Лекция 8.

Восстановление плотности.

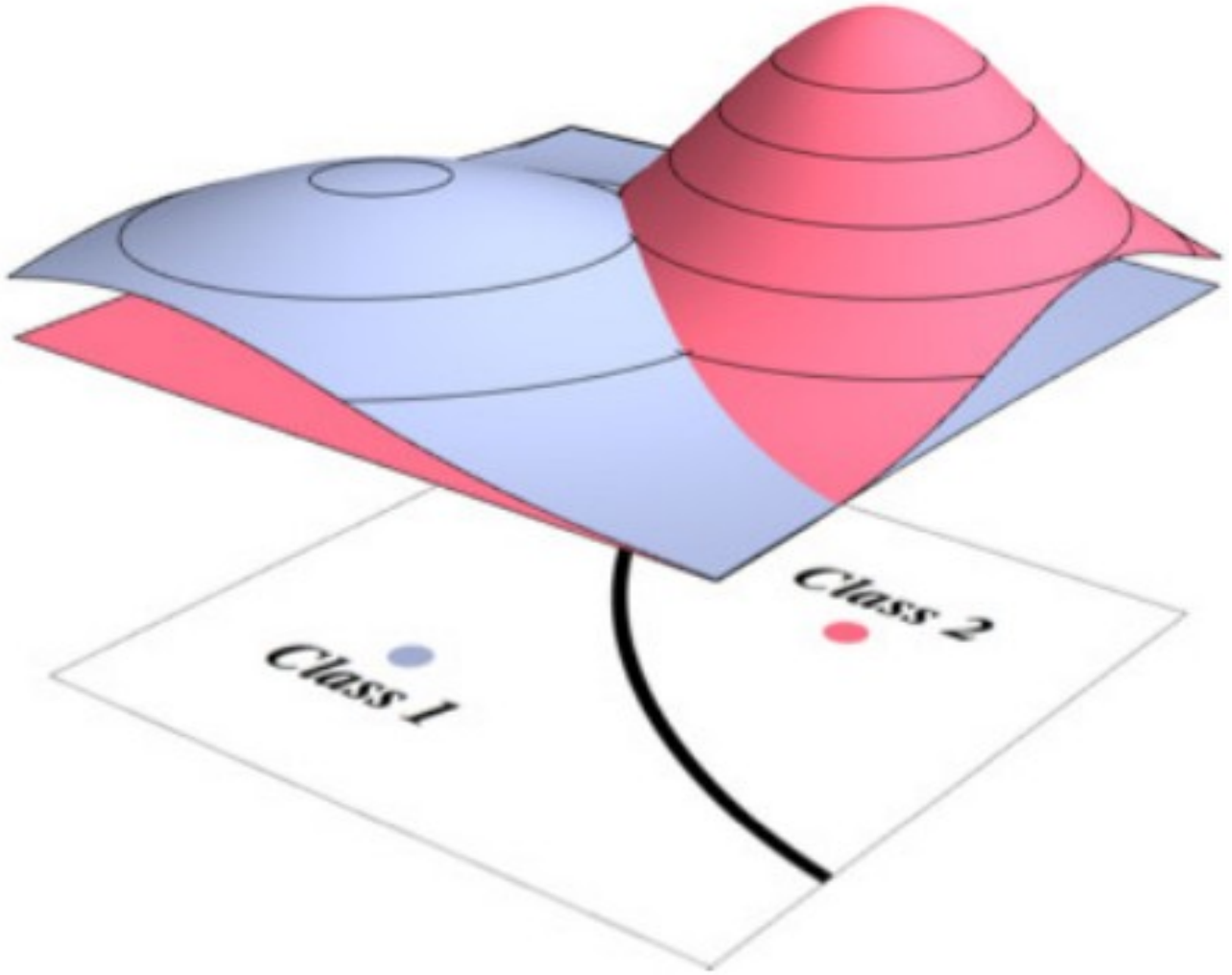
Павел Владимирович
Слипенчук
PavelMSTU@stego.su
ИУ-8

Москва
28.10.2016

Подходы к классификации

- **Первый подход**. Строим набор *гиперплоскостей* (возможно одну), разделяющую *пространство признаков*.
- **Первый подход А**. Строим дерево решений (по сути частный случай набора гиперплоскости)
- **Второй подход.** Байесовский-вероятностный
- **Второй подход А***. Восстанавливаем плотность.
(*Формально сводиться к Байесу)

