Статистические методы распознавания образов

Лекция N 7 курса “Современные задачи теоретической информатики”

Юрий Лифшиц
yura@logic.pdmi.ras.ru

ИТМО

Осень’2005

Чем более подходящими шаблонами вы пользуетесь, тем более успешными будут ваши решения. Это обнадеживающее наблюдения для поборников искусственного интеллекта, так как, разумеется, компьютеры могут быть обучены распознаванию образов. Ясно, что успешные компьютерные программы, помогающие банкам — рассматривать заявки на кредиты, докторам — ставить диагнозы, а пилотам — приземлять самолеты, в каком-то смысле основаны на распознавании образов... Мы должны уделить намного больше внимания непосредственно распознаванию образов.

Герберт Саймон, нобелевский лауреат

План лекции

1. **Общие принципы распознавания образов**
   - Постановка и применения
   - Методы распознавания
2. **Курс математической статистики в 5 слайдах**
3. **Статистические методы распознавания образов**
4. Задача

План лекции

1. **Общие принципы распознавания образов**
   - Постановка и применения
   - Методы распознавания
2. Курс математической статистики в 5 слайдах
3. Статистические методы распознавания образов
4. Задача

Приложения

Биоинформатика	Поиск шаблонов в ДНК
Базы данных	Поиск и классификация
Обработка текстов	Тематическая классификация
Анализ изображений	Распознавание букв
Производство	Контроль качества
Поиск по мультимедиа	Определение жанров, ...
Биометрия	Отпечатки пальцев,...
Прогнозирование	Погода, сейсмология, геология
Обработка речи	Перевод аудио в текст

Распознавание образов по шагам

1. Восприятие образа (измерения)
2. Предварительная обработка
3. Выделение характеристик (индексация)
4. Классификация (принятие решения)

Разработка системы распознавания

Что нужно сделать для построения системы распознавания образов?

Разработка системы распознавания

Что нужно сделать для построения системы распознавания образов?

1. Достать тренировочную коллекцию
2. Выбрать модель представления объектов
3. Выбрать значимые характеристики
4. Разработать классифицирующее правило
5. Обучение алгоритма
6. Проверить качество. Вернуться к шагу 2 (3,4)...
7. Оптимизация алгоритма

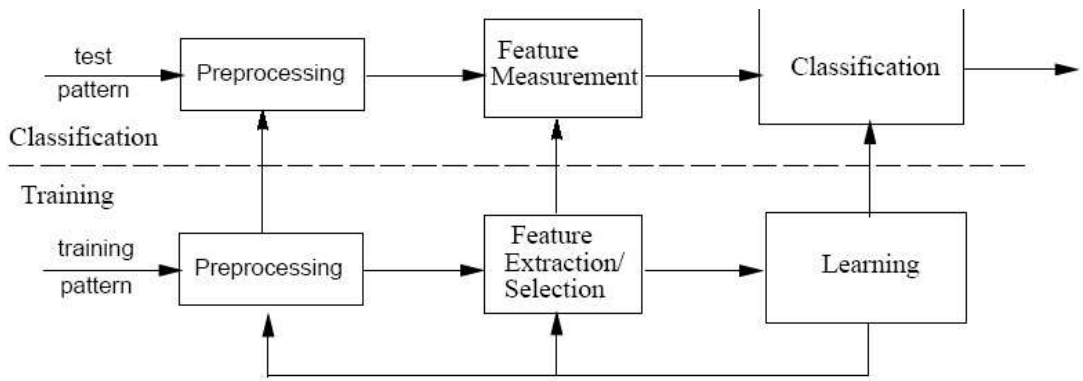


Figure 1: Model for statistical pattern recognition.

Основные методы

Выделяют 4 группы методов:

- Сравнение с образцом
  - Применяем геометрическую нормализацию и считаем расстояние до прототипа

Основные методы

Выделяют 4 группы методов:

- Сравнение с образцом
  - Применяем геометрическую нормализацию и считаем расстояние до прототипа
- Статистические методы
  - Строим распределение для каждого класса и классифицируем по правилу Байеса

Основные методы

Выделяют 4 группы методов:

- Сравнение с образцом
  - Применяем геометрическую нормализацию и считаем расстояние до прототипа
- Статистические методы
  - Строим распределение для каждого класса и классифицируем по правилу Байеса
- Нейронные сети
  - Выбираем вид сети и настраиваем коэффициэнты

Основные методы

Выделяют 4 группы методов:

- Сравнение с образцом
  - Применяем геометрическую нормализацию и считаем расстояние до прототипа
- Статистические методы
  - Строим распределение для каждого класса и классифицируем по правилу Байеса
- Нейронные сети
  - Выбираем вид сети и настраиваем коэффициэнты
- Структурные и синтаксические методы
  - Разбираем объект на элементы. Строим правило, в зависимости от вхождения/невхождения отдельных элементов и их последовательностей

