Домашнее задание №2

Иноземцев Игорь, 177 группа

2 ноября, 2014г.

1 Постановка задачи

Задача 1.

Визуализация набора данных "ирисы Фишера".

Задача 2.

Оценка плотности распределений с использованием разных типов ядер (применяются прямоугольное и гауссово ядра). Применение оценки плотности распределения к задаче классификации ирисов на три вида: Iris Setosa, Iris Versicolour, Iris Virginica.

2 Описание проведенных экспериментов

1. Загрузка данных, разделение на входные данные (4 признака: длина и ширина чашелистника, длина и ширина лепестков) и выходные (один из трех классов).
2. Визуализация
   (a) Построение гистограмм для каждого признака.
   (b) Построение scatter plot matrix.
3. Оценка плотностей различных распределений с помощью метода Парзена-Розенблатта, используется прямоугольное ядро.
   (a) Двумерное нормальное стандартное распределение
   (b) Смесь двух нормальных распределений.
4. Аналогичные действия для гауссова ядра.
5. Классификация ирисов на три класса с использованием Байесовского решающего правила.

3 Результаты

1. Визуализация набора данных.

Рис. 1: Гистограммы для набора данных "ирисы Фишера"@

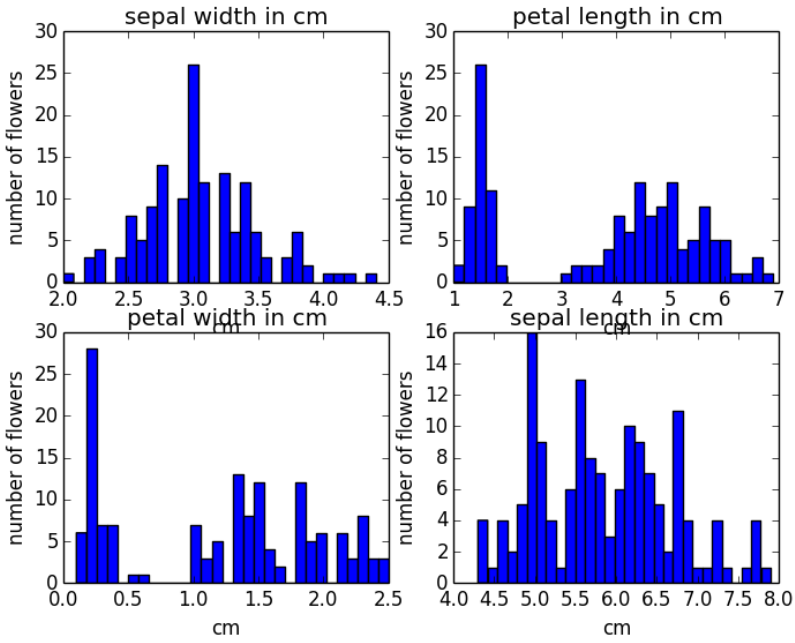
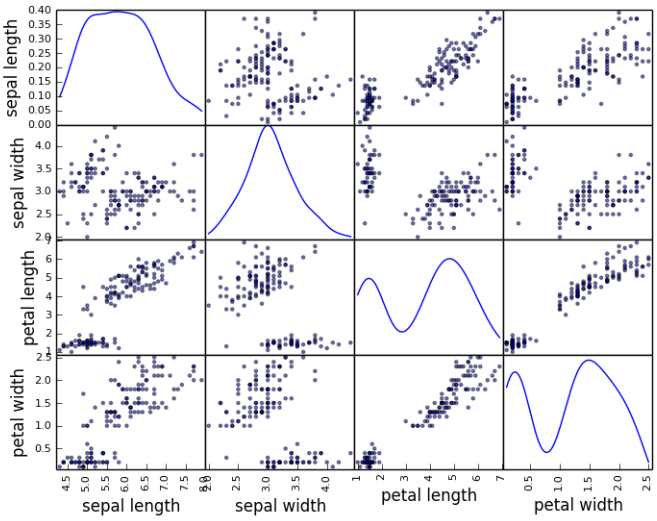


