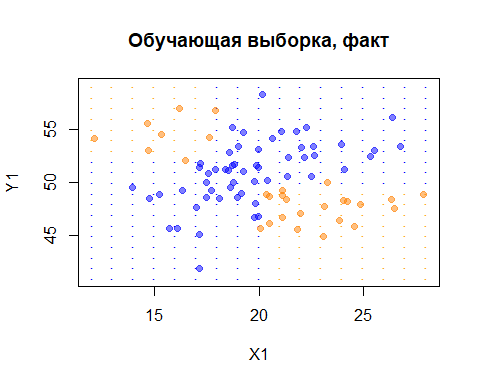
Упражнение 2

Кориба Марина

01 03 2021

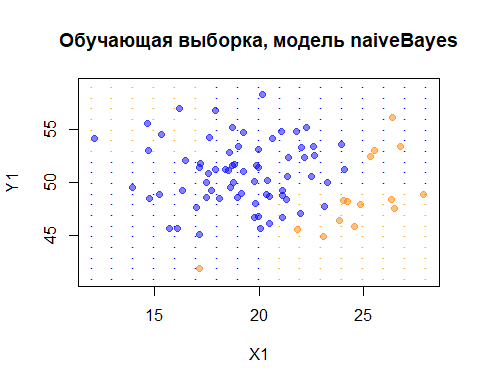
Сгенерированы данные, непрерывно объясняющие переменные.

Нарисуем обучающую выборку на графике. Сеткой точек показаны области классов, соответствующие истинным дискриминирующим правилам.



Обучим модель наивного байесовского классификатора и оценим её точность (верность) на обучающей выборке.

Байесовский классификатор



## y.nb.train  
## y.train 0 1  
## 0 51 5  
## 1 19 10

## [1] 0.7176471

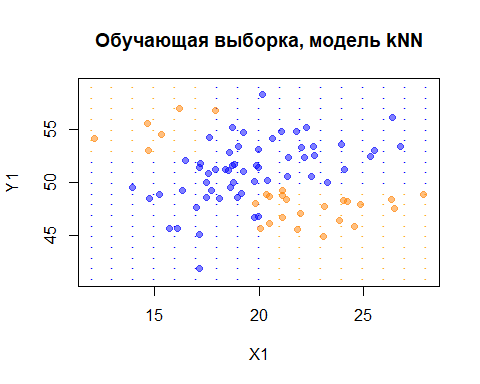
Как можно видеть на графике, ту часть жёлтого класса, которая расположена в левой верхней области пространства координат, модель классифицирует неверно. Таким образом, байесовская решающая граница не моделирует разрыв жёлтого класса синим. Это происходит потому, что в непрерывном случае наивный байесовский метод исходит из допущения о линейной разделимости двух классов и нормальности распределения объясняющих переменных в них. Однако в этом примере это допущение не выполняется.

Сделаем прогноз классов Y на тестовую выборку и оценим точность модели. Как можно убедиться, точность на тестовой оказывается ниже, чем на обучающей выборке. Учитывая, как ведёт себя классификатор на обучающей выборке, такой модели доверять не стоит.

## y.nb.test  
## y.test 0 1  
## 0 9 2  
## 1 3 1

## [1] 0.6666667

Построим модель kNN. Это “ленивый” классификатор, ему не требуется предварительное обучение. А ещё это непараметрический метод, и чем меньше количество ближайших соседей , тем гибче ведёт себя разделяющая граница. Метод хорошо работает с линейно неразделимыми классами.



## y.knn.train  
## y.train 0 1  
## 0 55 1  
## 1 3 26

## [1] 0.9529412

Можно видеть, что классификация обучающей выборки методом kNN не сильно отличается от фактических классов наблюдений. Оценим также точность модели на тестовой выборке.

