СТАТИСТИКА БОЛЬШИХ ДАННЫХ

Лекция 27. Задача классификации. Часть 2

Байесовская теория классификации

Байесовская теория классификации основана на применении оптимального байесовского классификатора и оценивании плотностей распределения классов по обучающей выборке. Различные методы оценивания плотности порождают большое разнообразие байесовских классификаторов.

- **Параметрическое оценивание плотности**
  - квадратичный дискриминант
  - линейный дискриминант Фишера
- **Непараметрическое оценивание плотности**
  - метод парзеновского окна
- **Оценивание плотности как смеси параметрических плотностей**
  - разделение смеси распределений, EM-алгоритм
  - метод радиальных базисных функций

Идея метода байесовской классификации

Поиск оптимального алгоритма $a(x)$ сводится к максимизации **апостериорной вероятности** класса y при условии x. Ставится задача определения максимально вероятного класса $y \in Y$, содержащего заданный объект классификации $x \in X$:

$$a(x) = \operatorname{argmax}_{y \in Y} p(y|x)$$

Воспользуемся известной формулой Байеса:

$$a(x) = \operatorname{argmax}_{y \in Y} p(y|x) = \operatorname{argmax}_{y \in Y} \frac{p(x|y)p(y)}{p(x)} = \operatorname{argmax}_{y \in Y} p(x|y)p(y).$$

Байесовские методы классификации

Пусть X - множество объектов, Y - конечное множество имён классов, множество $X \times Y$ является вероятностным пространством с плотностью распределения

$$p(x, y) = P(y)p(x|y).$$

Вероятности появления объектов каждого из классов $P_y = P(y)$ называются **априорными вероятностями классов**.

Плотности распределения $p_y(x) = p(x|y)$ называются **функциями правдоподобия классов.**

Задачи статистической классификации

Задача 1. Имеется обучающая выборка $X^l = (x_i, y_i)_{i=1}^l$ из неизвестного распределения $p(x, y) = P_y p_y(x)$. Требуется построить эмпирические оценки априорных вероятностей $\hat{P}_y$ и функций правдоподобия $\hat{p}_y(x)$ для каждого из классов $y \in Y$.

Задача 2. По известным плотностям распределения $p_y(x)$ и априорным вероятностям P_y всех классов $y \in Y$ построить алгоритм $a(x)$, минимизирующий вероятность ошибочной классификации.

Рассмотрим сначала решение задачи 2.

Решение задачи 2

Функционал среднего риска:

$$R(a) = \sum_{y \in Y} \lambda_y P_y P(a(x) \neq a^*(x)|y)$$

λ_y – плата за неверное распознавание класса y.

Апостериорная вероятность $P(y|x)$ класса y для объекта x вычисляется по формуле Байеса, если P_y и $p_y(x)$ известны:

$$P(y|x) = \frac{p(x,y)}{p(x)} = \frac{p_y(x)P_y}{\sum_{s \in Y} p_s(x)P_s}.$$

Оптимальное байесовское решающее правило

Байесовское решающее правило. Минимум среднего риска $R(a)$ реализуется на алгоритме:

$$a(x) = \arg\max_{y \in Y} \lambda_y P_y p_y(x)$$

Минимальное значение среднего риска $R(a)$ называется **байесовским риском.**

Если классы равнозначны $\lambda_y = 1$, то байесовское правило называется также **принципом максимума апостериорной вероятности**.

Если классы равновероятны, $P_y = \frac{1}{|Y|}$, то объект *x* просто относится к классу *y* с наибольшим значением плотности распределения $p_y(x)$ в точке *x*.

Задача восстановления плотности распределения

Требуется оценить плотность вероятностного распределения $p(x,y) = P_y p_y(x)$, сгенерировавшего обучающую выборку $(x_i, y_i)_{i=1}^{l}$

Априорные вероятности классов P_y:

$$\hat{P}_y = \frac{l_y}{l}$$

Задача восстановления плотности

Задано множество из m объектов $X^m = (x_1, \ldots, x_m) \subset X$, распределенных согласно неизвестному распределению $p(x)$. **Требуется** построить эмпирическую оценку плотности – функцию $\hat{p}(x)$, приближающую $p(x)$ на всём X.

Наивный байесовский классификатор

Обозначим через $\xi = (\xi_1, \dots, \xi_n)$ произвольный элемент пространства объектов X.