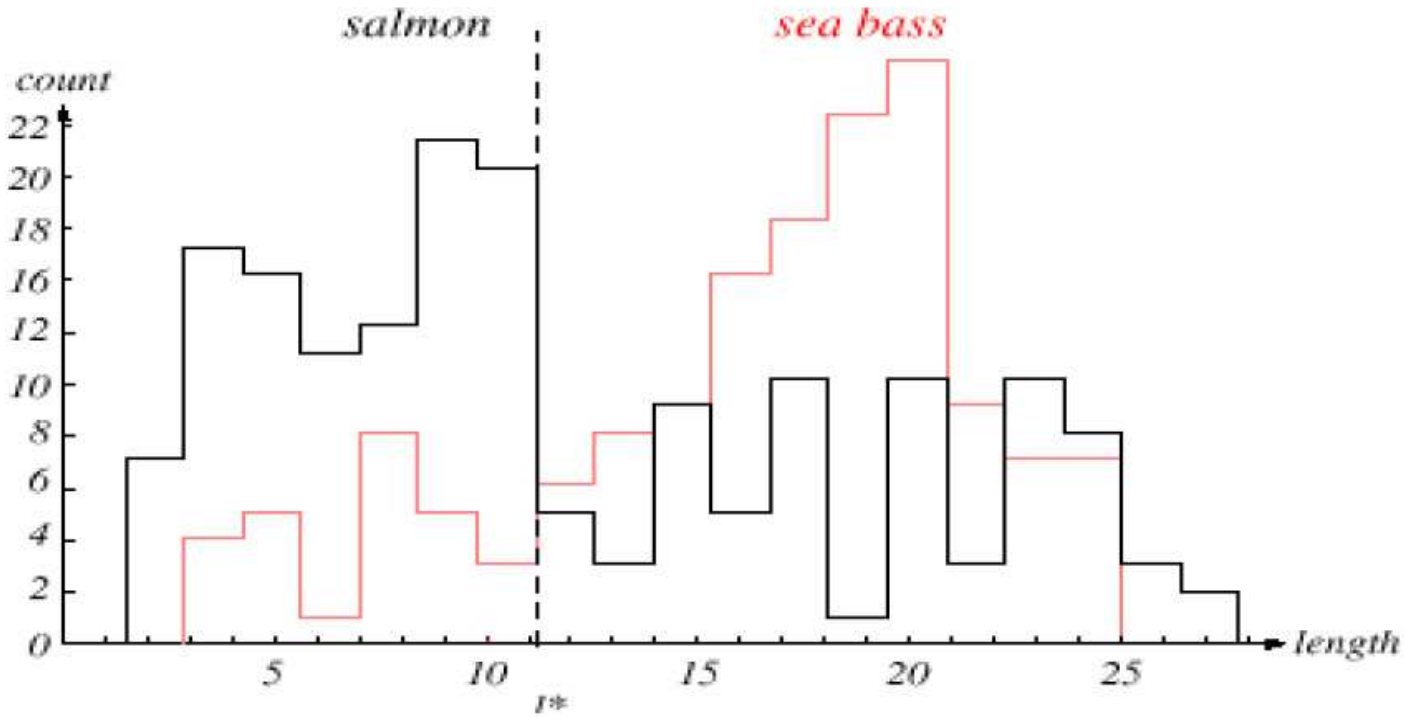
Основные методы

Выделяют 4 группы методов:

- Сравнение с образцом
  - Применяем геометрическую нормализацию и считаем расстояние до прототипа
- Статистические методы
  - Строим распределение для каждого класса и классифицируем по правилу Байеса
- Нейронные сети
  - Выбираем вид сети и настраиваем коэффициэнты
- Структурные и синтаксические методы
  - Разбираем объект на элементы. Строим правило, в зависимости от вхождения/невхождения отдельных элементов и их последовательностей

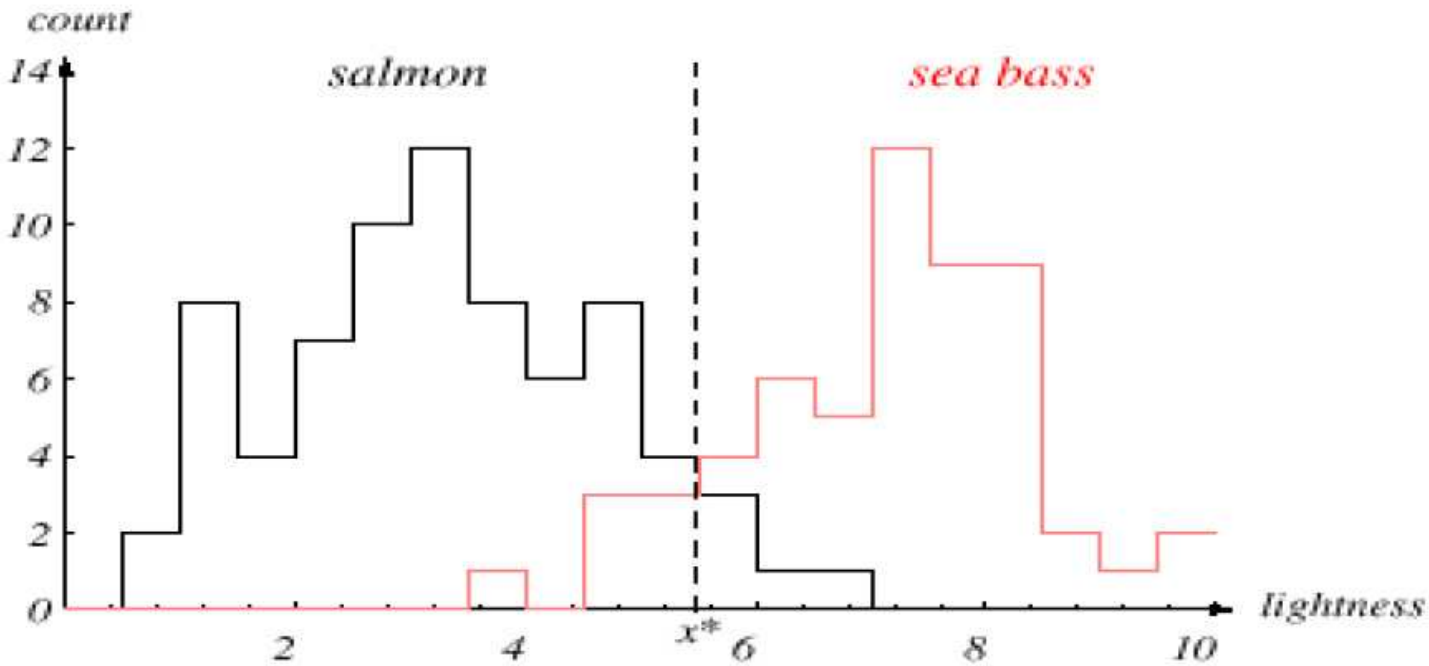
В какие группы можно отнести методы из прошлой лекции?

