросы в алгоритме AdaBoost?
- Достоинства и недостатки алгоритма AdaBoost.
- Основная идея алгоритма AnyBoost.
- Основная идея метода bagging.
- Основная идея метода случайных подпространств.
- Что такое смесь экспертов (помнить формулу)?
- Приведите примеры выпуклых функций потерь. Почему свойство выпуклости помогает строить смеси экспертов?

Логические алгоритмы классификации

- Что такое логическая закономерность? Приведите примеры закономерностей в задаче распознавания спама.
- Часто используемые типы логических закономерностей.
- Дайте определение эпсилон-дельта-логической закономерности (помнить формулы).
- Дайте определение статистической закономерности (помнить формулы).
- Сравните области статистических и логических закономерностей в (p,n)-плоскости.
- С какой целью делается бинаризация?

- В чём заключается процедура бинаризации признака?
- Как происходит перебор в жадном алгоритме синтеза информативных конъюнкций?
- Какие критерии информативности используются в жадном алгоритме синтеза информативных конъюнкций и почему?
- Как приспособить жадный алгоритм синтеза конъюнкций для синтеза информативных шаров?
- Что такое стохастический локальный поиск?
- В чём отличия редукции и стабилизации? В чём их достоинства и недостатки?
- Что такое решающий список?
- Какие критерии информативности используются при синтезе решающего списка и почему?
- Достоинства и недостатки решающих списков.
- Что такое решающее дерево?
- Какие критерии информативности используются при синтезе решающего дерева и почему?
- Достоинства и недостатки решающих деревьев.
- Зачем делается редукция решающих деревьев?
- Какие есть два основных типа редукции решающих деревьев?
- Как преобразовать решающее дерево в решающий список, и зачем это делается?
- Что такое ADT (alternating decision tree)? Как происходит построение ADT?
- Основная идея алгоритма КОРА.
- Почему возникает проблема предпочтения признаков с меньшими номерами в алгоритме КОРА? Как она решается?
- Основная идея алгоритма ТЭМП.
- Какие критерии информативности используются в алгоритме ТЭМП и почему?
- Почему возникает проблема дублирования закономерностей в алгоритме ТЭМП? Как она решается?
- Достоинства и недостатки алгоритма ТЭМП.
- Как использовать алгоритм AdaBoost для построения взвешенного голосования закономерностей?
- Какой критерий информативности используется в алгоритме AdaBoost?
- Структура алгоритма вычисления оценок (ABO).
- Что такое ассоциативное правило? Приведите пример ассоциативного правила в задаче анализа потребительских корзин.
- Основная идея алгоритма поиска ассоциативных правил APriory.

Кластеризация и таксономия

- Каковы основные цели кластеризации?
- Основные типы кластерных структур. Приведите для каждой из этих структур пример алгоритма кластеризации, который для неё НЕ подходит.
- В чём заключается алгоритм кратчайшего незамкнутого пути? Как его использовать для кластеризации? Как с его помощью определить число кластеров? Всегда ли это возможно?

- Основная идея алгоритма ФорЭл.
- Как вычисляются центры кластеров в алгоритме ФорЭл, если объекты — элементы метрического (не обязательно линейного векторного) пространства?
- Какие существуют функционалы качества кластеризации и для чего они применяются?
- Основные отличия алгоритма k-средних и EM-алгоритма. Кто из них лучше и почему?
- Основная идея иерархического алгоритма Ланса-Вильямса.
- Какие основные типы расстояний между кластерами применяются в алгоритме Ланса-Вильямса?
- Какие расстояния между кластерами, применяемые в алгоритме Ланса-Вильямса, лучше и почему?
- Что такое дендрограмма? Всегда ли её можно построить?
- Какой функционал качества оптимизируется сетью Кохонена? (помнить формулу)
- В чем отличия правил мягкой и жёсткой конкуренции? В чём преимущества мягкой конкуренции?
- Как устроена самооганизующаяся карта Кохоненеа?
- Как интерпретируются карты Кохонена?
- Почему задачи с частичным обучением выделены в отдельный класс? Приведите примеры, когда методы классификации и кластеризации дают неадекватное решение задачи с частичным обучением.
- Как приспособить графовые алгоритмы кластеризации для решения задачи с частичным обучением?
- Как приспособить EM-алгоритм для решения задачи с частичным обучением?
- Какие способы решения задачи с частичным обучением Вы знаете?