Гипотеза независимости. Признаки $\xi_1, \dots, \xi_n$ являются независимыми случайными величинами.

В этом случае функции правдоподобия классов представимы в виде произведения

$$p_y(\xi) = p_{y1}(\xi_1)p_{y2}(\xi_2)\cdots p_{yn}(\xi_n),\ y \in Y,$$

где $p_{yj}(\xi_j)$ – плотность распределения значений j-го признака для класса y.

Непараметрическая классификация

Одномерный непрерывный случай ($x \in R$):

$$\hat{p}(x) = \frac{1}{mh}\sum_{i=1}^{m} K\left(\frac{x - x_i}{h}\right).$$

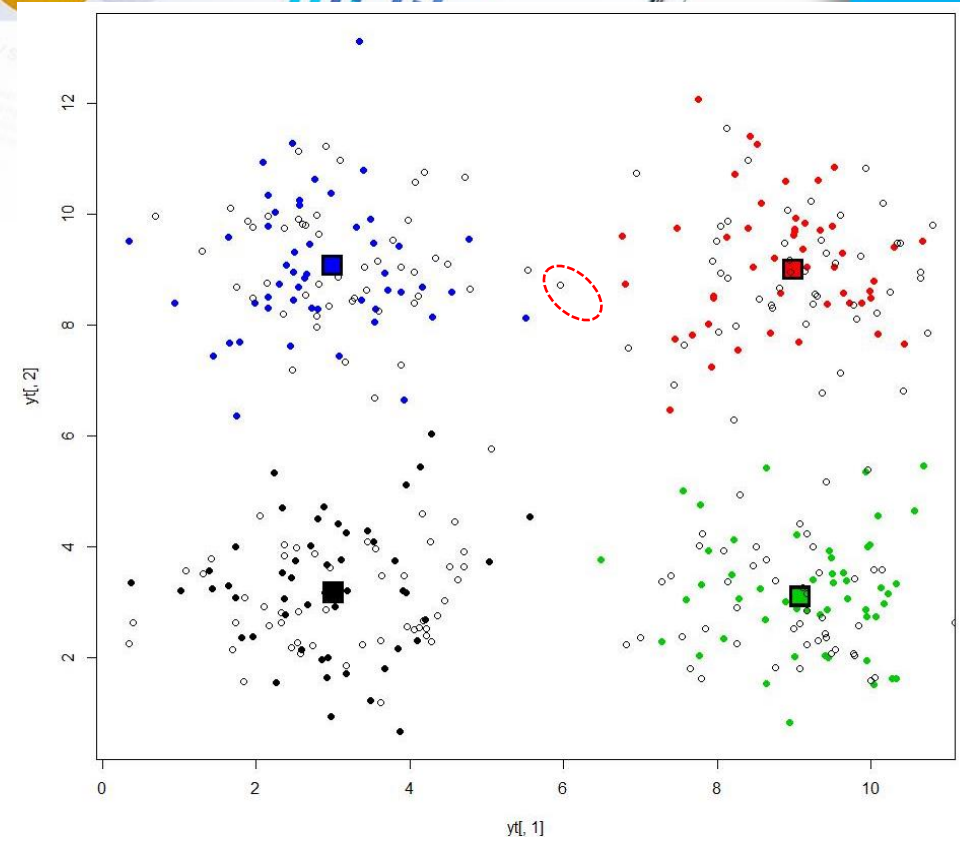
Многомерный непрерывный случай.

Пусть выполнена гипотеза независимости. Оценка многомерной плотности в точке $\xi = (\xi_1, \dots, \xi_n)$ имеет вид:

$$\hat{p}(\xi) = \prod_{j=1}^{n} \frac{1}{mh_j}\sum_{i=1}^{m} K\left(\frac{\xi_j - x_{ji}}{h_j}\right),$$