Пример



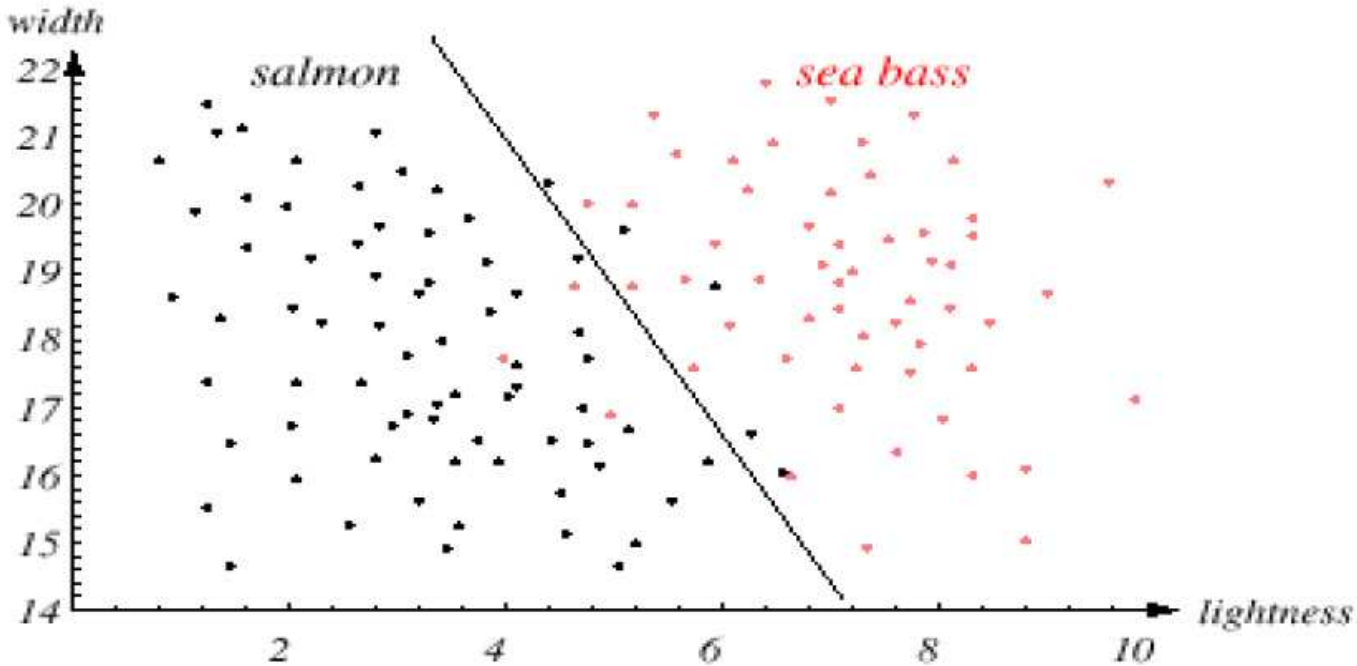
Пытаемся различить лосося и морского окуня. Длина — не идеальная характеристика.

Пример II



Окраска также не позволяет разделить эти два вида

Пример III



Комбинация двух характеристик дает лучший результат.

План лекции

Распределение (плотность распределения):

Каждому значению сопоставляет его вероятность

Основные понятия

Распределение (плотность распределения):

Каждому значению сопоставляет его вероятность

Выборка

Набор значений случайной величины $X_1, \ldots, X_n$

Основные понятия

Распределение (плотность распределения):
Каждому значению сопоставляет его вероятность

Выборка
Набор значений случайной величины $X_1, \ldots, X_n$

(Статистическая) оценка
Любая функция от выборки $T(X_1, \ldots, X_n)$

Две задачи мат. статистики

Оценка параметров

Предполагаем, что случайная величина принадлежит

Две задачи мат. статистики

Оценка параметров

Предполагаем, что случайная величина принадлежит известному распределению с неизвестными параметрами

Две задачи мат. статистики

Оценка параметров

Предполагаем, что случайная величина принадлежит известному распределению с неизвестными параметрами
Даем оценку параметрам

Две задачи мат. статистики

Оценка параметров

Предполагаем, что случайная величина принадлежит известному распределению с неизвестными параметрами

Даем оценку параметрам

Оценки: точечные и интервальны

Две задачи мат. статистики

Оценка параметров

Предполагаем, что случайная величина принадлежит известному распределению с неизвестными параметрами

Даем оценку параметрам

Оценки: точечные и интервальны

Требования к точечным: несмещенность и состоятельность

Две задачи мат. статистики

Оценка параметров

Предполагаем, что случайная величина принадлежит известному распределению с неизвестными параметрами

Даем оценку параметрам

Оценки: точечные и интервальны

Требования к точечным: несмещенность и состоятельность

Проверка гипотез

Не знаем распределения

Две задачи мат. статистики

Оценка параметров

Предполагаем, что случайная величина принадлежит известному распределению с неизвестными параметрами

Даем оценку параметрам

Оценки: точечные и интервальны

Требования к точечным: несмещенность и состоятельность

Проверка гипотез

Не знаем распределения

Делаем предположение о распределении (H_0)

Две задачи мат. статистики

Оценка параметров

Предполагаем, что случайная величина принадлежит известному распределению с неизвестными параметрами

Даем оценку параметрам

Оценки: точечные и интервальны

Требования к точечным: несмещенность и состоятельность

Проверка гипотез

Не знаем распределения

Делаем предположение о распределении (H_0)

По выборке принимаем/отвергаем H_0

Две задачи мат. статистики

Оценка параметров

Предполагаем, что случайная величина принадлежит известному распределению с неизвестными параметрами
Даем оценку параметрам
Оценки: точечные и интервальны
Требования к точечным: несмещенность и состоятельность

Проверка гипотез

Не знаем распределения
Делаем предположение о распределении (H_0)
По выборке принимаем/отвергаем H_0
Ошибка первого рода (отвергли правильное распределение) должна быть меньше α

Правило Байеса I

Данные:

Два распределения A и B

Значение X, порожденное одной из этих случайных величин

Определить: какой из них?

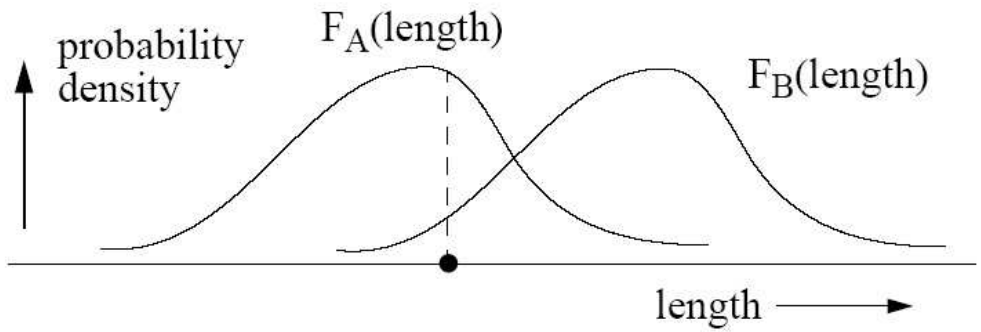
Правило Байеса I

Данные:

Два распределения A и B

Значение X, порожденное одной из этих случайных величин

Определить: какой из них?



