Федеральное государственное бюджетное образовательное учреждение
высшего образования
Московский государственный университет имени М.В. Ломоносова
Факультет вычислительной математики и кибернетики

УТВЕРЖДАЮ
декан факультета вычислительной
математики и кибернетики

________________/И.А. Соколов /
«27» сентября 2023г.

ФОНД ОЦЕНОЧНЫХ СРЕДСТВ

по дисциплине

Байесовские методы машинного обучения

Уровень высшего образования:

бакалавриат

Направление подготовки / специальность:

02.03.02 "Фундаментальная информатика и информационные технологии" (3++)

Направленность (профиль) ОПОП:

Искусственный интеллект и анализ данных

Форма обучения:

очная

Рассмотрен и утвержден
на заседании Ученого совета факультета ВМК
(протокол №7, от 27 сентября 2023 года)

Москва 2023

1. ФОРМЫ И ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ТЕКУЩЕГО КОНТРОЛЯ УСПЕВАЕМОСТИ И ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ

В процессе и по завершении изучения дисциплины оценивается формирование у студентов следующих компетенций:

Планируемые результаты обучения по дисциплине (модулю)		
Содержание и код компетенции.	**Индикатор (показатель) достижения компетенции**	**Планируемые результаты обучения по дисциплине, сопряженные с индикаторами достижения компетенций**
ПК-1. Способен классифицировать и идентифицировать задачи искусственного интеллекта, выбирать адекватные методы и инструментальные средства решения задач искусственного интеллекта	ПК-1.1. Классифицирует и идентифицирует задачи систем искусственного интеллекта в зависимости от особенностей проблемной и предметной областей ПК-1.2. Выбирает методы и инструментальные средства искусственного интеллекта для решения задач в зависимости от особенностей проблемной и предметной областей ПК-1.3. Собирает исходную информацию и формирует требования к решению задач с использованием методов искусственного интеллекта	знать основные понятия, концепции и проблемы байесовского анализа и его приложений в различных моделях, основные методы семплирования из апостериорного распределения; уметь выбирать подходящие под конкретную задачу априорные распределения для латентных переменных, применять на практике методы семплирования латентных переменных, делать статистические выводы на основе выборки из апостериорного распределения и интерпретировать полученные результаты; владеть методами сбора и подготовки данных с помощью пакетов программ на языке R, навыками проведения байесовского анализа в среде R с помощью библиотек JAGS и STAN, методами проверки адекватности выбороки из апостериорного распределения.

1.1. Текущий контроль успеваемости

Текущий контроль успеваемости осуществляется путем оценки результатов выполнения заданий практических (семинарских) занятий, самостоятельной работы, предусмотренных учебным планом и посещения занятий/активность на занятиях.

В качестве оценочных средств текущего контроля успеваемости предусмотрены:

выполнение заданий на практических (семинарски) занятиях

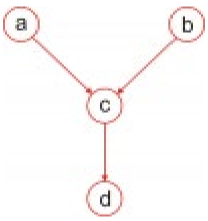
Примеры задач

Рассмотрим модель посещаемости студентами одного курса лекции. Пусть аудитория данного курса состоит из студентов профильной кафедры, а также студентов других кафедр. Обозначим через a количество студентов, распределившихся на профильную кафедру, а через b— количество студентов других кафедр на курсе. Пусть студенты профильной кафедры посещают курс с некоторой вероятностью p_1, а студенты остальных кафедр — с вероятностью p_2. Обозначим через c количество студентов на данной лекции. Тогда случайная величина $c|a,b$ есть сумма двух случайных величин, распределенных по биномиальному закону $B(a,p_1)$ и $B(b,p_2)$ соответственно. Пусть далее на лекции по курсу ведется запись студентов. При этом каждый студент записывается сам, а также, быть может, записывает своего товарища, которого на лекции на самом деле нет. Пусть студент записывает своего товарища с некоторой вероятностью p_3. Обозначим через d общее количество записавшихся на данной лекции. Тогда случайная величина $d|c$ представляет собой сумму c и случайной величины, распределенной по биномиальному закону $B(c,p_3)$. Для завершения задания вероятностной модели осталось определить априорные вероятности для a и для b. Пусть обе эти величины распределены равномерно в своих интервалах $[a_{min},a_{max}]$ и $[b_{min},b_{max}]$. Таким образом, мы определили следующую вероятностную модель:

Модель 1

$$p(a,b,c,d)=p(d|c)p(c|a,b)p(a)p(b),$$

$$d|c\sim c+B(c,p_3),$$
$$c|a,b\sim B(a,p_1)+B(b,p_2),$$
$$a\sim R[a_{min},a_{max}],$$
$$b\sim R[b_{min},b_{max}].$$



Графическая модель для вероятностной модели 1

Рассмотрим несколько упрощенную версию модели 1. Известно, что биномиальное распределение $B(n,p)$ при большом количестве испытаний и маленькой вероятности успеха может быть с высокой точностью приближено пуассоновским распределением $Poiss(\lambda)$ с $\lambda=np$. Известно также, что сумма двух пуассоновских распределений с параметрами λ_1 и λ_2 есть пуассоновское распределение с параметром $\lambda_1+\lambda_2$. Таким образом, мы можем сформулировать вероятностную модель, которая является приближенной версией модели 1:

Модель 2

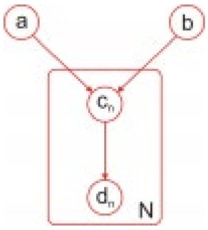
$$p(a,b,c,d)=p(d|c)p(c|a,b)p(a)p(b),$$
$$d|c\sim c+B(c,p_3),$$
$$c|a,b\sim Poiss(ap_1+bp_2),$$
$$a\sim R[a_{min},a_{max}],$$
$$b\sim R[b_{min},b_{max}].$$

Рассмотрим теперь модель посещаемости нескольких лекций курса. Будем считать, что

посещаемости отдельных лекций являются независимыми. Тогда:
Модель 3

$$p(a,b,c_1,\cdots,c_N,d_1,\cdots,d_N)=\prod_{n=1}^{N}p(d_n|c_n)p(c_n|a,b)p(a)p(b)$$
,

$d_n|c_n \sim c_n+B(c_n,p_3)$,
$c_n|a,b \sim B(a,p_1)+B(b,p_2)$,
$a \sim R[a_{min},a_{max}]$,
$b \sim R[b_{min},b_{max}]$.



Графическая модель для вероятностной модели 3

По аналогии с моделью 2 можно сформулировать упрощенную модель для модели 3:
Модель 4

$$p(a,b,c_1,\cdots,c_N,d_1,\cdots,d_N)=\prod_{n=1}^{N}p(d_n|c_n)p(c_n|a,b)p(a)p(b)$$
,

$d_n|c_n \sim c_n+B(c_n,p_3)$,
$c_n|a,b \sim Poiss(ap_1+bp_2)$,
$a \sim R[a_{min},a_{max}]$,
$b \sim R[b_{min},b_{max}]$.

Вариант 1

Рассматривается модель 2 с параметрами $a_{min}=15,a_{max}=30,b_{min}=250,b_{max}=350,p_1=0.5,p_2=0.05,p_3=0.5$. Провести на компьютере следующие исследования:

1. Найти математические ожидания и дисперсии априорных распределений для всех параметров a,b,c,d.
2. Пронаблюдать, как происходит уточнение прогноза для величины c по мере прихода новой косвенной информации. Для этого построить графики и найти мат.ожидание и дисперсию для распределений $p(c),p(c|b),p(c|a,b),p(c|a,b,d)$ при параметрах a,b,d, равных мат.ожиданиям своих априорных распределений, округленных до ближайшего целого.
3. Определить, какая из величин a,b,d вносит больший вклад в уточнение прогноза для величины c (в смысле дисперсии распределения). Для этого убедиться в том, что $\mathbb{D}[c|d]<\mathbb{D}[c|b]$ и $\mathbb{D}[c|d]<\mathbb{D}[c|a]$ для любых допустимых значений a,b,d. Найти множество точек (a,b) таких, что $\mathbb{D}[c|b]<\mathbb{D}[c|a]$. Являются ли множества $\{(a,b)|\mathbb{D}[c|b]<\mathbb{D}[c|a]\}$ и $\{(a,b)|\mathbb{D}[c|b]\geq\mathbb{D}[c|a]\}$ линейно разделимыми?
4. Провести временные замеры по оценке всех необходимых распределений $p(c),p(c|a),p(c|b),p(c|d),p(c|a,b),p(c|a,b,d),p(d)$.
5. Провести исследования из пп. 1-4 для точной модели 1 и сравнить результаты с аналогичными для модели 2. Привести пример оценки параметра, в котором разница между моделью 1 и 2 проявляется в большой степени.