Плотность вероятности



Способы восстановления плотности

- **«Честный способ»**. Для каждой точки пространства признаков строим «гравитационное поле» (аналог в физике) для **каждого** класса. Считаем «потенциал» для каждого «гравитационного поля». Какое «гравитационное поле» окажется сильнее – к тому классу и определяем точку.
- **Достоинства и недостатки?**

Способы восстановления плотности

- **«Честный способ»**. Для каждой точки пространства признаков строим «гравитационное поле» (аналог в физике) для **каждого** класса. Считаем «потенциал» для каждого «гравитационного поля». Какое «гравитационное поле» окажется сильнее – к тому классу и определяем точку.
- **Достоинства:**
  1) это «честный способ»;
  2) если признаки *информативны* – будет работать;
  3) нет проблемы разномощных классов – просто маломощным классам увеличивем вес («массу»);
  4) ясно и понятно.
  5) можно решить проблему «разряженных» выборок
- **Недостатки:**
  1) проклятье размерности.
  2) много точек – долго вычислять.

Способы восстановления плотности

- **Упрощение**: k ближайших соседей
- **Упрощение**: Берем точку. Берем расстояние **r** и все соседи заменяются на геометрический центр.
- **Упрощение**: «плотное» множество точек заменяем одной точкой и помещаем ее в *центр масс (**что это?**)*. Вес точки («масса») равен количеству точек.
- **Упрощение**: проводим разбиновку. В каждом бине определяем центр (геометрический или центр масс). Предварительно рассчитываем класс для каждого центра.

Способы восстановления плотности

- На самом деле не обязательно брать «закон всемирного тяготения».Можно взять любой другой закон, главное, чтобы «притяжение» монотонно ~~убывало~~ не увеличивалось от расстояния.
- **Упрощение.** Можно векторно не складывать, а складывать «потенциалы» от каждой точки
- **Упрощение**. Если классов два, то можно потенциал брать со знаком «+», если один класс и со знаком «-», если другой класс.
- Функцию «потенциала» называют **ядром**.

Метод парзеновского окна

- **Парзеновская плотность**:

$$p(\boldsymbol{x}; y_i, h) = c(h, y_i) \cdot \sum_{\forall \boldsymbol{x}: y(\boldsymbol{x_i}) = y_i} K\left(\frac{r(\boldsymbol{x}, \boldsymbol{x_i})}{h}\right)$$

h – «парзеновское окно». Параметр системы

$r(\boldsymbol{x}, \boldsymbol{x_i})$ – расстояние между двумя точками в пространстве признаков

$c(h, y_i)$ – функция нормировки. Обычно: $c(h, y_i) = \frac{1}{l_y \cdot V(h)}$

$K(r(\boldsymbol{x}, \boldsymbol{x_i})/h) = K(r_0)$ – ядро

Виды ядер

- Прямоугольное:

$$K(r_0)=\Pi(r_0)=\frac{1}{2},\quad r_0\le 1\ \ иначе\ \ 0$$

- Треугольное:

$$K(r_0)=T(r_0)=(1-r_0),\quad r_0\le 1\ \ иначе\ \ 0$$

- Гауссовское:

$$K(r_0)=G(r_0)=(2\pi)^{-½}\cdot e^{-½\cdot r_0^2}$$

- Квартическое:

$$K(r_0)=Q(r_0)=\frac{15}{16}\cdot(1-r_0^2)^2,\quad r_0\le 1\ \ иначе\ \ 0$$

- Епанчикова:

$$K(r_0)=E(r_0)=\frac{3}{4}\cdot(1-r_0^2),\quad r_0\le 1,\ \ иначе\ \ 0$$

Разбиновка

- **Достоинства**:

  1) нужно вычислить один раз, далее алгоритм работает быстро.

  2) можно выявить слабые бины. (далее: бэггинг)

- **Недостатки**:

  1) менее точный метод

  2) разбиновка – отдельная задача

Достоинства и недостатки
восстановления плотности
vs
разбиения пространства признаков

?

Достоинства и недостатки

- **Достоинства восстановления плотности**:

  1) нет предварительного обучения (нужны только данные и мы готовы к работе)

  2) быстрый учет feedback-а (нужно всего лишь добавить новую точку в данные)

  3) более точный метод чем разбиновка

- **Достоинства разбиения пространства**:

  1) быстрая работа после предварительного обучения.