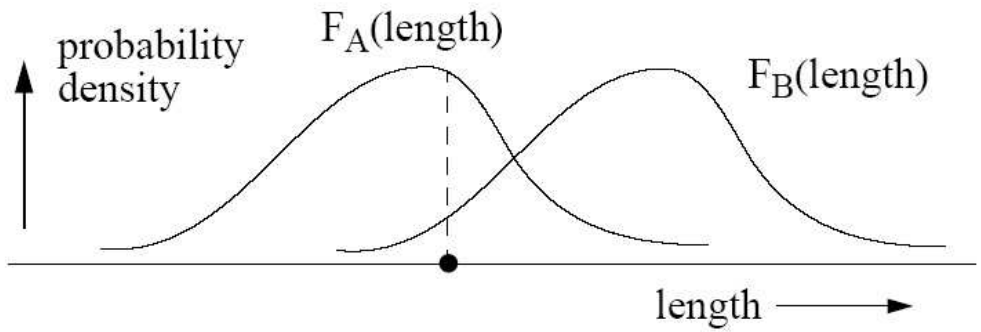
Правило Байеса I

Данные:

Два распределения A и B

Значение X, порожденное одной из этих случайных величин

Определить: какой из них?



Интуиция: Надо выбирать A, если

$$Prob(A|X) > Prob(B|X)$$

Правило Байеса II

Немного преобразований:

$$Prob(A|X)Prob(X) = Prob(A, X) = Prob(A)Prob(X|A)$$

Правило Байеса II

Немного преобразований:

$$Prob(A|X)Prob(X) = Prob(A, X) = Prob(A)Prob(X|A)$$

$$\begin{aligned} Prob(A|X) &= \frac{Prob(X|A)Prob(A)}{Prob(X)} \\ &= \frac{Prob(X|A)Prob(A)}{Prob(X|A)Prob(A) + Prob(X|B)Prob(B)} \end{aligned}$$

Правило Байеса II

Немного преобразований:

$$Prob(A|X)Prob(X) = Prob(A,X) = Prob(A)Prob(X|A)$$

$$\begin{aligned} Prob(A|X) &= \frac{Prob(X|A)Prob(A)}{Prob(X)} \\ &= \frac{Prob(X|A)Prob(A)}{Prob(X|A)Prob(A) + Prob(X|B)Prob(B)} \end{aligned}$$

$$\boxed{Prob(A|X) = \frac{F_A(X)P_A}{F_A(X)P_A + F_B(X)P_B}}$$

Правило Байеса II

Немного преобразований:

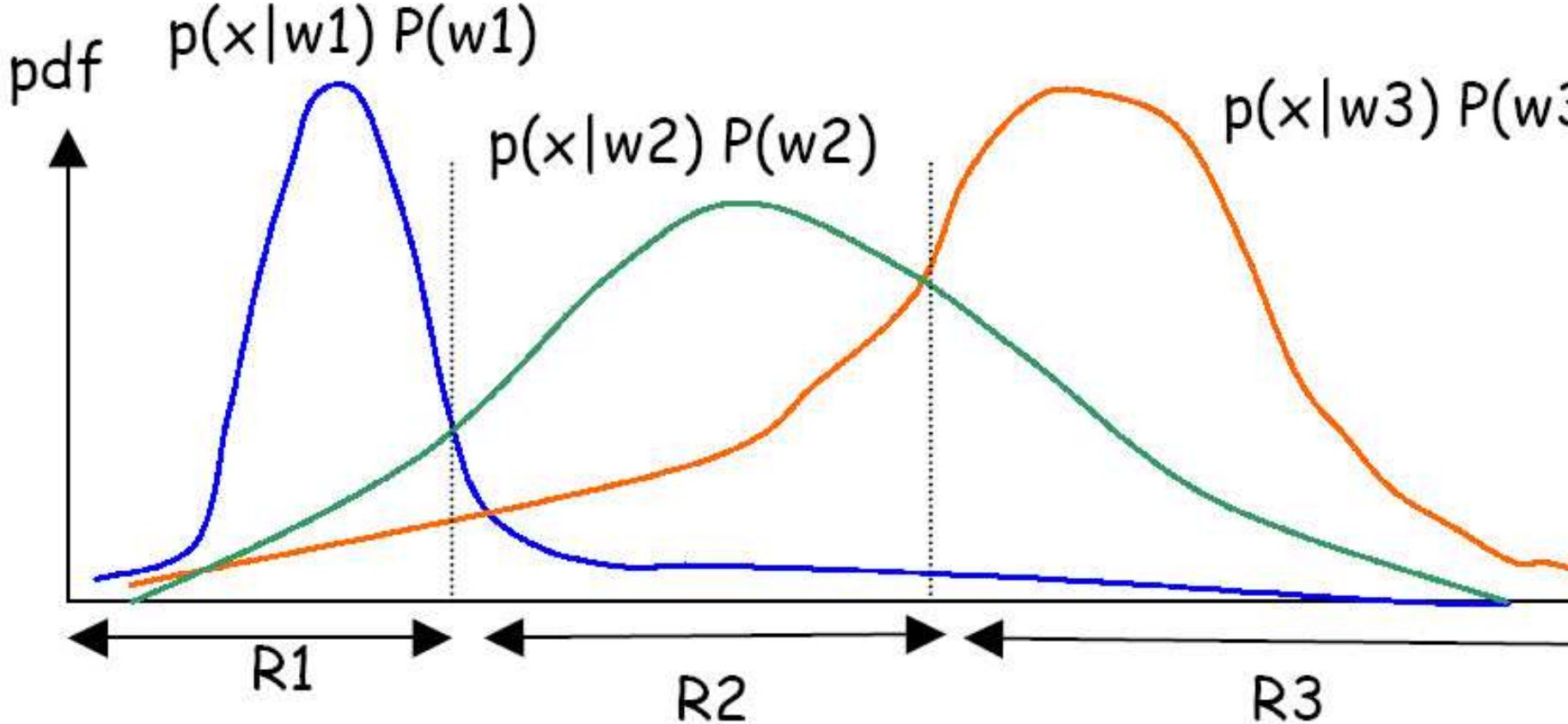
$$Prob(A|X)Prob(X) = Prob(A,X) = Prob(A)Prob(X|A)$$

$$\begin{aligned} Prob(A|X) &= \frac{Prob(X|A)Prob(A)}{Prob(X)} \\ &= \frac{Prob(X|A)Prob(A)}{Prob(X|A)Prob(A) + Prob(X|B)Prob(B)} \end{aligned}$$

$$\boxed{Prob(A|X) = \frac{F_A(X)P_A}{F_A(X)P_A + F_B(X)P_B}}$$

Правило Байеса: Надо выбирать A, если

$$F_A(X)P_A > F_B(X)P_B)$$


pdf
p(x|w1) P(w1)
p(x|w2) P(w2)
p(x|w3) P(w3
R1
R2
R3

Критерий χ^2

Постановка задачи

Проверяемой распределение D

Выборка $X_1, \ldots, X_n$

Принять/отвергнуть: “выборка принадлежит D”

Критерий χ^2

Постановка задачи

Проверяемой распределение D

Выборка $X_1, \ldots, X_n$

Принять/отвергнуть: "выборка принадлежит D"