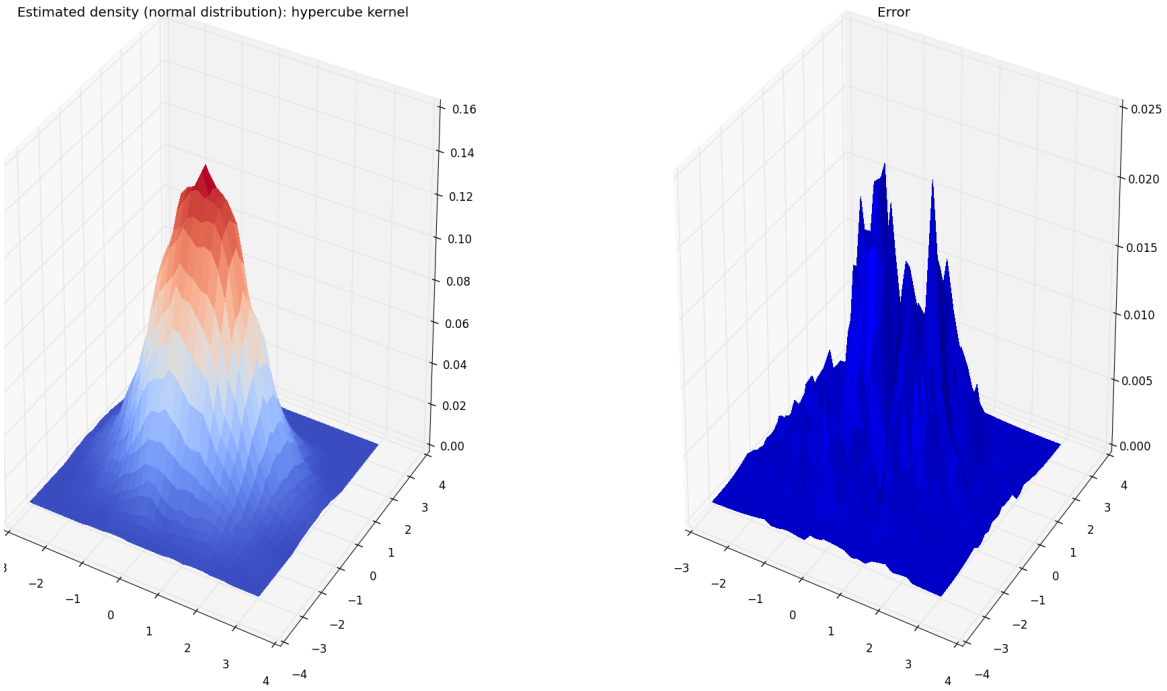
Рис. 2: Scatter plot matrix



2. Оценка плотностей распределений.

   (a) Оцениваем плотность стандартного нормального распределения, используя прямоугольное ядро

Рис. 3: Стандартное нормальное распределение, прямоугольное ядро. Справа изображен график модуля ошибки

