Статистические методы распознавания образов

Ю. Лифшиц*

6 декабря 2005 г.

План лекции

1. Общие принципы распознавания образов
   (a) Постановка и применения
   (b) Методы распознавания
2. Краткий курс математической статистики
3. Статистические методы распознавания образов
4. Задача

1 Общие принципы распознавания образов

1.1 Постановка и применения

Распознавание образов - процесс отнесения объекта по фиксированой группе его свойств к одному объекту из множества образов по заранее оговоренному правилу.

Например, пойманую сетями рыбу надо разделить на окуней и лососей. Предположим, что это делается по длине рыбы. Если вернуться к постановке задачи, то у нас есть объект "рыба", и по значению свойства "длина"мы относим рыбу либо к образу "лосось", либо к образу "окунь".

Распознование образов применяется в следующих областях:

- Биоинформатика: поиск шаблонов в ДНК.
- Базы данных: поиск и классификация.
- Обработка текстов: тематическая классификация.

*Законспектировал А. Вокин.

- Анализ изображений: распознавание символов, работа с картами, распознование лиц, разделение объектов.
- Производство: контроль качества (визуальная проверка корректности микросхем).
- Поиск по мультимедиа: определение жанров. К сожалению, на данных момент в этой области еще ничего не реализовано.
- Биометрия: Идентификация человека по отпечаткам пальцев, по радужной оболочке глаза.
- Прогнозирование: погода, сейсмология, геология.
- Обработка речи: перевод аудио в текст.

1.2 Процедура распознования

Выделим наиболее важные шаги в процедуре распознования:

1. Восприятие образа. На этом этапе производят получение значений характеристических свойств объекта (измерения линейных замеров, фотографирование, оцифровка звука).
2. Предварительная обработка (удаление шумов, представление изображения в чернобелом варианте, обрезание ненужных частей изображения).
3. Выделение характеристик (индексация). На этом этапе измеряются характеристический свойства объекта (измеряем длину рыбы и ее цвет).
4. Классификация (принятие решения).

1.3 Разработка системы распознавания

1. Достать тренировочную коллекцию
   Тренировочная коллекция - коллекция объектов для которых заведомо известны их образы. Например коллекция аудио записей для каждого звука, или коллекция изображений каждой буквы латинского алфавита.
2. Выбрать модель представления объектов
3. Выбрать значимые характеристики
   Один из самых важных этапов разработки системы распознавания. Например, если в случае идентификации рыбы окунь/лосось в качестве характеристики выбрать только длину рыбы, то никакое классифицирующее правило не сможет точно определить тип рыбы, поскольку весьма вероятно встретить лосося и окуня одинаковой длины.
4. Разработать классифицирующее правило
   Классифицирующее правило - правило, которое по значениям характеристических свойств объекта отнесет его к одному из образов.

5. Обучение алгоритма.
   На этом этапе алгоритм "собирает опыт"на основе респознавания тренировочной коллекции. Для того, чтобы правильно выставить коэффициенты (параметры) алгоритма его прогоняют на тренировочной коллекции контролируя результат работы алгоритма.

6. Проверить качество. Вернуться к шагу 2 (3, 4)...
   Если частота ошибок алгоритма не устраивает решаемую задачу, то необходимо вернуться к п. 2 (3, 4). Интуитивно понятно, что увеличение количества характеристических свойств, увеличение тренировочной коллекции улучшают качество работы алгоритма.

7. Оптимизация алгоритма
   После того, как качество работы алгоритма подходит под условие рассматриваемой задачи, иногда приходится произвести его оптимизацию. Изначальный алгоритм может быть слишком долгим или ресурсоемким. Ускорить алгоритм распознавания можно уменьшив количество характеристических свойств объекта, выбрав другие характеристические свойства, используя другое классифицирущее правило.

1.4 Методы распознавания

Выделяют 4 группы методов разпознавания:

1. Сравнение с образцом
   Применяем геометрическую нормализацию и считаем расстояние до прототипа. Наиболее наглядно применение этого метода в распознавании теста.

   Задача. У нас есть изображение сканированного символа и коллекция изображений образцов (всех букв алфавита), мы хотим определить, какой букве алфавита соответствует отсканированое изображение.