Вычисления

Разбиваем область значений на m классов

Пусть n_j — кол-во элементов выборки в классе j

Пусть p_j — вероятность попасть в класс j

согласно D. Обозначим $n'_j = np_j$

Вычисляем функцию:

$$T = \sum_{j-1}^{m} \frac{(n_j - n'_j)^2}{n'_j}$$

Критерий χ^2

Постановка задачи

Проверяемой распределение D

Выборка $X_1, \ldots, X_n$

Принять/отвергнуть: "выборка принадлежит D"

Вычисления

Разбиваем область значений на m классов

Пусть n_j — кол-во элементов выборки в классе j

Пусть p_j — вероятность попасть в класс j

согласно D. Обозначим $n'_j = np_j$

Вычисляем функцию:

$$T = \sum_{j-1}^{m} \frac{(n_j - n'_j)^2}{n'_j}$$

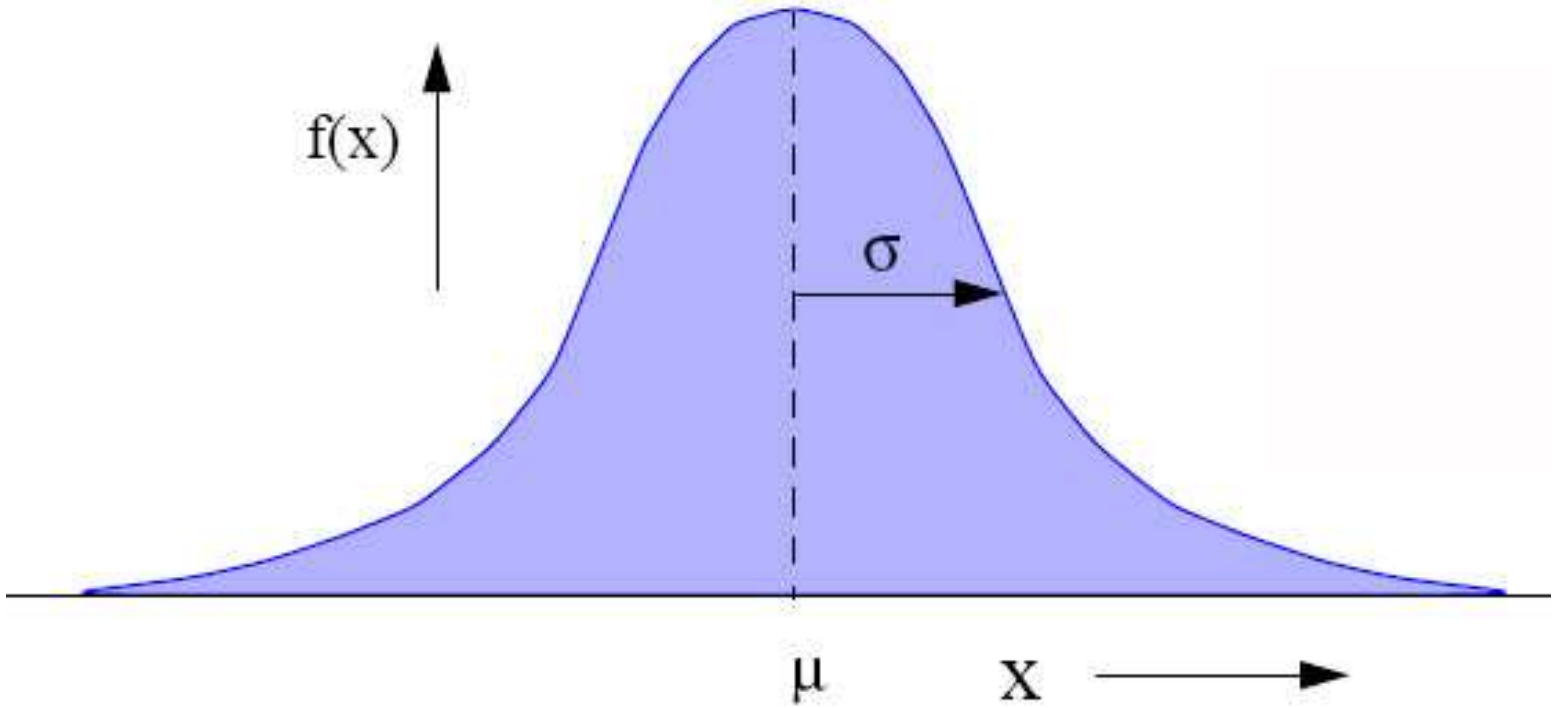
Критический уровень

Если $T < t$, принимаем гипотезу, $T \geq t$ — отвергаем

Значение t определяется как функция от α и k

Пример: $\alpha = 0.05,\ k = 5 \Rightarrow t = 11.1$

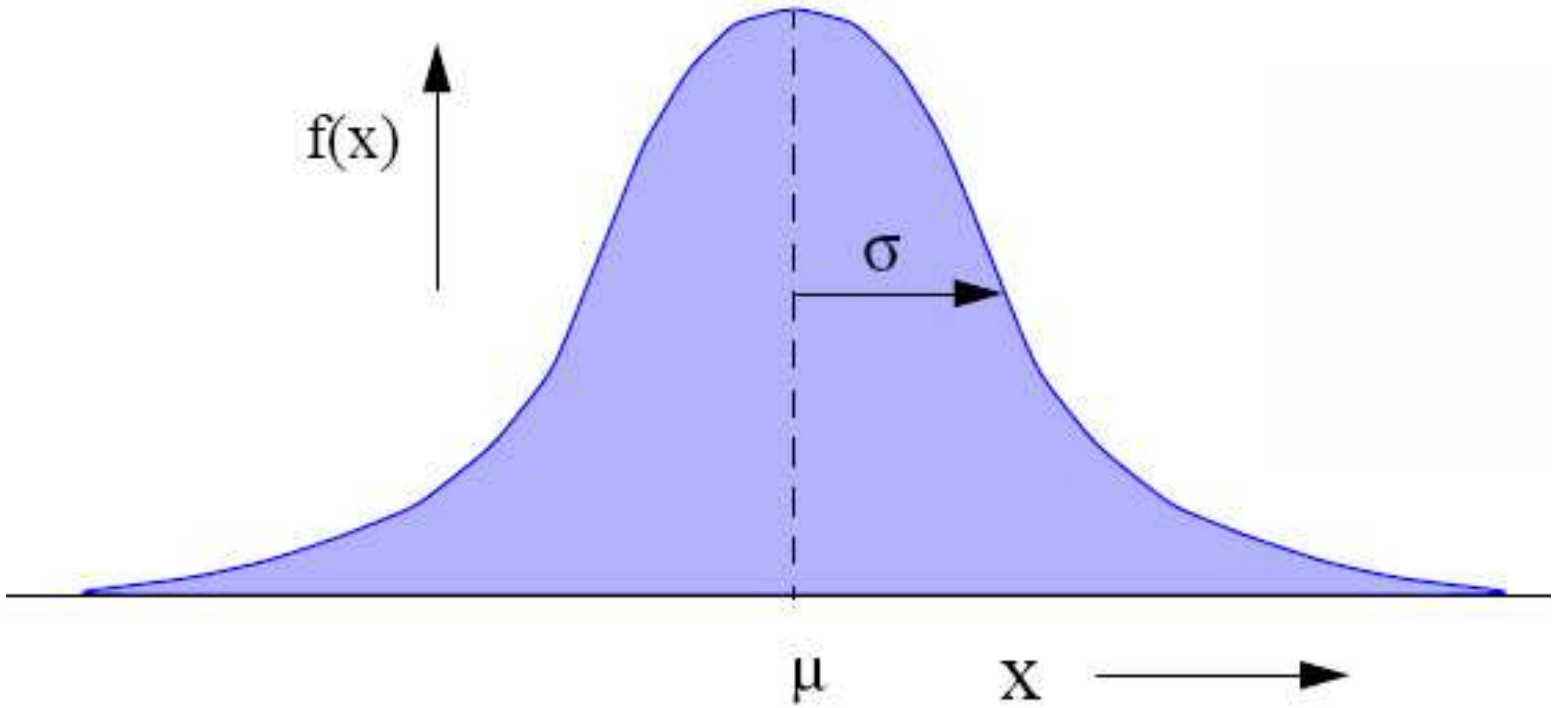
Нормальное распределение



Функция плотности нормального распределения:

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

Нормальное распределение



Функция плотности нормального распределения:

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

Два параметра: μ - математическое ожидание и σ — дисперсия

План лекции

Общая идеология

Задача

Дана тренировочная коллекция

Каждый объект = набор n характеристик = n-мерный вектор

Построить классифицирующее правило

Общая идеология

Задача

Дана тренировочная коллекция

Каждый объект = набор n характеристик = n-мерный вектор

Построить классифицирующее правило

Предпосылки

Считаем, что элементы каждой категории имеют свое распределение в n-мерном пространстве

Будем принимать решение по правилу Байеса!

Общая идеология

Задача

Дана тренировочная коллекция

Каждый объект = набор n характеристик = n-мерный вектор

Построить классифицирующее правило

Предпосылки

Считаем, что элементы каждой категории имеют свое распределение в n-мерном пространстве

