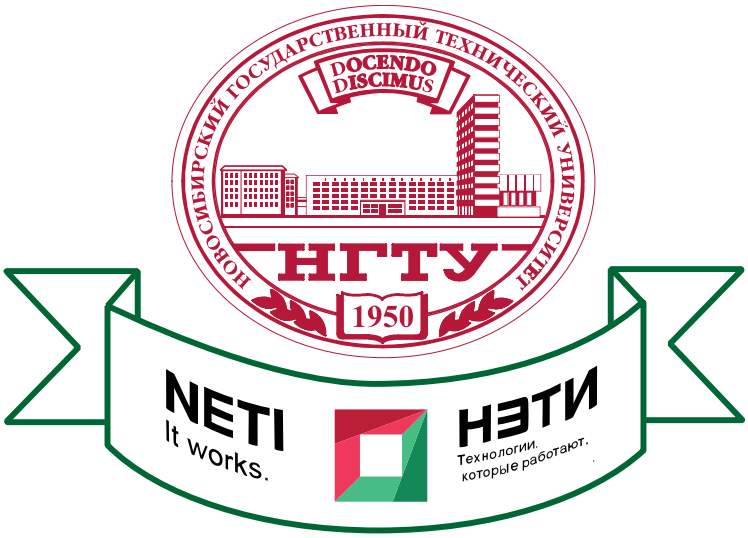
Министерство науки и высшего образования
Российской Федерации

Федеральное государственное бюджетное
образовательное учреждение высшего образования

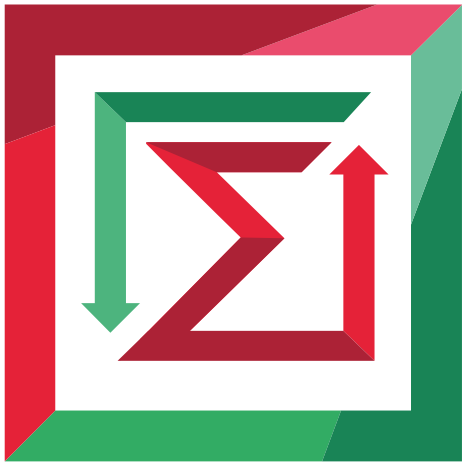
«Новосибирский государственный технический университет»



Кафедра теоретической и прикладной информатики

Лабораторная работа № 2

по дисциплине «Статистический анализ нечисловых данных»



Факультет:	ПМИ
Группа:	ПМИ-02
Вариант:	22
Студент:	Сидоров Даниил,
	Дюков Богдан
Преподаватель:	Тимофеева Анастасия Юрьевна.

Новосибирск

2026

Задание 1

С помощью метода на основе корреляций (CFS) выбрать из набора потенциальных объясняющих переменных признаки, которые будут использоваться для обучения наивного байесовского классификатора.

Решение

Для этого мы нашли набор S_k из k признаков, который обеспечивает максимум следующей функции:

$$F(S_k) = \frac{\sum_{i \in S_k} R_i}{\sqrt{k + 2 \sum_{i,j \in S_k, i \neq j} r_{ij}}}$$

где R_i – абсолютное значение показателя взаимосвязи между i-м признаком и откликом, r_{ij}– абсолютное значение показателя взаимосвязи между i-м и j-м признаком, S_k - подмножество из k признаков. Показатель взаимосвязи для нашего варианта – коэффициент корреляции Пирсона.

Чтобы найти искомое подмножество, мы перебрали все комбинации признаков, для каждой комбинации находили значение $F(S_k)$, в результате нашли лучшую комбинацию признаков: **('A1', 'A3', 'A6')**, которая обеспечивает максимум функции, равный 0.45869. Данные признаки будут являться нашими объясняющими переменными, которые будут использоваться для обучения наивного байесовского классификатора.

Задание 2

Разделить исходную выборку из 500 объектов (строк) на:

- обучающую – первые 400 объектов,
- контрольную – последние 100 объектов.