Критерии оценки:

- оценка «отлично» выставляется студенту, если изложение полученных знаний в устной форме полное, в системе, в соответствии с требованиями учебной программы; допускаются единичные несущественные ошибки, самостоятельно исправляемые учащимися;
- оценка «хорошо» выставляется студенту, если изложение полученных знаний в устной форме полное, в системе, в соответствии с требованиями учебной программы; допускаются, отдельные несущественные ошибки, исправляемые учащимися после указания преподавателя на них;
- оценка «удовлетворительно» выставляется студенту, если изложение полученных знаний неполное, однако это не препятствует усвоению последующего программного материала; допускаются отдельные существенные ошибки, исправляемые с помощью преподавателя;
- оценка «неудовлетворительно» выставляется студенту, если изложение учебного материала неполное, бессистемное, что препятствует усвоению последующей учебной информации; существенные ошибки, не исправляемые даже с помощью преподавателя;

Комплект тестовых заданий для контроля

Что, из ниже перечисленного, относится к обучающей выборке?

Ответ:

(1) классификация данных

(2) объекты с известными ответами

(3) алгоритм решающий функцию

Номер 2
Объекты состоят из признаков?

Ответ:
(1) Да

(2) Нет

Номер 3
Что называют данными в машинном обучении?

Ответ:
(1) матрицы

(2) объекты

(3) признаки

(4) алгоритм

(5) функция

Упражнение 2:

Номер 1
Выберите правильный ответ. Задача классификации - это:

Ответ:
(1) множество объектов, разделенных на классы

(2) исследование влияние одного или нескольких признаков на объект

(3) определение порядка признака согласно рангу

Номер 2
Выберите правильный ответ. Задача регрессии - это:

Ответ:
(1) множество объектов, разделенных на классы

(2) исследование влияние одного или нескольких признаков на объект

(3) определение порядка признака согласно рангу

Номер 3
Выберите правильный ответ. Задача ранжирования - это:

Ответ:

(1) множество объектов, разделенных на классы

(2) исследование влияние одного или нескольких признаков на объект

(3) определение порядка признака согласно рангу

Упражнение 3:

Номер 1
Что служит индикатором ошибки для задач классификации?

Ответ:
(1) $\varphi(a,x) = [a(x) \neq y^(*)(x)]$

(2) $\varphi(a,x) = | a(x) - y^(*)(x) |$

(3) $\varphi(a,x) = (a(x) - y^(*)(x))^(2)$

Номер 2
Как формула подходит для абсолютного значения ошибки для задач регрессии?

Ответ:
(1) $\varphi(a,x) = | a(x) - y^(*)(x) |$

(2) $\varphi(a,x) = (a(x) - y^(*)(x))^(2)$

(3) $\varphi(a,x) = [a(x) \neq y^(*)(x)]$

Номер 3
Что является квадратичной ошибкой для задачи регрессии?

Ответ:
(1) $\varphi(a,x) = [a(x) \neq y^(*)(x)]$

(2) $\varphi(a,x) = (a(x) - y^(*)(x))^(2)$

(3) $\varphi(a,x) =| a(x) - y^(*)(x) |$

Упражнение 4:

Номер 1
Эмпирический риск - это средняя потеря на одном объекте.

Ответ:
(1) Да

(2) Нет

Номер 2
Если происходит средняя потеря на всех объектах, то это есть:

Ответ:
(1) переобучение

(2) эмпирический риск

(3) оценка релевантности

Номер 3
Верно ли утверждение? Всякая оптимизация по неполной информации и избыточная сложность параметров приводит в переобучению.

Ответ:
(1) Да

(2) Нет

Упражнение 5:

Номер 1
Выберите верные утверждения.

Ответ:
(1) класс - это множество всех объектов с определенным значением.

(2) в задачах регрессии допустимым ответом является действительное число или числовой вектор.

(3) в задачах ранжирования ответы получают сразу на множестве объектов.

(4) области минимального объёма с достаточно гладкой границей являются основной составляющей задач ранжирования

Номер 2
Верно ли следующее утверждение? Многие виды задач медицинской диагностики решаются задачами классификации.

Ответ:
(1) Да

(2) Нет

Номер 3
В задачах классификации признаки могут быть строковыми, вещественными, числовыми.

Ответ:
(1) Да

(2) Нет

Упражнение 6:

Номер 1
Какие задачи из ниже перечисленных относятся к задачам классификации?

