Неверова Елена Григорьевна
Университет Нархоз, Республика Казахстан

Применение байесовой вероятности при анализе больших данных с использованием различных пакетов языка R

Язык R давно вышел за рамки заявленной в его названии статистической обработки данных и стал инструментом всестороннего и глубокого анализа, предлагая к использованию пакеты для создания нейросетей, исследования эмоциональной окраски текстов, построения социальных графов.

Метод Байеса или **«**наивный Байес**»,** как один из самых распространенных вероятностных подходов, является очень простым и эффективным инструментом для работы, связанной с анализом больших данных[1].

Следует заметить, что большое разнообразие применения данного метода в среде языка R благодаря множеству пакетов, содержащих данную модель.

Рассмотрим вариант применения байесовой сети в задаче определения успешности прохождения дисциплины **«**Основы R**»** студентами нашего университета, базируясь на реальных данных:

- результатах первого внутрисеместрового контроля,
- результатах второго внутрисеместрового контроля,

и, при среднем арифметическом обоих ВСК больше 50% для получения допуска

- результатах сдачи экзамена (больше 50% - сдано).

Итоговая оценка - результат среднего арифметического ВСК1 и ВСК2 (60%) и оценка за экзамен (40%).

Взаимосвязь этих показателей изображена на рисунке 1.

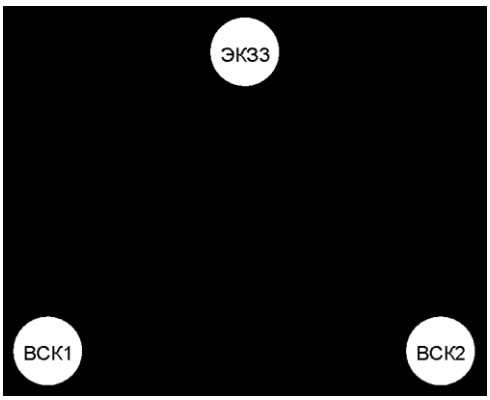

Рисунок 1. Граф взаимосвязи контрольных точек.

Первый вариант - применение пакета **«**e1071**»** [2]. Используем категориальные переменные следующего вида:

	ВСК1	ВСК2	ЭКЗ3	РЕЗУЛЬТАТ
1	хорошо	хорошо	удовлетворительно	Сдал
2	отлично	отлично	отлично	Сдал
3	удовлетворительно	удовлетворительно	хорошо	Сдал
4	хорошо	хорошо	хорошо	Сдал
5	неудовлетворительно	неудовлетворительно	неудовлетворительно	Не сдал
6	удовлетворительно	удовлетворительно	удовлетворительно	Сдал
7	хорошо	хорошо	хорошо	Сдал

Реализуем байесову модель, применив библиотеку «e1071»:

```
> library(e1071)
>rating<-read.csv("C:/Users/Home/Documents/OsnovyR.csv",header= T)
>model<-naiveBayes(РЕЗУЛЬТАТ~.,data=rating)
```

Функция naiveBayes() генерирует таблицу апиорных и апостеорных вероятностей (см рис. 2):

```
Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
  Не сдал      Сдал
0.2307692 0.7692308

Conditional probabilities:
         ВСК1
Y         неудовлетворительно   отлично удовлетворительно    хорошо
  Не сдал           0.3333333 0.0000000         0.3333333 0.3333333
  Сдал              0.1000000 0.1000000         0.4000000 0.4000000

         ВСК2
Y          удовлетворительно неудовлетворительно   отлично
  Не сдал          0.0000000           0.3333333 0.0000000
  Сдал             0.1000000           0.1000000 0.1000000
         ВСК2
Y         удовлетворительно    хорошо
  Не сдал         0.6666667 0.0000000
  Сдал            0.3000000 0.4000000

         ЭКЗ3
Y         неудовлетворительно отлично удовлетворительно хорошо
  Не сдал                 1.0     0.0                0.0    0.0
  Сдал                    0.0     0.1                0.6    0.3
```

Рисунок 2. Результат применения функции naiveBayes из пакета «e1071».