По данным из обучающей выборки исходя из полученного в п. 1 набора признаков обучить наивный байесовский классификатор в предположении многомерной категориальной вероятностной модели. При оценивании вероятностей использовать поправку Лапласа.

Решение

По 400 объектам обучающей выборки строим таблицы сопряженности между откликом и каждой объясняющей переменной в отдельности:

	A21	
A1	**0**	**1**
A11	58	52
A12	41	67
A13	8	20
A14	22	132

	A21	
A3	**0**	**1**
A30	9	2
A31	13	9
A32	69	150
A33	16	22
A34	22	88

	A21	
A6	**0**	**1**
A61	97	155
A62	10	38
A63	6	13
A64	3	10
A65	13	55

При оценивании вероятностей воспользуемся поправкой Лапласа. При использовании поправки Лапласа, мы добавляем 1 к числителю (частоте каждого класса для данного значения признака), и добавляем количество возможных классов к знаменателю (общему количеству наблюдений каждого класса). Это делается для того, чтобы избежать проблемы с нулевыми вероятностями при отсутствии определенного класса для данного значения признака в обучающих данных. Например, для A11-0 имеем формулу:

$$\frac{58+1}{58+41+8+22+3} = \frac{59}{133} = 0{,}44361$$

	A21	
A1	**0**	**1**
A11	0,44361	0,19273
A12	0,31579	0,24727
A13	0,06767	0,07636
A14	0,17293	0,48364

	A21	
A3	**0**	**1**
A30	0,07463	0,01087
A31	0,10448	0,03623
A32	0,52239	0,5471
A33	0,12687	0,08333
A34	0,17164	0,32246

	A21	
A6	**0**	**1**
A61	0,73134	0,56522
A62	0,08209	0,1413
A63	0,05224	0,05072
A64	0,02985	0,03986
A65	0,10448	0,2029

Доля объектов A21 = 1:

$$\frac{271+1}{400+2} = 0{,}67662$$

Доля объектов A21 = 0:

$$\frac{129+1}{400+2} = 0{,}32338$$

Задание 3

Для объектов из контрольной выборки построить прогноз отклика с помощью правил максимального правдоподобия и апостериорного максимума.

Решение

Подставляем соответствующие оценки условных вероятностей и перемножаем их. Для правила апостериорного максимума дополнительно умножаем на долю объектов A21 = 0 и A21 = 1 в обучающей выборке. Выбираем класс, соответствующий максимальному значению. Первые 20 строк контрольной выборки и их прогнозы двумя правилами:

Правило максимального правдоподобия					
A1	A3	A6	No	Yes	Prediction
A14	A32	A65	0,17293*0,52239*0,10448=0,00944	0,48364*0,5471*0,2029=0,05369	1
A14	A34	A64	0,17293*0,17164*0,02985=0,00089	0,48364*0,32246*0,03986=0,00622	1
A13	A34	A65	0,06767*0,17164*0,10448=0,00121	0,07636*0,32246*0,2029=0,005	1
A12	A34	A62	0,31579*0,17164*0,08209=0,00445	0,24727*0,32246*0,1413=0,01127	1
A12	A34	A61	0,31579*0,17164*0,73134=0,03964	0,24727*0,32246*0,56522=0,04507	1
A14	A34	A63	0,17293*0,17164*0,05224=0,00155	0,48364*0,32246*0,05072=0,00791	1
A11	A32	A61	0,44361*0,52239*0,73134=0,16948	0,19273*0,5471*0,56522=0,0596	0
A14	A34	A65	0,17293*0,17164*0,10448=0,0031	0,48364*0,32246*0,2029=0,03164	1
A12	A32	A62	0,31579*0,52239*0,08209=0,01354	0,24727*0,5471*0,1413=0,01912	1
A14	A34	A61	0,17293*0,17164*0,73134=0,02171	0,48364*0,32246*0,56522=0,08815	1
A12	A32	A61	0,31579*0,52239*0,73134=0,12065	0,24727*0,5471*0,56522=0,07646	0
A11	A32	A61	0,44361*0,52239*0,73134=0,16948	0,19273*0,5471*0,56522=0,0596	0
A12	A31	A62	0,31579*0,10448*0,08209=0,00271	0,24727*0,03623*0,1413=0,00127	0
A11	A34	A61	0,44361*0,17164*0,73134=0,05569	0,19273*0,32246*0,56522=0,03513	0
A14	A34	A61	0,17293*0,17164*0,73134=0,02171	0,48364*0,32246*0,56522=0,08815	1
A11	A32	A61	0,44361*0,52239*0,73134=0,16948	0,19273*0,5471*0,56522=0,0596	0
A11	A32	A61	0,44361*0,52239*0,73134=0,16948	0,19273*0,5471*0,56522=0,0596	0
A14	A34	A65	0,17293*0,17164*0,10448=0,0031	0,48364*0,32246*0,2029=0,03164	1
A14	A32	A65	0,17293*0,52239*0,10448=0,00944	0,48364*0,5471*0,2029=0,05369	1
A14	A34	A65	0,17293*0,17164*0,10448=0,0031	0,48364*0,32246*0,2029=0,03164	1
A12	A32	A61	0,31579*0,52239*0,73134=0,12065	0,24727*0,5471*0,56522=0,07646	0