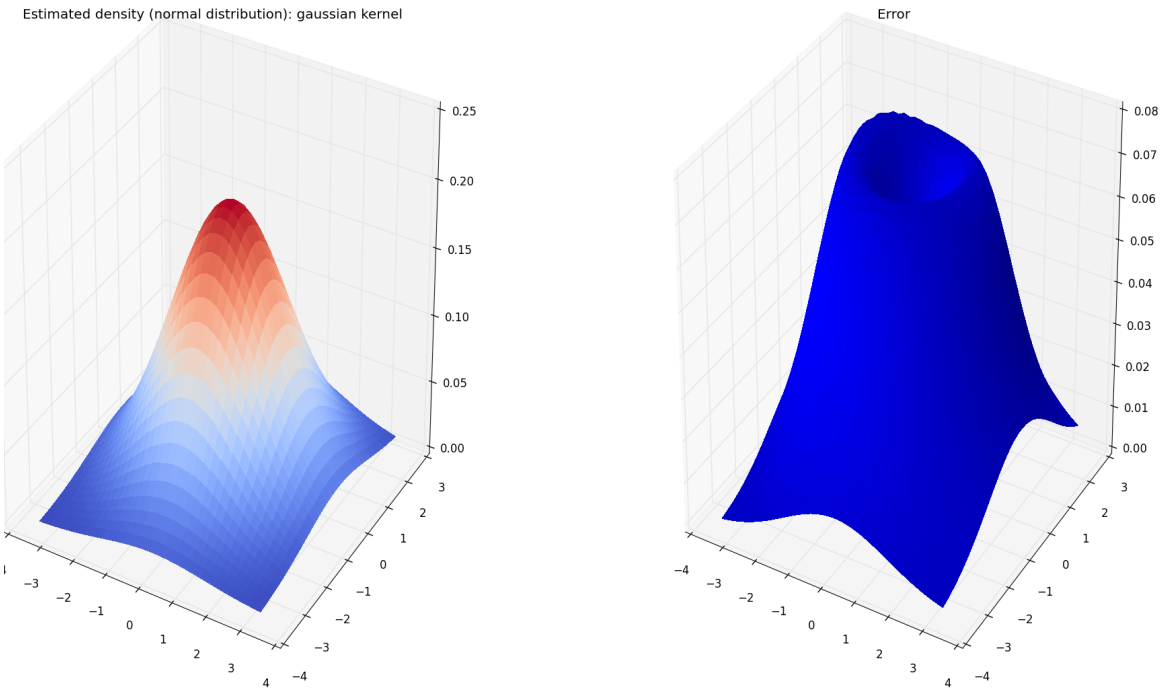


Ширина окна	1
Количество элементов в выборке	2000

Ошибка довольно большая (порядка 15

(b) Оцениваем плотность того же распределения, используя гауссово ядро

Рис. 4: Стандартное нормальное распределение, гауссово ядро. Справа изображен график модуля ошибки

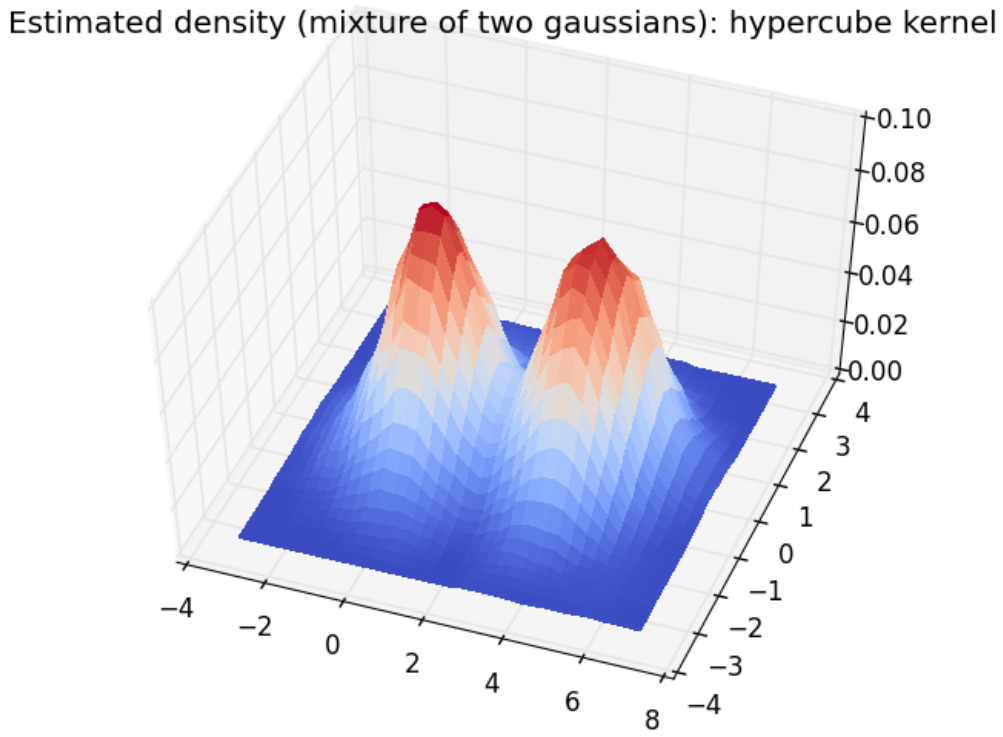


Ширина окна	1.4
Количество элементов в выборке	2000

Ошибка еще больше, чем в предыдущем пункте.