Прогнозные значения вероятности успешной сдачи другой дисциплины для первых 5-ти студентов выглядят следующим образом:

```
> predict(model,rating[sample(1:13,10,replace=FALSE),],type="raw")
        Не сдал        Сдал
[1,] 3.690037e-03 0.996309963
[2,] 9.250694e-04 0.999074931
[3,] 1.386963e-03 0.998613037
[4,] 1.041666e-06 0.999998958
[5,] 9.982032e-01 0.001796766
```

Следующий пакет bnlearn [3] предпочитает представление исходных данных в следующем виде:

	subject	type	id	rate
1	Основы R	1	1	85
2	Основы R	1	2	100
3	Основы R	1	3	70
4	Основы R	1	4	75
5	Основы R	1	5	0
...				
16	Основы R	2	8	65
17	Основы R	3	1	80
18	Основы R	3	2	97
19	Основы R	3	3	70
...				

В результате простейших преобразований, получим байесову модель для анализа и прогноза успеваемости в различных разрезах запросов.

```
>library(bnlearn)
>rating<-read.csv("C:/Users/Home/Documents/rating.csv")
>net <- model2network("[ВСК1][ВСК2|ВСК1][ЭК33|ВСК1:ВСК2]")
>ВСК1 <- rating$rate[rating$subj == "Основы R" & rating$type == 1]
>ВСК2 <- rating$rate[rating$subj == "Основы R" & rating$type == 2]
>ЭК33 <- rating$rate[rating$subj == "Основы R" & rating$type == 3]
>rate <- data.frame(ВСК1 = as.numeric(ВСК1), ВСК2 = as.numeric(ВСК2), ЭК33
= as.numeric(ЭК33))
>net.rate <- bn.fit(net, rate)
```

Такая сеть, обученная на байесовой модели, может рассуждать следующим образом: например, какова вероятность получить итоговую оценку «удовлетворительно» или «хорошо», если результаты экзамена пока неизвестны?

```
> cpquery(net.rate, (ЭК33 > 50 & ЭК33 < 90), TRUE)
[1] 0.4257778
```

Либо, анализируя причинность, можно задать вопрос: «Если получена экзаменационная оценка «неудовлетворительно», с какой долей вероятности можно предположить, что какая-либо из его оценок по ВСК меньше 50%?».

```
> cpquery(net.rate,(ВСК1 < 50 | ВСК2 < 50), (ЭК33<50))
[1] 0.8403318
```

Еще один пакет, позволяющий построить и визуализировать байесову модель - «CausalImpact»от Google [4].

Пакет CausalImpact R реализует байесовский подход к оценке влияния вмешательств во временном ряду. Учитывая временные ряды ответов (например, текущие оценки) и набор контрольных временных рядов (например, результаты), пакет создает байесовскую структурную модель временного ряда с встроенными эффектами spike-and-slab prior для автоматического выбора переменных. Затем эта модель используется для прогнозирования контрфактов, то есть выявления тенденций изменения ответа байесовой сети после вмешательства, если вмешательство не произошло.

Требуемый вид исходных данных, выглядит следующим образом:

	ВСК1	ВСК2	ЭКЗз	РЕЗУЛЬТАТ
1	85	75	70	Сдан
2	100	95	90	Сдан
3	70	65	75	Сдан
4	75	75	80	Сдан
5	0	0	0	Не Сдан
6	65	55	50	Сдан

Текст программы:

```
>library(CausalImpact)
>rating<-read.csv("C:/Users/Home/Documents/OsnoviR.csv")
>sem<-(rating$ВСК1+rating$ВСК2)/2
>itog<-c(sem,rating$ЭК33)
>pre.period <- c(1,8)
>post.period <- c(9,20)
>impact <- CausalImpact(itog, pre.period, post.period)
>plot(impact)
```

Вывод графика (см рис.3):