Взять в качестве диапазона допустимых значений для величины c интервал $[0, a_{max}+b_{max}]$, а для величины d — интервал $[0, 2*(a_{max}+b_{max})]$.

При оценке выполнения задания будет учитываться эффективность программного кода. В частности, временные затраты на расчет отдельного распределения не должны превышать одной секунды.

Вариант 2

Рассматривается модель 2 с параметрами $a_{min}=15, a_{max}=30, b_{min}=250, b_{max}=350, p_1=0.5, p_2=0.05, p_3=0.5$. Провести на компьютере следующие исследования:

1. Найти математические ожидания и дисперсии априорных распределений для всех параметров a, b, c, d.
2. Пронаблюдать, как происходит уточнение прогноза для величины b по мере прихода новой косвенной информации. Для этого построить графики и найти мат.ожидание и дисперсию для распределений $p(b), p(b|a), p(b|a,d)$ при параметрах a, d, равных мат.ожиданиям своих априорных распределений, округленных до ближайшего целого.
3. Определить, при каких соотношениях параметров p_1, p_2 изменяется относительная важность параметров a, b для оценки величины c. Для этого найти множество точек $\{(p_1, p_2) | \mathbb{D}[c|b] < \mathbb{D}[c|a]\}$ при a, b, равных мат.ожиданиям своих априорных распределений, округленных до ближайшего целого. Являются ли множества $\{(p_1, p_2) | \mathbb{D}[c|b] < \mathbb{D}[c|a]\}$ и $\{(p_1, p_2) | \mathbb{D}[c|b] \geq \mathbb{D}[c|a]\}$ линейно разделимыми?
4. Провести временные замеры по оценке всех необходимых распределений $p(c), p(c|a), p(c|b), p(b|a), p(b|a,d), p(d)$.
5. Провести исследования из пп. 1-4 для точной модели 1 и сравнить результаты с аналогичными для модели 2. Привести пример оценки параметра, в котором разница между моделью 1 и 2 проявляется в большой степени.

Взять в качестве диапазона допустимых значений для величины c интервал $[0, a_{max}+b_{max}]$, а для величины d — интервал $[0, 2*(a_{max}+b_{max})]$.

При оценке выполнения задания будет учитываться эффективность программного кода. В частности, временные затраты на расчет отдельного распределения не должны превышать одной секунды.

Вариант 3

Рассматривается модель 4 с параметрами $a_{min}=15, a_{max}=30, b_{min}=250, b_{max}=350, p_1=0.5, p_2=0.05, p_3=0.5, N=50$. Провести на компьютере следующие исследования:

1. Найти математические ожидания и дисперсии априорных распределений для всех параметров a, b, c_n, d_n.
2. Реализовать генератор выборки $d_1, \cdots, d_N$ из модели при заданных значениях параметров a, b.
3. Пронаблюдать, как происходит уточнение прогноза для величины b по мере прихода новой косвенной информации. Для этого построить графики и найти мат.ожидание и дисперсию

для распределений $p(b), p(b|d_1), \cdots, p(b|d_1, \cdots, d_N)$, где выборка $d_1, \cdots, d_N$ 1) сгенерирована из модели при параметрах a, b, равных мат.ожиданиям своих априорных распределений, округленных до ближайшего целого и 2) $d_1 = \cdots = d_N$, где d_n равно мат.ожиданию своего априорного распределения, округленного до ближайшего целого. Провести аналогичный эксперимент, если дополнительно известно значение a. Сравнить результаты двух экспериментов.