Правило апостериорного максимума					
A1	A3	A6	No	Yes	Prediction
A14	A32	A65	0,00944*0,32338=0,00305	0,05369*0,67662=0,03633	1
A14	A34	A64	0,00089*0,32338=0,00029	0,00622*0,67662=0,00421	1
A13	A34	A65	0,00121*0,32338=0,00039	0,005*0,67662=0,00338	1
A12	A34	A62	0,00445*0,32338=0,00144	0,01127*0,67662=0,00762	1
A12	A34	A61	0,03964*0,32338=0,01282	0,04507*0,67662=0,03049	1
A14	A34	A63	0,00155*0,32338=0,0005	0,00791*0,67662=0,00535	1
A11	A32	A61	0,16948*0,32338=0,05481	0,0596*0,67662=0,04032	0
A14	A34	A65	0,0031*0,32338=0,001	0,03164*0,67662=0,02141	1
A12	A32	A62	0,01354*0,32338=0,00438	0,01912*0,67662=0,01293	1
A14	A34	A61	0,02171*0,32338=0,00702	0,08815*0,67662=0,05964	1
A12	A32	A61	0,12065*0,32338=0,03901	0,07646*0,67662=0,05174	1
A11	A32	A61	0,16948*0,32338=0,05481	0,0596*0,67662=0,04032	0
A12	A31	A62	0,00271*0,32338=0,00088	0,00127*0,67662=0,00086	0
A11	A34	A61	0,05569*0,32338=0,01801	0,03513*0,67662=0,02377	1
A14	A34	A61	0,02171*0,32338=0,00702	0,08815*0,67662=0,05964	1
A11	A32	A61	0,16948*0,32338=0,05481	0,0596*0,67662=0,04032	0
A11	A32	A61	0,16948*0,32338=0,05481	0,0596*0,67662=0,04032	0
A14	A34	A65	0,0031*0,32338=0,001	0,03164*0,67662=0,02141	1
A14	A32	A65	0,00944*0,32338=0,00305	0,05369*0,67662=0,03633	1
A14	A34	A65	0,0031*0,32338=0,001	0,03164*0,67662=0,02141	1
A12	A32	A61	0,12065*0,32338=0,03901	0,07646*0,67662=0,05174	1

Задание 4

Оценить качество классификации с помощью следующих показателей:

- частота истинно положительных результатов (чувствительность),
- частота истинно отрицательных результатов (специфичность),
- частота ошибок,
- точность.

Сравнить результаты, полученные с помощью правил максимального правдоподобия и апостериорного максимума.