   Решение. Смасштабируем изображение символа до размеров образцов и выберем тот, до которого растояние минимально.

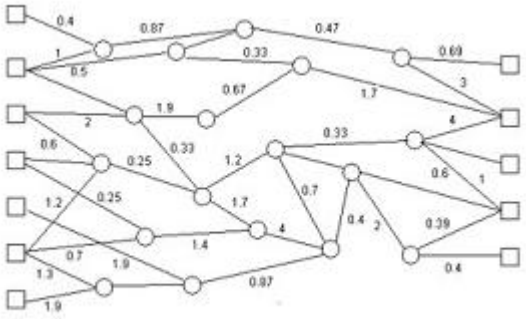
2. Статистические методы
   Строим распределение для каждого класса и классифицируем по правилу Байеса.

   Распределение можно построить используя тренировочную коллекцию.

3. Нейронные сети
   Выбираем вид сети и настраиваем коэффициенты.

   На рисунке представлена простейшая нейронная сеть. На вход нейронной сети подается распознаваемый объект. Слева на рисунке расположена группа **рецепторов**, каждый из которых отвечает за прием своего характеристического свойства распознаваемых объектов. Справа на рисунке расположенна группа **эффекторов**, каждый из которых соответствуею

одному из образов. Выбирается тот из эффекторов значение в котором максимально.

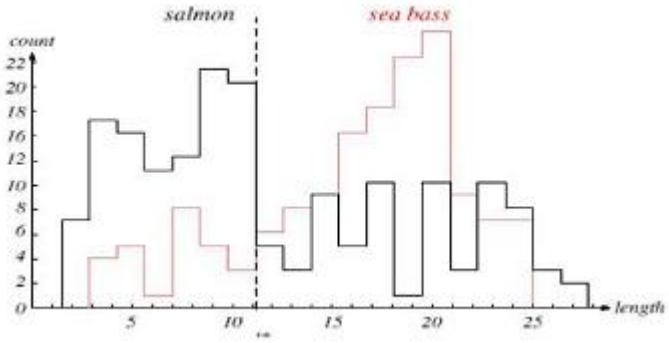


Настройка коэффицентов является фазой обучения алгоритма. На этом этапе мы настраиваетм коэффициенты таким образом, чтобы алгоритм правильно работал на образцах. Чем больше образцов, тем больше вероятность того, что алгоритм примет верное решение на остальных данных.

4. Структурные и синтаксические методы
   Разбираем объект на элементы. Строим правило, в зависимости от вхождения/невхождения отдельных элементов и их последовательностей

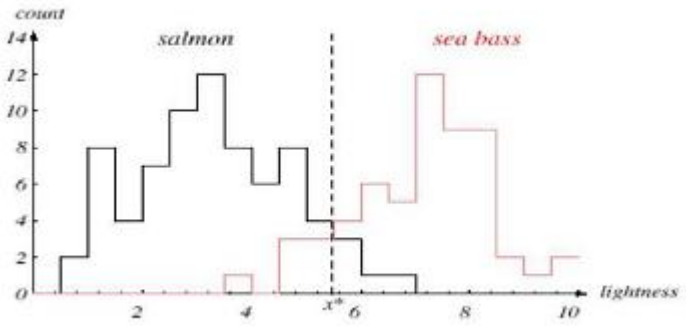
1.5 Пример

Предположим, нам необходимо определить вид рыбы, поданной на вход. Упростим задачу до случая 2 видов (окунь и лосось). Первое решение, которое приходит нам в голову - делать предположение о виде рыбы по ее длине.



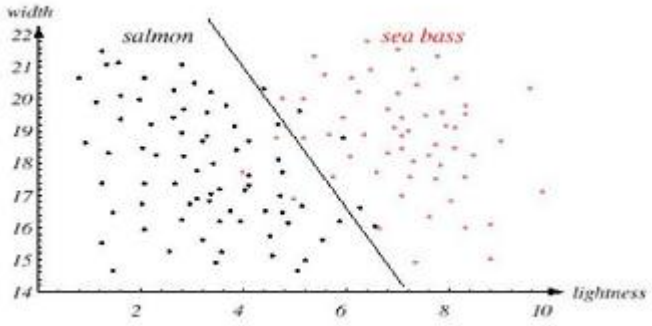
Как мы видем на графике, принимая решение по длине, у нас не получится достоверно определить вид рыбы. ть, длина - не идеальная характеристика.