## y.knn.test  
## y.test 0 1  
## 0 10 1  
## 1 2 2

## [1] 0.8

Модель kNN оказалась точной на этих данных, чего не скажешь о байесовском классификаторе

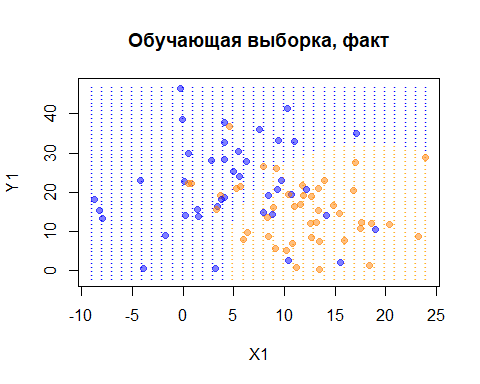
**Задача**

Построить модели на данных *примера 3* с параметрами распределений, соответствующими своему варианту. На графики нанести сетку истинных классов. Определить, какой из методов срабатывает на этих данных лучше, и почему.

* n = 100, доля обучающей выборки: 85%
* класс
* класс

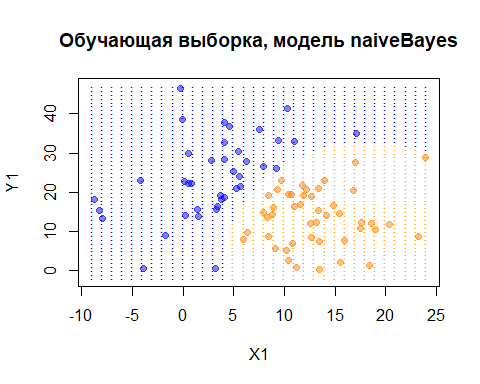
Сгенерированы данные, непрерывно объясняющие переменные.

Нарисуем обучающую выборку на графике. Сеткой точек показаны области классов, соответствующие истинным дискриминирующим правилам. Это правило создаём, зная истинные законы распределения классов, как максимум из двух плотностей распределения (плотность многомерного закона считаем функцией dmvnorm(), классы точкам сетки присваиваем пользовательской функцией rules.mv()).



Обучим модель наивного байесовского классификатора и оценим её точность (верность) на обучающей выборке. Поскольку объясняющие переменные для классов сгенерированы как двумерные нормальные распределения и сами классы не перекрываются, следует ожидать, что эта модель окажется точной.

### Байесовский классификатор



## y.nb.train  
## y.train 0 1  
## 0 30 11  
## 1 9 35

## [1] 0.7647059

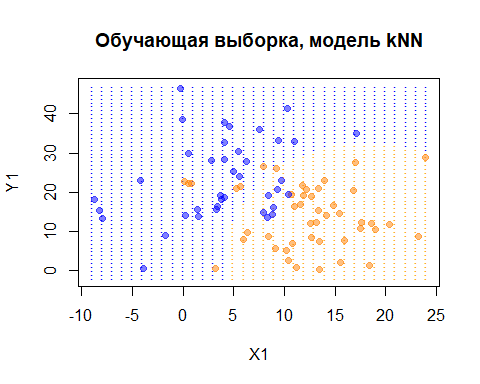
Точность на обучающей выборке недостаточно высокая. Сделаем прогноз классов Y на тестовую выборку и оценим точность модели.

## y.nb.test  
## y.test 0 1  
## 0 4 0  
## 1 2 9

## [1] 0.8666667

Байесовский метод разделяет классы на обучающей выборке, практически не ошибаясь, делая всего две ошибки.

Построим модель kNN. С этими данными у метода не должно возникнуть проблем, так как он не проводит чёткой границы между классами, а в каждом случае ориентируется на соседние наблюдения.



## y.knn.train  
## y.train 0 1  
## 0 33 8  
## 1 6 38

## [1] 0.8352941

Точность на обучающей выборке высока, но не идеальна. Оценка точности на тестовой выборке также показывает, что модель классифицирует верно все наблюдения, кроме шести, близких к границе разделения классов.

|  |  |
| --- | --- |
|  | Значение характеристики |
| TPR | 0.80 |
| SPC | 0.73 |
| PPV | 0.76 |
| NPV | 0.77 |
| FNR | 0.20 |
| FPR | 0.27 |
| FDR | 0.24 |
| MMC | 0.53 |

Вывод: характеристики качества должны быть приближены к единице, а ошибки - минимальными.