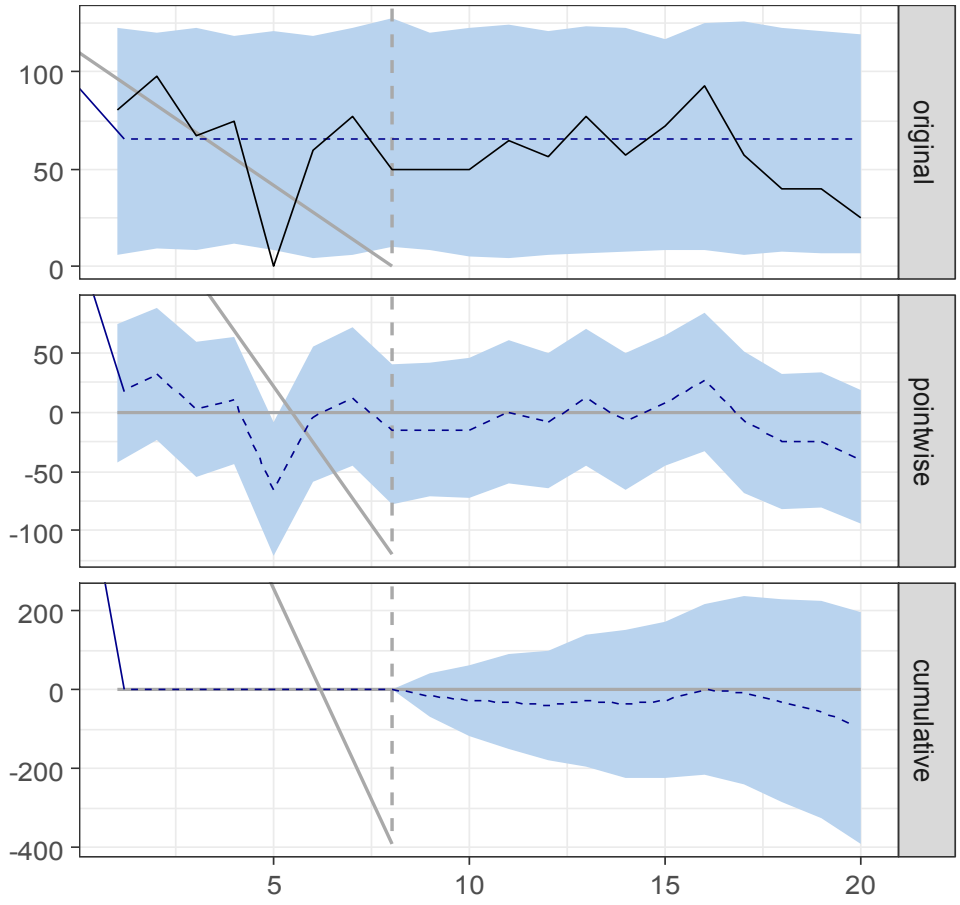

Рисунок 3. Результат применения пакета «CausalImpact»

График состоит из трех частей. В верхней части показаны данные и контрфактическое предсказание для последующих подобных случаев. Вторая панель показывает разницу между наблюдаемыми данными и контрафактическими предсказаниями. Это точечный причинный эффект, как оценивается процесс построенный байесовской моделью. Третья панель складывает поточные вклады со второй панели, в результате чего получается

график кумулятивного эффекта вмешательства. На нем видна основная причина изменений.

Мы рассмотрели несколько библиотек, позволяющих строить и обследовать вероятностные модели практически любой предметной области.

Каждый из них обладает своими выразительными средствами и набором инструментов. Предпочтение в выборе пакета отдается исходя из качества полученных наборов данных и поставленных перед исследователем задач.

Литература:

1. Эфрон Б. Математика. Теорема Байеса в XXI веке. Science 2013; стр. 340 : 1177-8. 10.1126 / science.1236536
2. Различные функции Департамента статистики, группы теории вероятностей (ранее: E1071), TU Wien [R пакет e1071 версии 1.6-7]. Полная архивная сеть (CRAN); 2014. Доступно в Интернете URL: https://CRAN.R-project.org/package=e1071
3. Олег Шмелев. Байесовы сети в R: пакет bnlearn. Блог "Электронные семечки". Дата обращения 25.03.2018. URL: http://electronicseeds.blogspot.ru/2014/05/
4. Портал Google Open Source. CausalImpact: A new open-source package for estimating causal effects in time series. September 10, 2014. URL: https://opensource.googleblog.com/2014/09/causalimpact-new-open-source-package.html?m=1