Ответ:
(1) определение наиболее целесообразного способа лечения;

(2) определение длительности и исхода заболевания;

(3) оценивание кредитоспособности заёмщика;

(4) задачи поискового вывода

Номер 2
Какие задачи, из ниже перечисленных, являются задачами ранжирования?

Ответ:

(1) обнаружение спама

(2) задачи поискового вывода;

(3) определение наиболее целесообразного способа лечения;

Номер 3
Какие задачи, из ниже перечисленных, являются задачами прогнозирования?

Ответ:
(1) математический прогноз даты сильных землетрясений;

(2) определение длительности и исхода заболевания;

(3) обнаружение спама;

(4) прогнозирование вероятности летельного исхода;

(5) задачи поискового вывода.

Упражнение 7:

Номер 1
Какая, из ниже перечисленных задач, является задачей классификации на 4

класса?

Ответ:

(1) $Y = \{0, 1\}^M$

(2) $Y = \{0, 1\}$

(3) $Y = \{-1; +1\}$

(4) $Y = \{1, 2, 3, 4\}$

Номер 2

Какой пример подходит для задачи восстановления регрессии?

Ответ:

(1) $Y = \{0, 1\}^M$

(2) $Y = R^m$

(3) $Y = \{-1; +1\}$

(4) $Y = \{1, 2, 3, 4\}$

Номер 3

Какие, из ниже перечисленных задач, являются задачами классификации?

Ответ:

(1) $Y = \{0, 1\}^M$

(2) $Y = R$

(3) $Y = \{-1; +1\}$

(4) $Y = R^m$

(5) $Y = \{1, 2, 3, 4\}$

Упражнение 8:

Номер 1
Какой тип экспериментального исследования имеет цель - понимание, на что влияют параметры метода обучения?

Ответ:
(1) исследование задач ранжирования

(2) исследование задач классификации

(3) исследование на модельных данных

Номер 2
Какой тип экспериментального исследования имеет цель - либо решение конкретной прикладной задачи, либо выявление «слабых мест»?

Ответ:
(1) исследование задач ранжирования

(2) исследование на реальных данных

(3) исследование на модельных данных

Номер 3
Что, из ниже перечисленного, не относится к типу экспериментального исследования?

Ответ:
(1) исследование задач ранжирования

(2) исследование на реальных данных

(3) исследование на модельных данных

Критерии оценки:

- оценка «отлично» выставляется студенту, если верно и правильно выполнено 90%-100% заданий;
- оценка «хорошо» выставляется студенту, если верно и правильно выполнено 70%-80% заданий;
- оценка «удовлетворительно» выставляется студенту, если верно и правильно решено 50%-60% заданий, возможны некоторые исправления при решении;
- оценка «неудовлетворительно» выставляется студенту, если верно выполнено менее 50% заданий;

Темы эссе (рефератов, докладов, сообщений)

1. Тема: Системы автоматизации проектных работ (САПР).
2. Тема: Экспертные системы, их применение для решения задач различных предметных областей.
3. Тема: Системы искусственного интеллекта, классификация, особенности.
4. Тема: Роль автоматизированных систем поддержки принятия решений в управлении экономическими объектами.
5. Тема: Области применения нейронных сетей, классы задач, решаемых благодаря их использованию.
6. Тема: Формализация и структурирование знаний при проектировании баз знаний. Модели знаний.
7. Тема: Автоматизированные информационные технологии и системы для интеллектуальной поддержки финансового управления и проведения финансового анализа состояния предприятия.
8. Тема: Назначение и области применения правовых информационно – поисковых справочных систем.
9. Тема: Электронные программы – словари.
10. Тема: Программы перевода текстов с одних языков на другие.
11. Тема: Инструментальные средства и языки программирования, применяемые для разработки систем искусственного интеллекта.
12. Тема: Общая характеристика классов задач, решаемых с помощью систем искусственного интеллекта.
13. Тема: Общая характеристика и основные компоненты автоматизированных систем поддержки принятия решений модельного типа.
14. Тема: Гипертекстовые поисковые Internet – системы.
15. Тема: Интеллектуальные обучающие программы по дисциплинам средней и высшей школы, специальным курсам.
16. Тема: Основные понятия теории предикатов, её использование для представления знаний.
17. Тема: Нечёткие множества, операции над ними. Использование нечётких выводов в экспертных системах.
18. Тема: Определение и методы построения когнитивных карт. Принятие решений с помощью когнитивных карт.
19. Тема: Применение автоматизированных систем поддержки принятия решений модельного типа в управлении предприятиями.
20. Тема: Применение систем искусственного интеллекта для статистического анализа данных и прогнозирования поведения объектов и систем.
21. Тема: OLAP – технологии.
22. Тема: Информационные хранилища: принципы построения, основные компоненты.
23. Тема: CASE – технологии: назначение, примеры.
24. Тема: Классификация систем искусственного интеллекта.