Будем принимать решение по правилу Байеса!

Чего не хватает для полного счастья?

Общая идеология

Задача

Дана тренировочная коллекция

Каждый объект = набор n характеристик = n-мерный вектор

Построить классифицирующее правило

Предпосылки

Считаем, что элементы каждой категории имеют свое распределение в n-мерном пространстве

Будем принимать решение по правилу Байеса!

Чего не хватает для полного счастья?

Ответ: функций распределения для каждой категории

Строим функции распределения

Нам известны функции распределения

Просто используем правило Байеса

Строим функции распределения

Нам известны функции распределения

Просто используем правило Байеса

Известен тип, но не параметры

Пример: нормальное распределение с неизвестными μ, σ

Используем точечные оценки: $\hat{\mu} = \frac{X_1 + \cdots + X_n}{n}$

$\hat{\sigma} = \sqrt{\frac{1}{n-1} \sum (X_i - \hat{\mu})^2}$

Строим функции распределения

Нам известны функции распределения

Просто используем правило Байеса

Известен тип, но не параметры

Пример: нормальное распределение с неизвестными μ, σ

Используем точечные оценки: $\hat{\mu} = \frac{X_1 + \cdots + X_n}{n}$

$\hat{\sigma} = \sqrt{\frac{1}{n-1} \sum (X_i - \hat{\mu})^2}$

В чем разница с одномерным методом Rocchio?

Строим функции распределения

Нам известны функции распределения

Просто используем правило Байеса

Известен тип, но не параметры

Пример: нормальное распределение с неизвестными μ, σ

Используем точечные оценки: $\hat{\mu} = \frac{X_1 + \cdots + X_n}{n}$

$\hat{\sigma} = \sqrt{\frac{1}{n-1} \sum (X_i - \hat{\mu})^2}$

В чем разница с одномерным методом Rocchio?

Неизвестное распределение

Мы должны построить его по тренировочной коллекции!

Метод гистограмм

Построим функцию распределения “в лоб”

Разобъем n-мерное пространство на клеточки

для каждой клетки определим плотность распределения

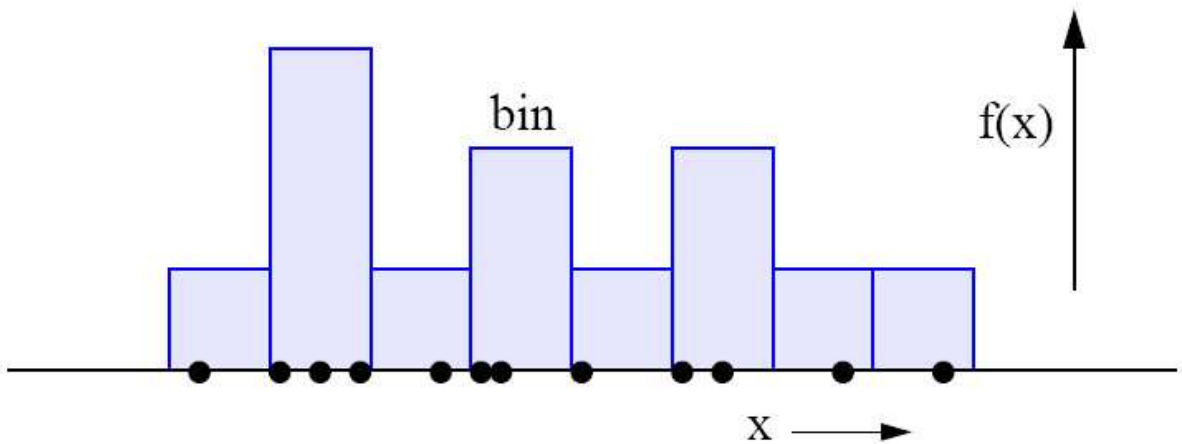
как долю всех документов, попавших в клетку