Попробуем выбрать другое характериститческое свойство рыбы. Рассмотрим окрас.



Как мы видим на графике, руководствуясь только окрасом, мы все равно не научились определять вид рыбы.

Но если учитывать сразу два вышеперечисленных характеристических свойст рыбы, то мы сможем делать предположение о виде рыбы намного точнее.



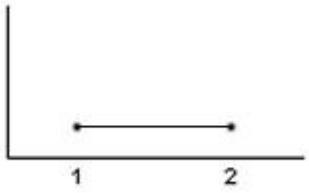
Дело в том, что когда мы рассматривали только одну характеристику рыбы мы могли проводить эту разделяющую прямую параллельно одной из осей координат. В то время, как в случае двух характеристик, мы можем контролировать и наклон этой разделяющей прямой.

2 Краткий курс математической статистики

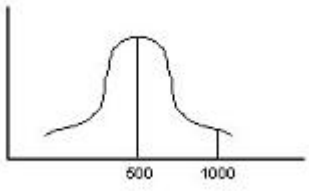
2.1 Основные понятия

1. Распределение (плотность распределения) - функция, которая каждому значению сопоставляет его вероятность.

   (a) Равномерное распределение

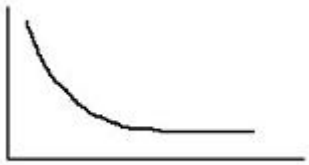


   (b) Нормальное распределение



   Например, вероятность того, что бросив монету 1000 раз, что выпало ровно n орлов?

   (c) Геометрическое распределение



   Например, вероятность того, что бросив монету 1000 раз, орел выпадет только на n-ый

2. Выборка - Набор значений случайной величины $X_1, \ldots, X_n$
3. (Статистическая) оценка - любая функция от выборки $T(X_1, \ldots, X_n)$

2.2 Две задачи математической статистики

1. **Оценка параметров**
   Пусть выборка принадлежит известному распределению с неизвестными параметрами. Нам необходимо дать оценку параметрам распределения. Например, дана выборка (7, 8, 5), необходимо выяснить на каком отрезке производился выбор случайных велечин, если известно, что распределение равномерно.

   Ответ на этот вопрос можно давать в двух видах: точечном и интервальном. Пример точной оценки параметров: [5, 8]. Пример интервальной оценки: левая граница отрезка лежит в интервале [3-5], правая в [8-12].

   Основные критерии для точечных оценок:

   (a) Несмещенность

   (b) Состоятельность

2. **Проверка гипотез**
   Часто на основе данных наблюдения нужно проверить те или иные предположения о распределении вероятностей экспериментальных данных. Например, является ли распределение нормальным, имеет ли заданное распределение фиксированные характеристики. Будем рассматривать двухальтернативные задачи проверки гипотез. В этом случае одна из гипотез называется **основной гипотезой**, а другая **альтернативной**. При этом верна только одна из них.

   Вернемся к выяснению нормальности распределение. Ошибки могут быть 2 типов.

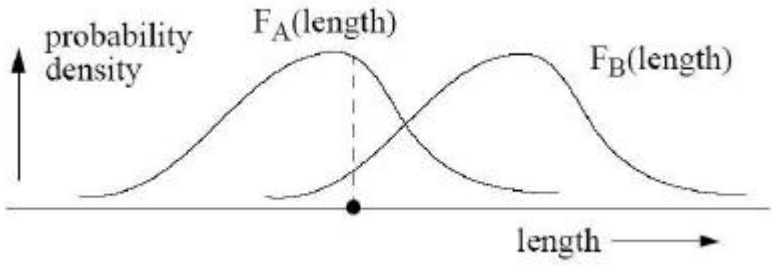
   (a) **Ошибка 1 рода** - отклонение основной гипотезы, в то время как она справедлива.