$x_i = (x_{1i}, \dots, x_{ni})$ – i-ый элемент обучающей выборки X^m

Пример

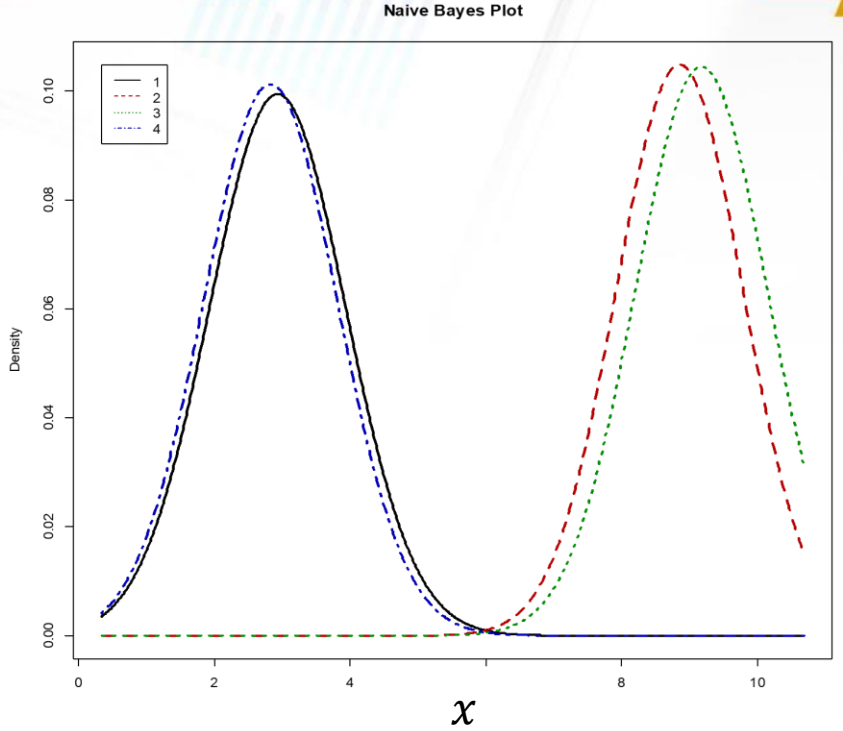


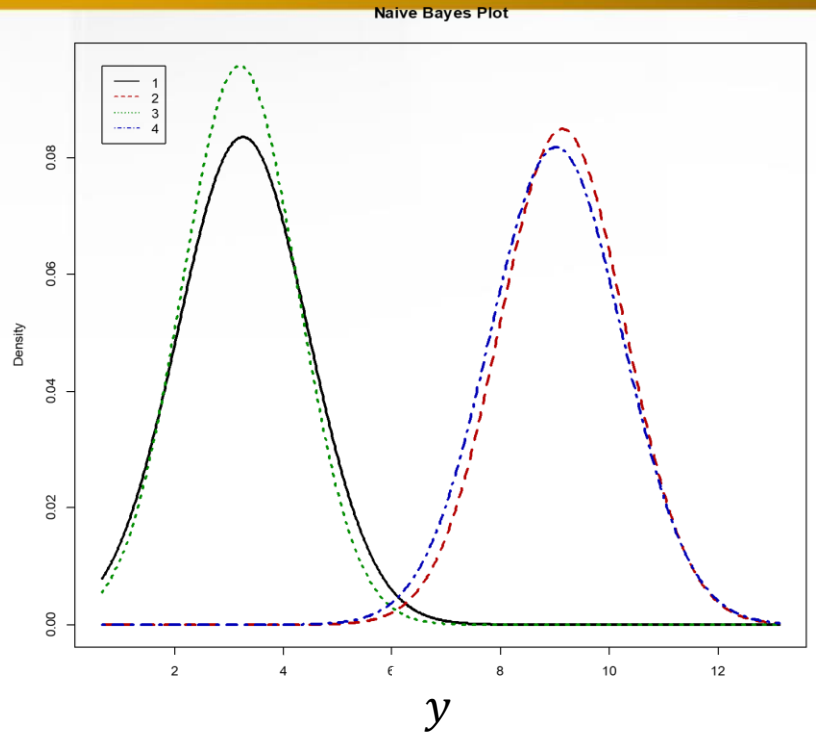

```
library(klaR)
…
nb<-NaiveBayes(zlab~., data=z)
pre<- predict (nb,ze)$class
…
plot(nb,col = 1:4,lwd=3)
```