(c) Проделаем то же самое для смеси двух нормальных распределений, применяя прямоугольное ядро

Рис. 5: Смесь двух нормальных распределений, прямоугольное ядро

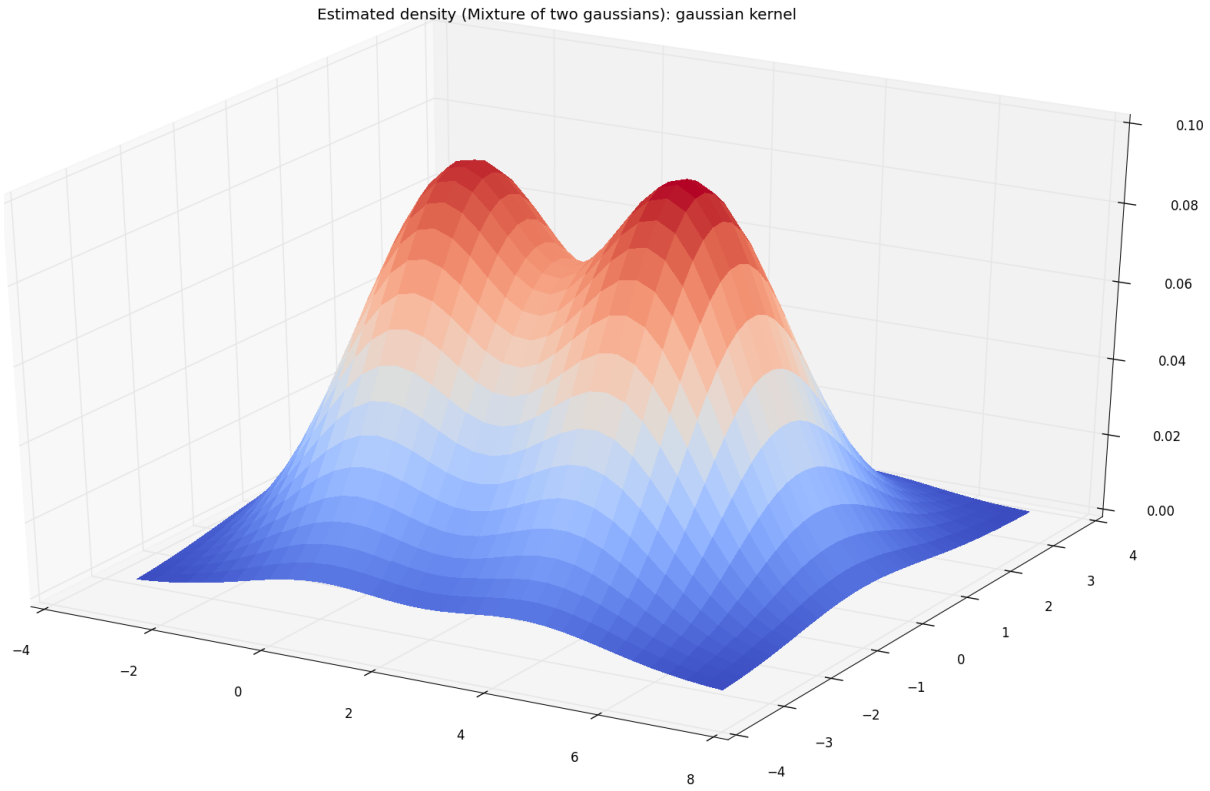


$$\mu_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mu_2 = \begin{pmatrix} 4 \\ 0 \end{pmatrix}, \Sigma_1 = \Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Ширина окна	1.4
Количество элементов в выборке	2000 (1000 для одной гауссианы и 1000 для другой)

(d) Оцениваем плотность для смеси двух нормальных распределений, используя гауссово ядро

