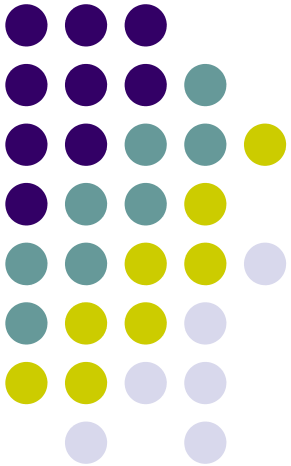
Информативность признаков

Корлякова М.О.

Калуга, 2019

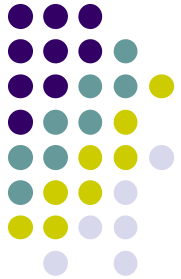


Литература

Хайкин С. Нейрокомпьютеры: полный курс. – М.:Вильямс – 2006

- Математические методы распознавания образов. Курс лекций. МГУ, ВМиК, кафедра «Математические методы прогнозирования» Местецкий Л.М., 2002–2004

План

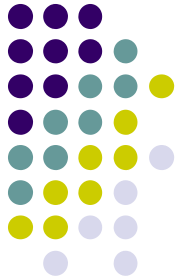


- Основные проблемы формирования информации для обучаемых систем
- Меры информативности независимой системы признаков. Информативность признака по Шеннону.
- Меры информативности независимой системы признаков. Дихотомия выборки. Геометрическая мера информативности.

План

- Выбор наиболее информативного набора признаков в зависимой системе характеристик. Проблемы и подходы к их решению.
- Критерии разделимости классов. Сепарабельность классов.
- Алгоритмы выбора информативного набора признаков ADD
- Алгоритмы выбора информативного набора признаков «случайный поиск с адаптацией»
- Алгоритмы выбора информативного набора признаков «таксономия признаков»

Обучение



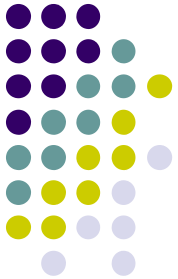
- С учителем (отнесение к классу)
- Без учителя (выделение классов)