25. Тема: Контекстные системы поиска: назначение, примеры.

Реферат оценивается следующим образом:

- соответствие содержания теме- 4 балла;
- глубина проработки материала, 3 балла;
- грамотность и полнота использования источников, 1 балл;
- соответствие оформления реферата требованиям, 2 балла;
- доклад, 5 баллов;
- умение вести дискуссию и ответы на вопросы, 5 баллов.

Максимальное количество баллов: 20.

Критерии оценки:

- оценка «отлично» выставляется студенту, если набрал 19-20 баллов;
- оценка «хорошо» выставляется студенту, если набрал 15-18 баллов;
- оценка «удовлетворительно» выставляется студенту, если набрал 10-14 баллов;
- оценка «неудовлетворительно» выставляется студенту, если набрал менее 10 баллов;

Вопросы к зачету

Байесовская классификация

- Записать общую формулу байесовского классификатора (надо помнить формулу).
- Какие вы знаете три подхода к восстановлению плотности распределения по выборке?
- Что такое наивный байесовский классификатор?
- Что такое оценка плотности Парзена-Розенблатта (надо помнить формулу). Выписать формулу алгоритма классификации в методе парзеновского окна.
- На что влияет ширина окна, а на что вид ядра в методе парзеновского окна?
- Многомерное нормальное распределение (надо помнить формулу). Вывести формулу квадратичного дискриминанта. При каком условии он становится линейным?
- На каких предположениях осован линейный дискриминант Фишера?
- Что такое «проблема мультиколлинеарности», в каких задачах и при использовании каких алгоритмов она возникает? Какие есть подходы к её решению?

- Что такое «смесь распределений» (надо помнить формулу)?
- Что такое EM-алгоритм, какова его основная идея? Какая задача решается на E-шаге, на M-шаге? Каков вероятностный смысл скрытых переменных?
- Последовательное добавление компонент в EM-алгоритме, основная идея алгоритма.
- Что такое стохастический EM-алгоритм, какова основная идея? В чём его преимущество (какой недостаток стандартного EM-алгоритма он устраняет)?
- Что такое сеть радиальных базисных функций?
- Что такое «выбросы»? Как осуществляется фильтрация выбросов?

Метрическая классификация

- Что такое обобщённый алгоритм классификации (надо помнить формулу)? Какие вы знаете частные случаи?
- Как определяется понятие отступа в метрических алгоритмах классификации?
- Что такое окно переменной ширины, в каких случаях его стоит использовать?
- Что такое метод потенциальных функций? Идея алгоритма настройки. Сравните с методом радиальных базисных функций.
- Зачем нужен отбор опорных объектов в метрических алгоритмах классификации?
- Основная идея алгоритма СТОЛП.
- Что такое функция конкурентного сходства? Основная идея алгоритма FRiS-СТОЛП.
- Приведите пример метрического алгоритма классификации, который одновременно является байесовским классификатором.
- Приведите пример метрического алгоритма классификации, который одновременно является линейным классификатором.

Линейная классификация

- Что такое модель МакКаллока-Питтса (надо помнить формулу)?
- Метод стохастического градиента. Расписать градиентный шаг для квадратичной функции потерь и сигмоидной функции активации.
- Недостатки метода SG и как с ними бороться?
- Что такое линейный адаптивный элемент ADALINE?
- Что такое правило Хэбба?
- Что такое «сокращение весов»?
- Обоснование логистической регрессии (основная теорема), основные посылки (3) и следствия (2). Как выражается апостериорная вероятность классов (надо помнить формулу).

- Как выражается функция потерь в логистической регрессии (надо помнить формулу).
- Две мотивации и постановка задачи метода опорных векторов. Уметь вывести постановку задачи SVM (рекомендуется помнить формулу постановки задачи).
- Какая функция потерь используется в SVM? В логистической регрессии? Какие ещё функции потерь Вы знаете?
- Что такое ядро в SVM? Зачем вводятся ядра? Любая ли функция может быть ядром?
- Какое ядро порождает полимиальные разделяющие поверхности?
- Что такое ROC-кривая, как она определяется? Как она эффективно вычисляется?
- В каких алгоритмах классификации можно узнать не только классовую принадлежность классифицируемого объекта, но и вероятность того, что данный объект принадлежит каждому из классов?
- Каков вероятностный смысл регуляризации? Какие типы регуляризаторов Вы знаете?
- Что такое принцип максимума совместного правдоподобия данных и модели (надо помнить формулу)?