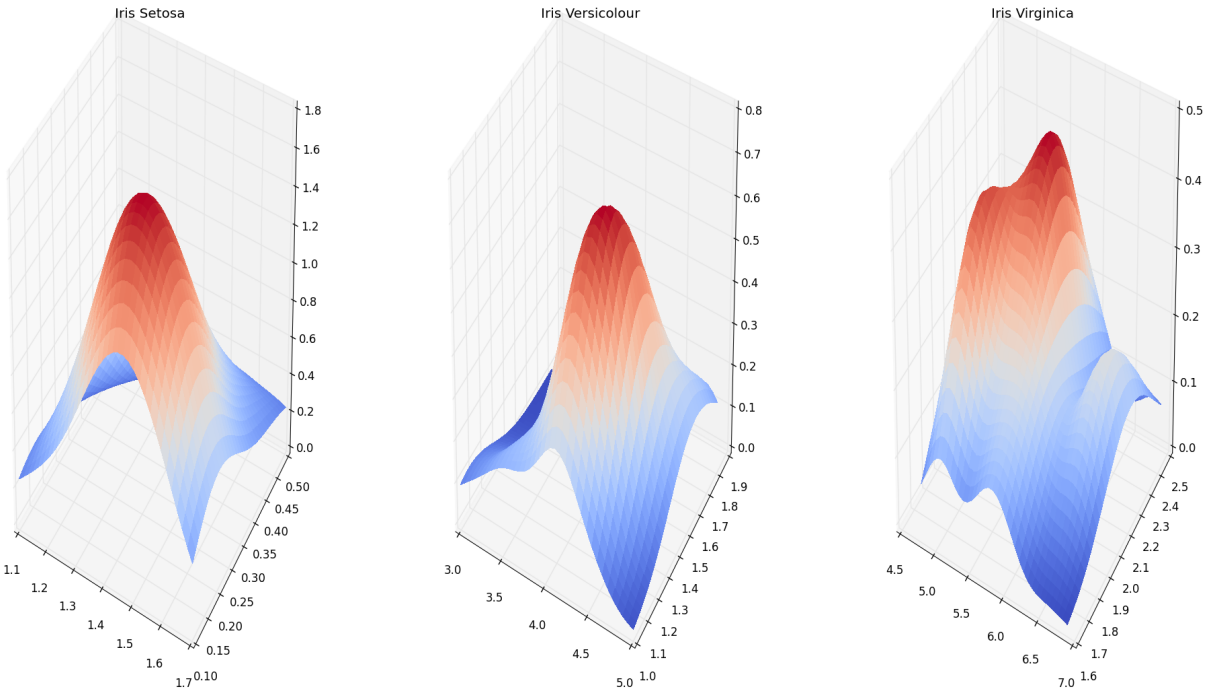
Рис. 6: Стандартное нормальное распределение, гауссово ядро. Справа изображен график модуля ошибки



Ширина окна	1.2
Количество элементов в выборке	2000 (1000 для одной гауссианы и 1000 для другой)

3. Двумерные плотности распределения для ирисов Фишера. Исходя из вида гистограмм 1 и 2, в качестве признаков взяты длина и ширина лепестка.

	Ширина окна
Iris Setosa	0.1
Iris Versicolour	0.2
Iris Virginica	0.2



4. Классификация ирисов с помощью Байесовского решающего правила. В качестве признаков взяты длина и ширина лепестка. Тестовая выборка - 60 элементов (40% представленных данных).
Результат: 4 ошибки на тестовой выборке.

4 Выводы

Решены поставленные задачи визуализации набора данных "ирисы Фишера"и оценки плотностей распределений с использованием разных типов ядер. Также достигнуто хорошее качество классификации ирисов.
Качество оценки плотности зависит от выбора ядра. Точность оценки во многом определяется шириной окна. При малых значениях ширины окна зависимость претерпевает резкие скачки (т.к. в этом случае оценка опирается лишь на небольшое число наблюдений из небольшой окрестности точки). В противном случае, при увеличении ширины окна функция сильно сглаживается.