- Обучение с подкреплением

- Supervised learning
- Unsupervised learning

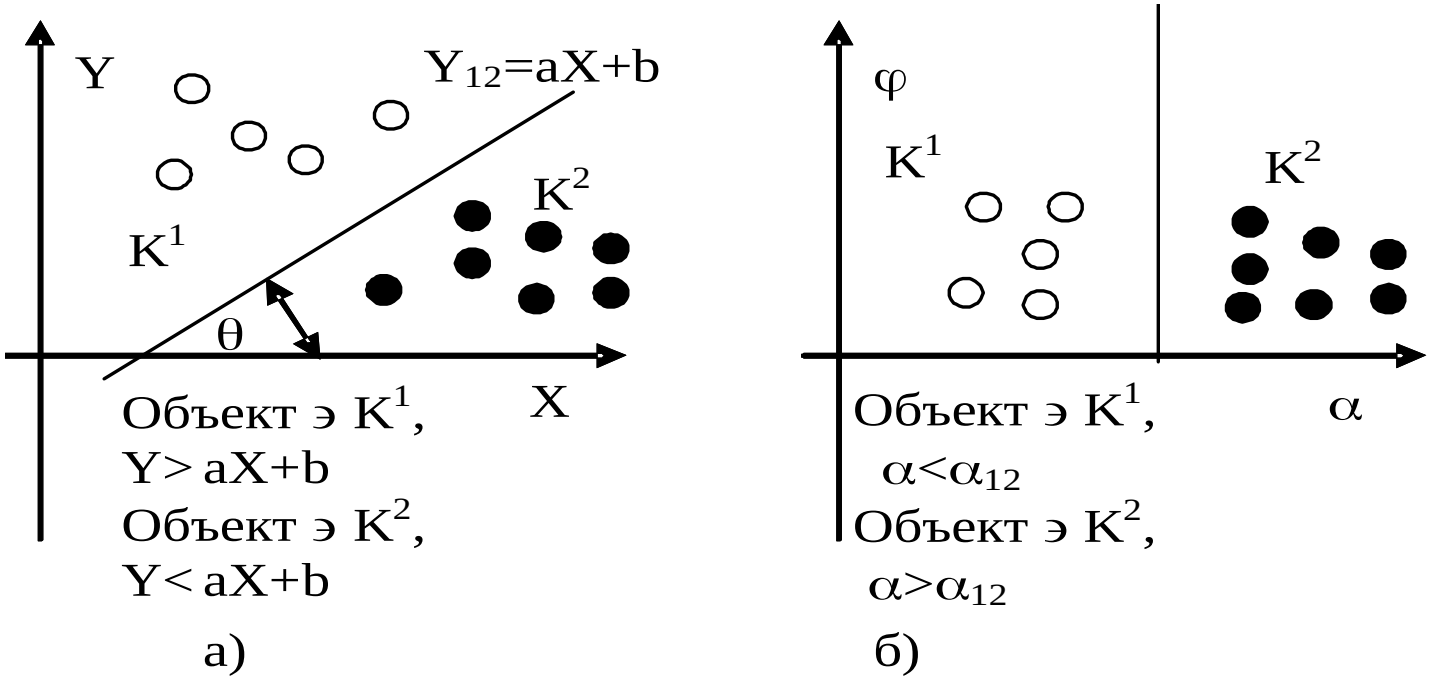
- Reinforcement learning

процедура представления информации



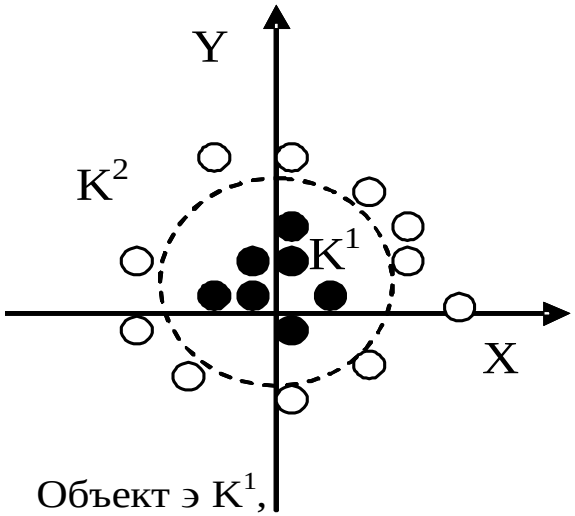
- Какие составляющие входной информации следует учитывать?
- Какой объем информации необходимо и достаточно сохранять для адекватной работы нейронной сети?
- Какие методы следует применять для решения вопросов информативности единиц данных?

Изменение координат



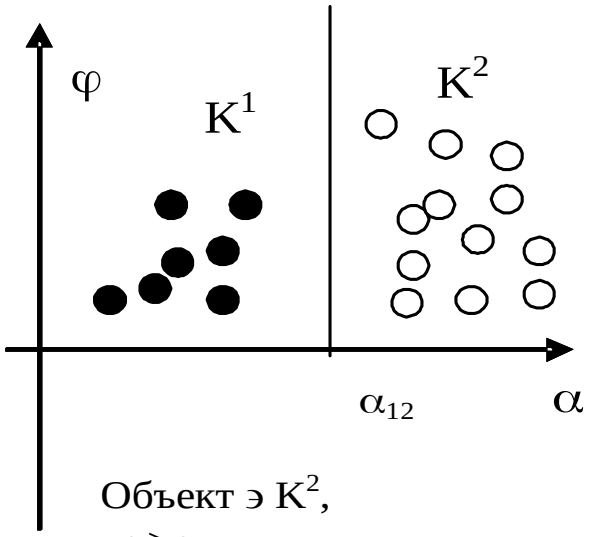
- **пространство объектов в исходной а) и развернутой на $\pi/2$- б) системе координат.**

Изменение координат



Объект э K^1,
$b2(Y-b1)^2+a2(X-a1)^2<C$
Объект э K^2,
$b2(Y-b1)^2+a2(X-a1)^2>C$

а)