   (b) **Ошибка 2 рода** - принятие основной гипотезы, в то время как она неверна.

   Обычно вероятности ошибок 1 и 2 рода взаимосвязаны. То есть уменьшение вероятности ошибки 1 рода влечет за собой увеличение вероятности ошибки 2 рода, и наоборот, уменьшение вероятности ошибки 2 рода влечет за собой увеличение вероятности ошибки 1 рода. Необходимо найти компромис, делают это следующим образом: выбирают α такую, что вероятность ошибки 1 рода меньше α, α в таком случае называют **уровнем значимости**

2.3 Правило Байеса

Пусть даны два распределения A, B и значение X, порожденное одним из этих распределений. Наша задача определить каким из этих распределений было порождено значение X.



Интуиция нам подсказывает, что надо выбрать распределение A, если:

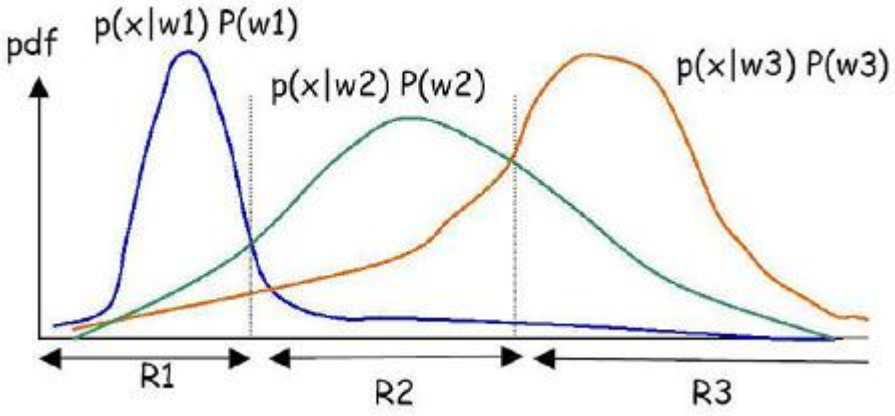
$$Prob(A|X) > Prob(B|X)$$

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

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

Надо выбирать A, если

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



2.4 Критерий χ^2

Постановка задачи
Дано:

1. Проверяемое распределение D
2. 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:

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

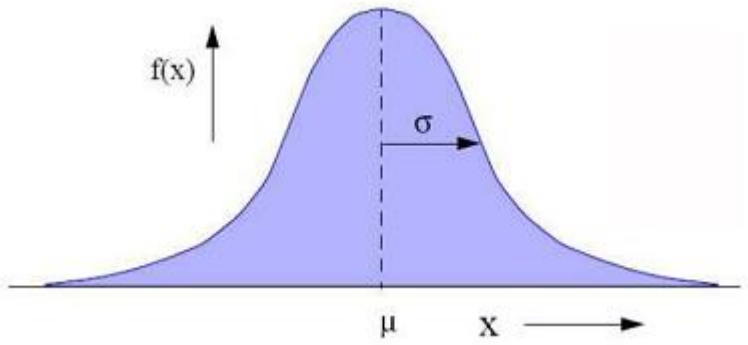
Критический уровень

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

Значение t определяется как функция от α и k

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

2.5 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

3 Статистические методы распознавания образов

3.1 Общая идеология

Постановка задачи

Дана тренировочная коллекция. Каждый объект представляется ввиде набора n характеристик = n-мерный вектор. Необходимо построить классифицирующее правило.

Предпосылки

Считаем, что элементы каждой категории имеют свое распределение в n-мерном пространстве. Будем принимать решение по правилу Байеса! Но для этого необходимо знать функции распределения каждой категории. Далее возможны три варианта: функции распределениея нам известны; нам известен тип распределения, но не параметры; нам ничего неизвестно о функциях распределния.