Регрессия

- Что такое ядерное сглаживание?
- Что есть общего между ядром в непараметрической регрессии и ядром SVM?
- На что влияет ширина окна, а на что вид ядра в непараметрической регрессии?
- Что такое окна переменной ширины, и зачем они нужны?
- Что такое «выбросы»? Как осуществляется фильтрация выбросов в непараметрической регрессии?
- Постановка задачи многомерной линейной регрессии. Матричная запись.
- Что такое сингулярное разложение? Как оно используется для решения задачи наименьших квадратов?
- Что такое «проблема мультиколлинеарности» в задачах многомерной линейной регрессии? Какие есть три подхода к её устранению?
- Сравнить гребневую регрессию и лассо. В каких задачах предпочтительнее использовать лассо?
- Какую проблему решает метод главных компонент в многомерной линейной регрессии? Записать матричную постановку задачи для метода главных компонент.

- Как свести задачу многомерной нелинейной регрессии к последовательности линейных задач?
- Метод настройки с возвращениями (backfitting): постановка задачи и основная идея метода.
- Какие методы построения логистической регрессии Вы знаете?
- Приведите примеры неквадратичных функций потерь в регрессионных задачах. С какой целью они вводятся?

Выбор модели и отбор признаков

- В чём отличия внутренних и внешних критериев?
- Разновидности внешних критериев.
- Разновидности критерия скользящего контроля.
- Что такое критерий непротиворечивости? В чём его недостатки?
- Что такое многоступенчатый выбор модели по совокупности критериев?
- Основная идея отбора признаков методом полного перебора. Действительно ли это полный перебор?
- Основная идея отбора признаков методом добавлений и исключнений.
- Что такое шаговая регрессия? Можно ли её использовать для классификации, в каком методе?
- Основная идея отбора признаков методом поиска в глубину.
- Основная идея отбора признаков методом поиска в ширину.
- Что такое МГУА?
- Основная идея отбора признаков с помощью генетического алгоритма.
- Основная идея отбора признаков с помощью случайного поиска.
- В чём отличия случайного поиска от случайного поиска с адаптацией?

Нейронные сети

- Приведите пример выборки, которую невозможно классифицировать без ошибок с помощью линейного алгоритма классификации. Какова минимальная длина выборки, обладающая данным свойством? Какие существуют способы модифицировать линейный алгоритм так, чтобы данная выборка стала линейно разделимой?
- Почему любая булева функция представима в виде нейронной сети? Сколько в ней слоёв?
- Метод обратного распространения ошибок. Основная идея. Основные недостатки и способы их устранения.
- Как можно выбирать начальное приближение в градиентных методах настройки нейронных сетей?
- Как можно ускорить сходимость в градиентных методах настройки нейронных сетей?
- Что такое диагональный метод Левенберга-Марквардта?

- Что такое «паралич» сети, и как его избежать?
- Как выбирать число слоёв в градиентных методах настройки нейронных сетей?
- Как выбирать число нейронов скрытого слоя в градиентных методах настройки нейронных сетей?
- В чём заключается метод оптимального прореживания нейронной сети? Какие недостатки стандартного алгоритма обратного распространения ошибок позволяет устранить метод ODB?

Критерии оценки:

- оценка «отлично» выставляется студенту, если изложение полученных знаний в устной форме полное, в системе, в соответствии с требованиями учебной программы; допускаются единичные несущественные ошибки, самостоятельно исправляемые учащимися;
- оценка «хорошо» выставляется студенту, если изложение полученных знаний в устной форме полное, в системе, в соответствии с требованиями учебной программы; допускаются, отдельные несущественные ошибки, исправляемые учащимися после указания преподавателя на них;
- оценка «удовлетворительно» выставляется студенту, если изложение полученных знаний неполное, однако это не препятствует усвоению последующего программного материала; допускаются отдельные существенные ошибки, исправляемые с помощью преподавателя;
- оценка «неудовлетворительно» выставляется студенту, если изложение учебного материала неполное, бессистемное, что препятствует усвоению последующей учебной информации; существенные ошибки, не исправляемые даже с помощью преподавателя;