Метод гистограмм

Построим функцию распределения “в лоб”

Разобъем n-мерное пространство на клеточки

для каждой клетки определим плотность распределения как долю всех документов, попавших в клетку

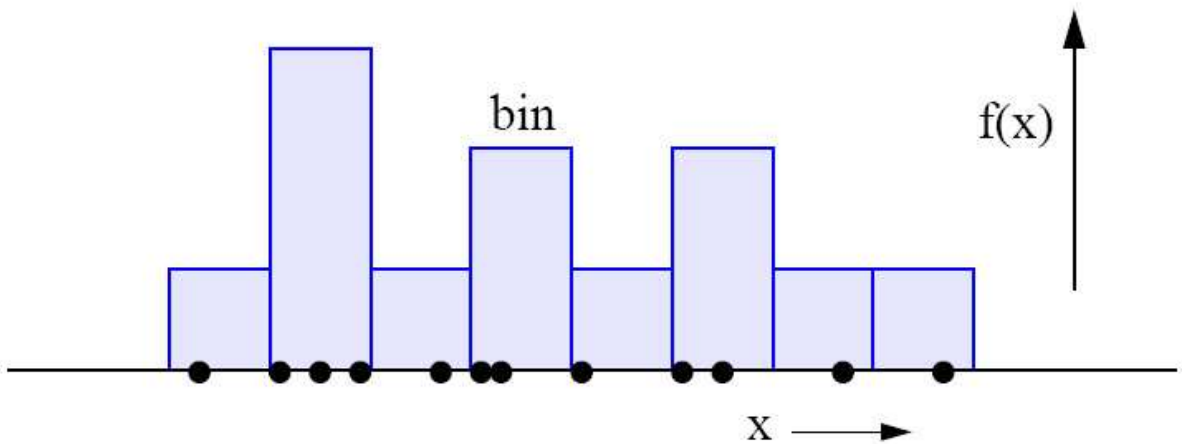


Метод гистограмм

Построим функцию распределения “в лоб”

Разобъем n-мерное пространство на клеточки

для каждой клетки определим плотность распределения как долю всех документов, попавших в клетку



Недостаток: для большой размерности нужна огромная тренировочная коллекция

Метод Парзена

Идея метода

Для каждой точки из класса построим функцию, достигающую максимума в этой точке и быстро убывающей при удалении от нее

Сложим такие функции для всех точек

Метод Парзена

Идея метода

Для каждой точки из класса построим функцию, достигающую максимума в этой точке и быстро убывающей при удалении от нее
Сложим такие функции для всех точек

Что это вам напоминает?

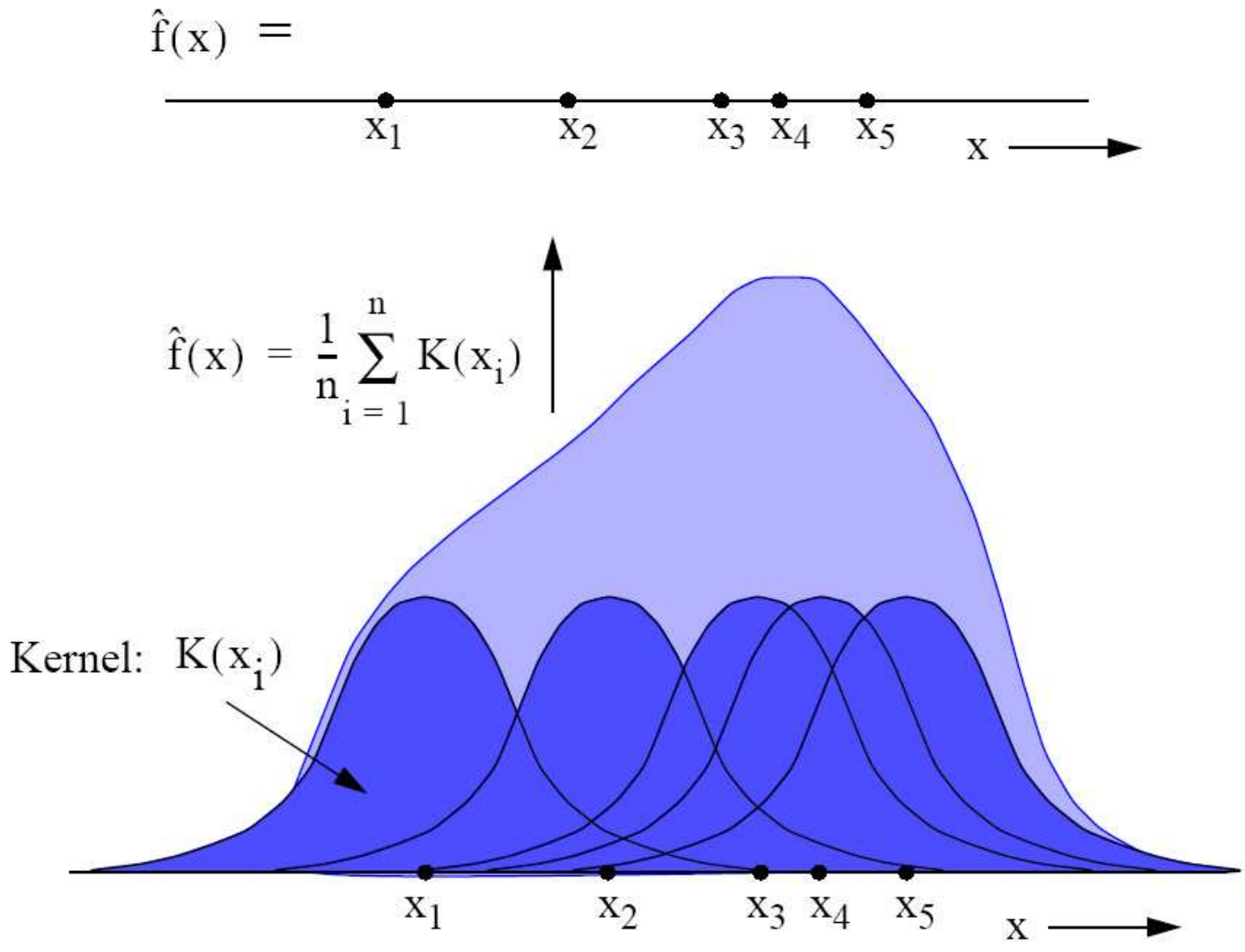
Метод Парзена

Идея метода

Для каждой точки из класса построим функцию, достигающую максимума в этой точке и быстро убывающей при удалении от нее

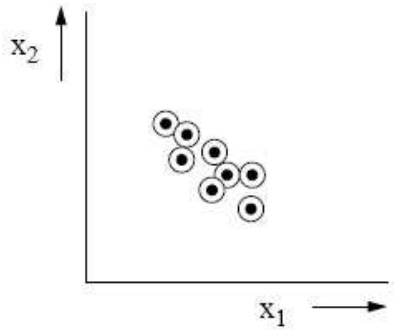
Сложим такие функции для всех точек

Что это вам напоминает? Сравните с kNN-методом.