Предсказанное

Фактическое	1	2	3	4
1	51	0	0	0
2	0	50	0	0
3	0	0	50	0
4	0	1	0	48

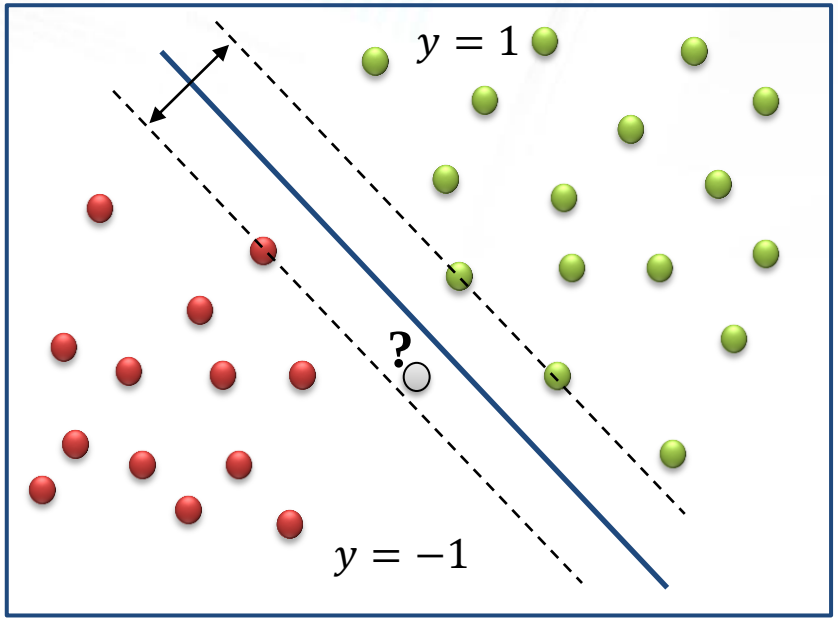
Пример


Naive Bayes Plot
1
2
3
4
Density
0.00
0.02
0.04
0.06
0.08
0.10
0
2
4
8
10
x


Naive Bayes Plot
1
2
3
4
Density
0.00
0.02
0.04
0.06
0.08
2
4
8
10
12
y

Метод опорных векторов (SVM)

$x \in R^n, y \in Y = \{-1,1\}.$



Строим разделяющую гиперплоскость с максимальным зазором:

$$(w^*, x) - b = 0,$$

где w^* – решение задачи оптимизации:

$$(w^*)^2 = \min w^2$$
$$y_i[(w, x_i) - b] \geq 1, \qquad i = 1, \dots, l,$$

$(x_i, y_i)_{i=1}^{l}$ – обучающая выборка

Уравнение классификатора:

$$y = a(x) = \text{sign}\big((w^*, x) - b\big)$$

Метод деревьев решений

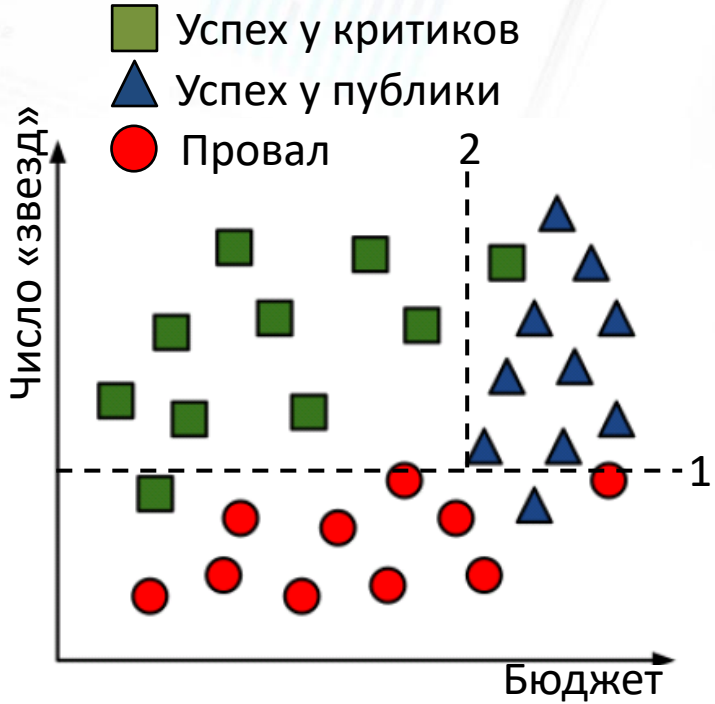
Метод деревьев решений (decision tree) для задачи классификации состоит в делении исходных данных на группы до тех пор, пока не будут получены однородные подмножества.