Решение

Сопоставим отклик контрольной выборки и полученные оценки с помощью правил максимального правдоподобия и апостериорного максимума:

Факт	МП	АМ
1	1	1
1	1	1
1	1	1

1	1	1
1	1	1
0	1	1
0	0	0
1	1	1
1	1	1
1	1	1
0	0	1
1	0	0
0	0	0
1	0	1
1	1	1
0	0	0
0	0	0
1	1	1
1	1	1
1	1	1

Построим матрицы неточностей для прогнозов отклика, построенных с помощью правил максимального правдоподобия и апостериорного максимума:

МП		
Fact	**1**	**0**
1	48	19
0	9	24

АМ		
Fact	**1**	**0**
1	59	8
0	16	17

Вычисляем показатели:

- Чувствительность. Это доля истинно положительных результатов среди всех действительно положительных случаев. Она вычисляется как TP / (TP + FN), где TP - количество истинно положительных результатов, а FN - количество ложно отрицательных результатов.
- Специфичность. Это доля истинно отрицательных результатов среди всех действительно отрицательных случаев. Она вычисляется как TN / (TN + FP), где TN - количество истинно отрицательных результатов, а FP - количество ложно положительных результатов.
- Частота ошибок. Это доля неправильных предсказаний среди всех случаев. Она вычисляется как (FP + FN) / (TP + TN + FP + FN).
- Точность (Precision): Это доля истинно положительных результатов среди всех положительных предсказаний. Она вычисляется как TP / (TP + FP).

Вычисляем:

$$\text{Чувствительность МП } = \frac{48}{48+19} = 0{,}71642$$

$$\text{Чувствительность АМ } = \frac{59}{59+8} = 0{,}8806$$

$$\text{Специфичность МП } = \frac{24}{24+9} = 0{,}72727$$

$$\text{Специфичность АМ } = \frac{17}{17+16} = 0{,}51515$$

$$\text{Частота ошибок МП } = \frac{19+9}{48+19+9+24} = 0{,}28$$

$$\text{Частота ошибок АМ } = \frac{16+8}{59+8+16+17} = 0{,}24$$

$$\text{Точность МП } = \frac{48}{48+9} = 0{,}8421$$

$$\text{Точность АМ } = \frac{59}{59+16} = 0{,}78667$$

Сравнивая результаты, полученные с помощью правил максимального правдоподобия (МП) и апостериорного максимума (АМ), можно сделать следующие выводы:

- Чувствительность. Метод апостериорного максимума (0.8806) имеет более высокую чувствительность по сравнению с методом максимального правдоподобия (0.71642). Это означает, что метод АМ лучше определяет положительные случаи.

- Специфичность. Метод максимального правдоподобия (0.72727) имеет более высокую специфичность по сравнению с методом апостериорного максимума (0.51515). Это означает, что метод МП лучше определяет отрицательные случаи.

- Частота ошибок. Метод апостериорного максимума (0.24) имеет более низкую частоту ошибок по сравнению с методом максимального

правдоподобия (0.28). Это означает, что метод АМ делает меньше ошибок в общем.

- Точность. Метод максимального правдоподобия (0.8421) имеет более высокую точность по сравнению с методом апостериорного максимума (0.78667). Это означает, что из всех положительных прогнозов, которые делает метод МП, большая доля действительно является положительными.

Выводы

В результате выполнения лабораторной работы мы успешно использовали метод на основе корреляций (CFS) для выбора признаков для обучения наивного байесовского классификатора. Также обучили наивный байесовский классификатор, используя многомерную категориальную вероятностную модель и поправку Лапласа. После мы использовали два разных правила для прогнозирования отклика: правило максимального правдоподобия и правило апостериорного максимума. Это позволило нам сравнить эти два подхода и понять, какой подход лучше в зависимости от ситуации. По итогу нам удалось оценить качество классификации с помощью различных показателей, таких как чувствительность, специфичность, частота ошибок и точность. Это позволило нам количественно оценить производительность нашего классификатора и определить его сильные и слабые стороны.