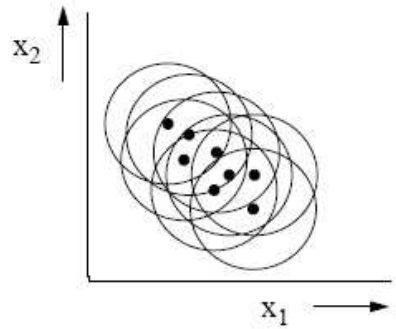

$\hat{f}(x) =$
x_1 x_2 x_3 x_4 x_5 x
$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K(x_i)$
Kernel: $K(x_i)$
x_1 x_2 x_3 x_4 x_5 x

Форма ядра

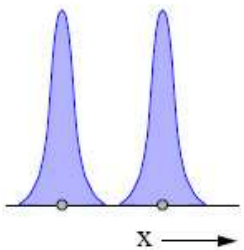
Как выбрать форму ядра? Слишком маленькая/большая форма может ухудшить качество распознавания:

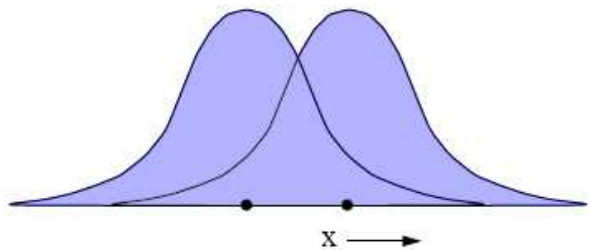


Too small: lot of empty space.



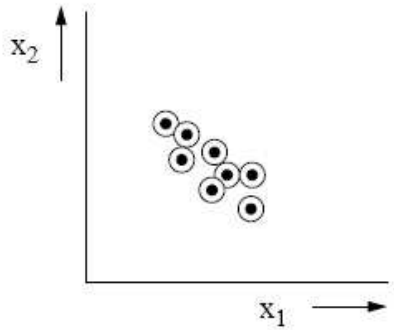
Too wide: estimated density is dominated by kernel shape.



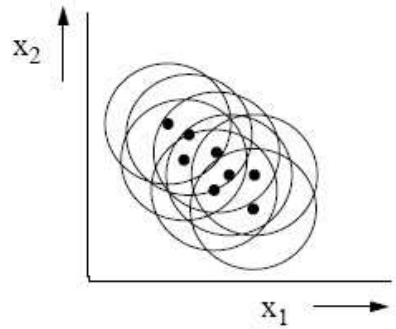


Форма ядра

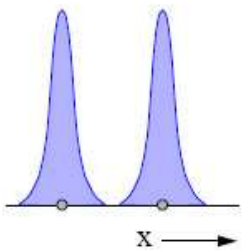
Как выбрать форму ядра? Слишком маленькая/большая форма может ухудшить качество распознавания:

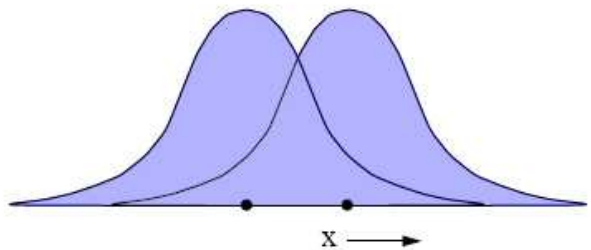


Too small: lot of empty space.



Too wide: estimated density is dominated by kernel shape.





Решение: выбирать радиус ядра, чтобы в него попало 5 ближайших тренировочных документов

План лекции

Открытый вопрос от А.Куликова

Пусть есть граф из n вершин, степень каждой вершины не больше трех. Для какой наименьшей функции $f(n)$ всегда можно разбить вершины на две группы по $n/2$ так, чтобы между ними было не более $f(n)$ ребер?

Открытый вопрос от А.Куликова

Пусть есть граф из n вершин, степень каждой вершины не больше трех. Для какой наименьшей функции $f(n)$ всегда можно разбить вершины на две группы по $n/2$ так, чтобы между ними было не более $f(n)$ ребер?

Гипотеза: $f(n) = c \cdot n$ для некоторого c

Открытый вопрос от А.Куликова

Пусть есть граф из n вершин, степень каждой вершины не больше трех. Для какой наименьшей функции $f(n)$ всегда можно разбить вершины на две группы по $n/2$ так, чтобы между ними было не более $f(n)$ ребер?

Гипотеза: $f(n) = c \cdot n$ для некоторого c

Нижние оценки. Можете ли придумать граф, в котором в любом разрезе будет хотя бы $\log n$ ребер?

Открытый вопрос от А.Куликова

Пусть есть граф из n вершин, степень каждой вершины не больше трех. Для какой наименьшей функции $f(n)$ всегда можно разбить вершины на две группы по $n/2$ так, чтобы между ними было не более $f(n)$ ребер?

Гипотеза: $f(n) = c \cdot n$ для некоторого c

Нижние оценки. Можете ли придумать граф, в котором в любом разрезе будет хотя бы $\log n$ ребер?

Задача имеет приложения в разработке эффективных алгоритмов

Последний слайд

Если не запомните ничего другого:

- Распознавание образов: восприятие, характеристик, классификация

Последний слайд

Если не запомните ничего другого:

- Распознавание образов: восприятие, характеристик, классификация
- Статистический метод: угадываем распределение и применяем правило Байеса

Последний слайд

Если не запомните ничего другого:

- Распознавание образов: восприятие, характеристик, классификация
- Статистический метод: угадываем распределение и применяем правило Байеса
- Метод Парзена: суммируем ядра вокруг всех тренировочных документов

Последний слайд

Если не запомните ничего другого:

- Распознавание образов: восприятие, характеристик, классификация
- Статистический метод: угадываем распределение и применяем правило Байеса
- Метод Парзена: суммируем ядра вокруг всех тренировочных документов

Последний слайд

Если не запомните ничего другого:

- Распознавание образов: восприятие, характеристик, классификация
- Статистический метод: угадываем распределение и применяем правило Байеса
- Метод Парзена: суммируем ядра вокруг всех тренировочных документов

Вопросы?