4. Провести временные замеры по оценке всех необходимых распределений $p(c_n), p(d_n), p(b|d_1, \cdots, d_n), p(b|a, d_1, \cdots, d_n)$.
5. Провести исследования из пп. 1-4 для точной модели 3 и сравнить результаты с аналогичными для модели 4.

Взять в качестве диапазона допустимых значений для величины c интервал $[0, a_{max} + b_{max}]$, а для величины d — интервал $[0, 2*(a_{max} + b_{max})]$.

При оценке выполнения задания будет учитываться эффективность программного кода. В частности, временные затраты на расчет отдельного распределения не должны превышать одной секунды.

1.2. Промежуточная аттестация

Промежуточная аттестация осуществляется в форме экзамена
В качестве средств, используемых на промежуточной аттестации предусматривается:
Билеты

1.3. Типовые задания для проведения промежуточной аттестации

Вопросы к экзамену

1. Байесовский подход к теории вероятностей. Оценка параметров в байесовском и частотном подходе. Примеры байесовских рассуждений.
2. Сопряжённые распределения. Примеры. Экспоненциальный класс распределений, его свойства.
3. Решение задачи выбора модели по Байесу. Обоснованность модели. Полный байесовский вывод.
4. Вероятностная модель линейной регрессии. Метод релевантных векторов для задачи регрессии.
5. Логистическая регрессия. Метод релевантных векторов для задачи классификации.
6. EM-алгоритм в общем виде. Примеры применения.
7. Вариационный подход для приближенного байесовского вывода.
8. Вариационная линейная регрессия.
9. Задача уменьшения размерности в данных. Вероятностная модель главных компонент, ее обучение с помощью метода максимального правдоподобия и EM-алгоритма.
10. Распределение Дирихле. Свойства накопления и нейтральности. Генерация выборки из Дирихле через гамма-распределения и через stick-breaking.
11. Байесовская модель разделения смеси гауссиан. Вариационный вывод для неё.
12. Тематическая модель LDA. Обучение и вывод в модели.
13. Методы MCMC для оценки статистик вероятностных распределений. Теоретические свойства марковских цепей.
14. Схема Метрополиса-Хастингса и схема Гиббса. Примеры использования.
15. Гауссовские процессы для задачи регрессии. Подбор параметров ковариационной функции.
16. Гауссовские процессы для задачи классификации.
17. Процессы Дирихле. Представление процесса Дирихле с помощью процесса китайского ресторана. Схема Гиббса для разделения смеси распределений с процессом Дирихле.
18. Процессы Дирихле. Представление процесса Дирихле с помощью stick-breaking. Вариационный вывод для разделения смеси распределений с процессом Дирихле.

Пример экзаменационного билета

1. Вероятностная модель линейной регрессии. Метод релевантных векторов для задачи регрессии.
2. Схема Метрополиса-Хастингса и схема Гиббса. Примеры использования.

2. КРИТЕРИИ ОЦЕНКИ ПО ДИСЦИПЛИНЕ

ШКАЛА И КРИТЕРИИ ОЦЕНИВАНИЯ результатов обучения (РО) по дисциплине				
Оценка виды оценочных средств	**2 (не зачтено)**	**3 (зачтено)**	**4 (зачтено)**	**5 (зачтено)**
Знания (виды оценочных средств: приведены в п. 1.2.)	Отсутствие знаний	Фрагментарные знания	Общие, но не структурированные знания	Сформированные систематические знания
Умения (виды оценочных средств: приведены в п. 1.2.)	Отсутствие умений	В целом успешное, но не систематическое умение	В целом успешное, но содержащее отдельные пробелы умение (допускает неточности непринципиального характера)	Успешное и систематическое умение
Навыки (владения, опыт деятельности) (виды оценочных средств: приведены в п. 1.2..)	Отсутствие навыков (владений, опыта)	Наличие отдельных навыков (наличие фрагментарного опыта)	В целом, сформированные навыки (владения), но используемые не в активной форме	Сформированные навыки (владения), применяемые при решении задач