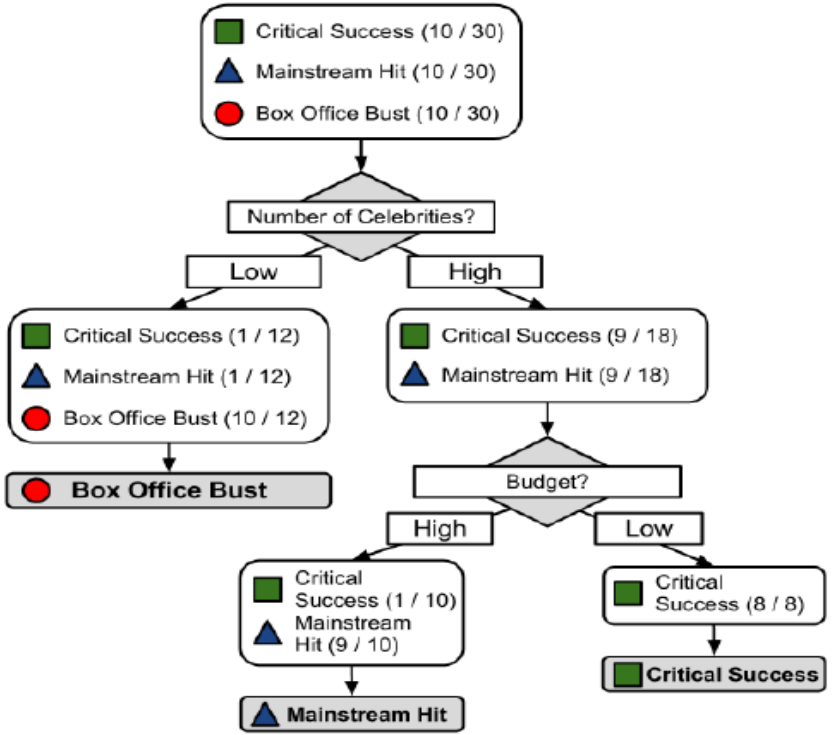
Идея метода

Метод деревьев решений реализует принцип «рекурсивного разделения». В узлах, начиная с корневого, выбирается признак, значение которого используется для разбиения всех данных на 2 класса. Процесс продолжается до тех пор, пока не выполнится один из критерий остановки:

- Все (или почти все) данные данного узла принадлежат одному классу
- Не осталось признаков, по которым можно построить новое разбиение
- Дерево превысило заранее заданный «предел роста»

Пример


Успех у критиков
Успех у публики
Провал
Число «звезд»
2
1
Бюджет


Critical Success (10 / 30)
Mainstream Hit (10 / 30)
Box Office Bust (10 / 30)
Number of Celebrities?
Low
High
Critical Success (1 / 12)
Mainstream Hit (1 / 12)
Box Office Bust (10 / 12)
Box Office Bust
Critical Success (9 / 18)
Mainstream Hit (9 / 18)
Budget?
High
Low
Critical Success (1 / 10) Mainstream Hit (9 / 10)
Mainstream Hit
Critical Success (8 / 8)
Critical Success

Контрольные вопросы и задания

1. Сформулируйте существенные признаки задач кластеризации и классификации. В чем их сходство и различие?
2. Классифицируйте методы классификации. Какие из методов классификации можно отнести к статистическим?
3. Сгенерируйте 3 выборки по 200 элементов из двумерных нормальных распределений с центрами в точках (3,3), (9,2), (9,6) и диагональными ковариационными матрицами с элементами (1.5,1.5), (1,1), (1,1). Решить задачу кластеризации методом k средних. Нарисуйте результирующие графики. Сделайте выводы.
4. Решить задачу классификации методом k ближайших соседей и наивным байесовским методом, используя решение предыдущей задачи по приведенным образцам в лекции. Построить таблицу результатов.

Заключение

Данный курс лекций включил в себя только основные идеи и подходы применения статистических методов в больших данных. Готовя этот курс, мы, коллектив авторов, многому научились. Мы надеемся, что и для вас он будет полезен и станет отправной точкой для продвижения в области обработки больших данных.

Спасибо за внимание!