Вариант 26

Упражнение 2

Степанов Г.А.

26 Февраль 2021

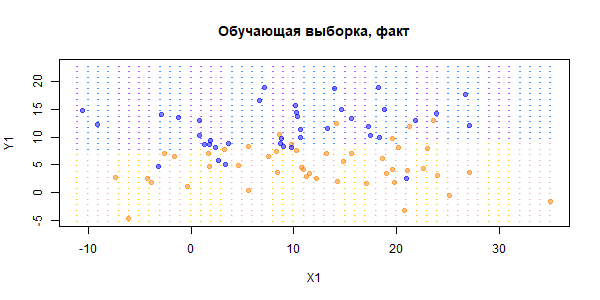
**Задача 1**

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.

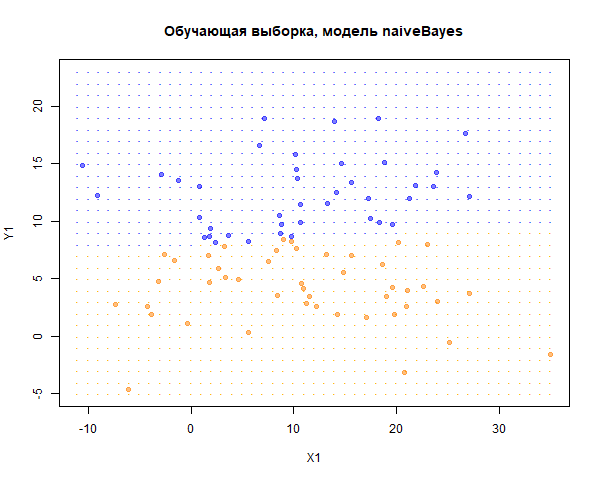
Сгенерированы данные, непрерывно объясняющие переменные.

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

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.



Матрица неточностей на обучающей выборке:

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 33 | 6 |
| 1 | 7 | 39 |

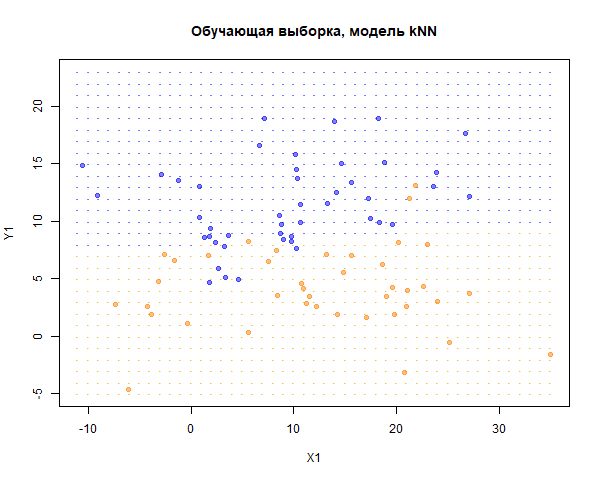
Точность на обучающей выборке высокая = 0.8470588. Сделаем прогноз классов Y на тестовую выборку и оценим точность модели.

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 5 | 1 |
| 1 | 2 | 7 |

Наивный байесовский метод разделяет классы на обучающей выборке, но с ошибками в тестовой в 3 наблюдениях, от этого её точность равна 0.8.

Точность на тестовой оказывается ниже, чем на обучающей выборке. Учитывая, как ведёт себя классификатор на обучающей выборке, такой модели доверять не стоит.

Построим модель kNN (метод ближайшего соседа). 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.



Матрица неточностей на обучающей выборке:

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 36 | 3 |
| 1 | 9 | 37 |

Можно видеть (0.8588235), что классификация обучающей выборки методом kNN не сильно отличается от фактических классов наблюдений. Оценим также точность модели на тестовой выборке, на ней получаются приблизтельн такте же значения.

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 4 | 2 |
| 1 | 0 | 9 |

**Вывод**: На обеих тестовых выборках точность оказалась одинаковой и не очень высокой, поэтому выбор метода является затруднительным. Выберем метод ближайшего соседа для дальнейших расчётов характеристик качества.

**Задача 2**

По матрице неточностей той модели, которая оказалась лучше по Acc, рассчитать характеристики качества и ошибки из лекции: TPR, SPC, PPV, NPV, FNR, FPR, FDR, MCC.

Матрица неточностей:

Чувствительность (sensitivity, TP rate, recall):

Специфичность (specificity, TN rate):

Ценность положительного прогноза (positive predictive value, precision):

Ценность отрицательного прогноза (negative predictive value):

*Ошибки (чем ниже, тем лучше):*

* Доля ложноотрицательных исходов (FN rate):
* Доля ложных срабатываний (fall-out rate, FP rate):
* Доля ложного обнаружения (false discovery rate):

Корреляция Мэтьюса:

Построим таблицу со всеми характеристиками качества для обучающей и тестовой выборок.

|  |  |  |
| --- | --- | --- |
|  | Обучающая | Тестовая |
| TPR | 0.804 | 1.000 |
| SPC | 0.923 | 0.667 |
| PPV | 0.925 | 0.818 |
| NPV | 0.800 | 1.000 |
| FNR | 0.196 | 0.000 |
| FPR | 0.077 | 0.333 |
| FDR | 0.075 | 0.182 |
| MMC | 0.726 | 0.739 |

Вывод: Таким образом, характеристики качества TPR, SPC, PPV, NPV и MMC желательно должны быть приближены к единице, в то время как ошибки - быть минимальными, как и получилось у нашей модели.