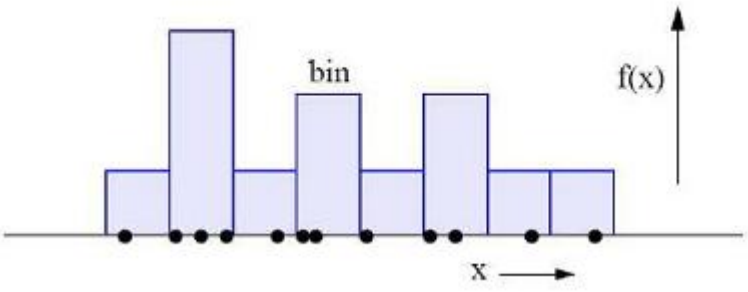
1. Функции распределения известны
   Просто используем правило Байеса.

2. Известен тип, но не параметры
   Будем использовать точечные оценки для параметров распределения.

   Пример. В случае нормального распределения будем использовать: $\hat{\mu} = \frac{X_1+\cdots+X_n}{n}$ $\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum(X_i - \hat{\mu})^2}$

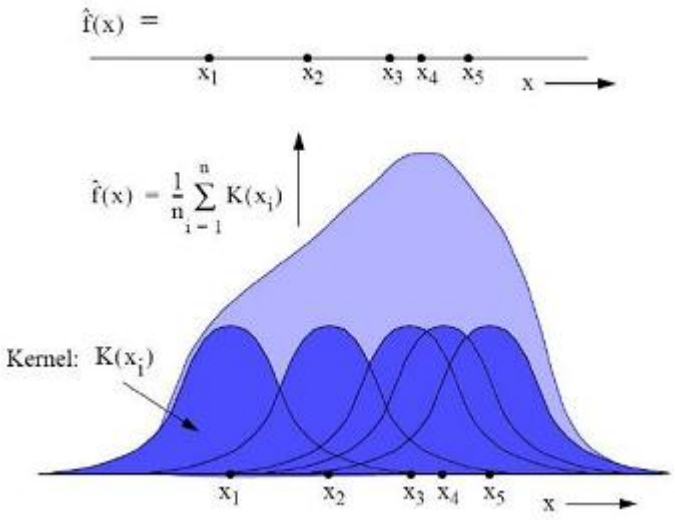
3. Неизвестное распределение
   Для того чтобы построить функцию распределения, нам придется воспользоваться тренировочной коллекцией. Рассмотрим два метода: метод гистограмм и метод Парзена.

   (a) Метод гистограмм
   Разобьем все n-мерное пространство на клетки. Каждой клетки определим плотность распределения как долю всех документов, попавших в клетку.

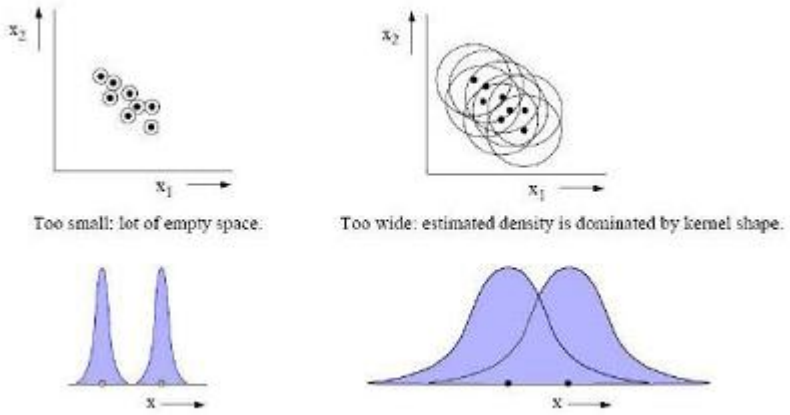


   Но у этого метода есть один недостаток. При большом количестве рассматриваемых характеристических свойств объектов, то есть когда n велико, необходима огромная тренировочная коллекция.

   (b) Метод Парзена
   Для каждой точки из класса построим функцию, достигающую максимума в этой точке и быстро убывающей при удалении от нее
   И в качестве функции распределения возьмем среднее арифметическое построенных функций.



Но что делать с размером "ядра"? И слишком маленькие ядра и слишком большие могут ухудшить качество распознавания.



Будем выбирать размер ядра для каждого объекта из тренировочной коллекции отдельно. Размер ядра будем рассчитывать таким образом, чтобы в него попало 5 соседей.

4 Задача

4.1 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?