Объект э K^2,
$\alpha>\alpha_{12}$
Объект э K^1,
$\alpha<\alpha_{12}$

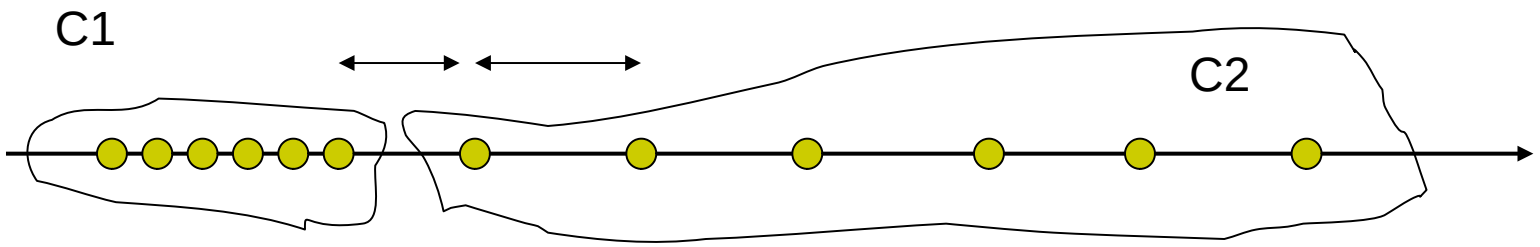
б)

- **пространство объектов в исходной а) и сферической б) системе координат.**

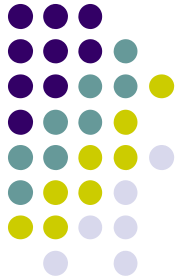
Гипотеза компактности

- *Классическая. Реализация одного и того же образа, обычно, отображается признаком пространства геометрически близкими точками.*
- Гипотеза -компактности

Расстояние мало, но есть неоднородность.



Рабочие утверждения

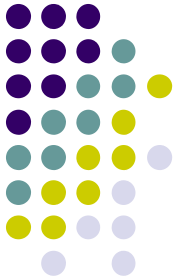


- Необработанное представление информации увеличивает ошибку обобщения нейронной сети и время на ее обучение.
- Состав и порядок представления объектов значительно влияет на результат обучения нейронной сети.

Проблема

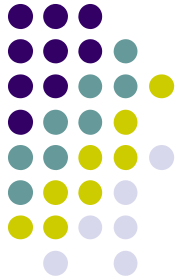
- **Необходимо отобрать интересные составляющие описания объекта – селекция :А КАК?**
- **Необходимо определить правильное преобразование описания объектов – выбор способа обработки : А КАКОЕ?**
- **Реализация дополнительного алгоритма преобразования описания объектов увеличивает время обработки данных : ВСЕ ПРОПАЛО?**

Этапы формирования системы распознавания



- Генерация признаков – выявление признаков, которые наиболее полно описывают объект.
- Селекция признаков – выявление признаков, которые имеют наилучшие классификационные свойства для конкретной задачи.
- Построение классификатора.
- Оценка классификатора.

Задача селекции признаков

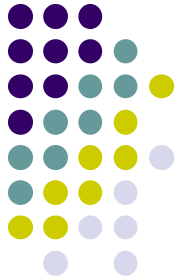


- ***Определение.*** *Процедура выделения из множества признаков меньшего подмножества с наилучшим сохранением информативности для классификации называется селекцией признаков.*

Задача селекции признаков

- $X \ R^m$ – множество признаков,
- $Y \ R^i$ – множество признаков, которые нужно отобрать в процессе селекции, причем
- $i < m$.
- Тогда задача селекции задается следующим образом: $X \ Y$.

Зачем тратим время на отбор признаков?



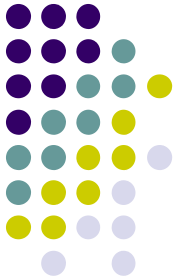
- Снижение сложности
- Повышение общности

Общность модели

- N – число прецедентов,
- k – число степеней свободы модели
- $\frac{N}{k}$ характеристика общности модели.

- Скалярная селекция (независимая система признаков)
- Векторная селекция (зависимые признаки)

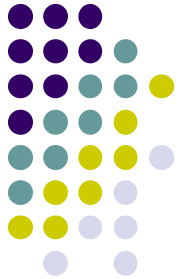
Построение информативных наборов признаков



- Решение лежит в поиске комбинаций удовлетворяющих гипотезе компактности.

- Для отбора из 20 исходных признаков пяти наиболее информативных приходится иметь дело примерно с $15{,}5\ 10^3$ вариантами.

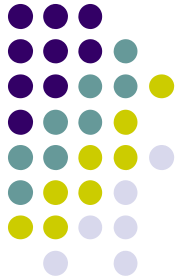
Типы селекции признаков



- Полный перебор

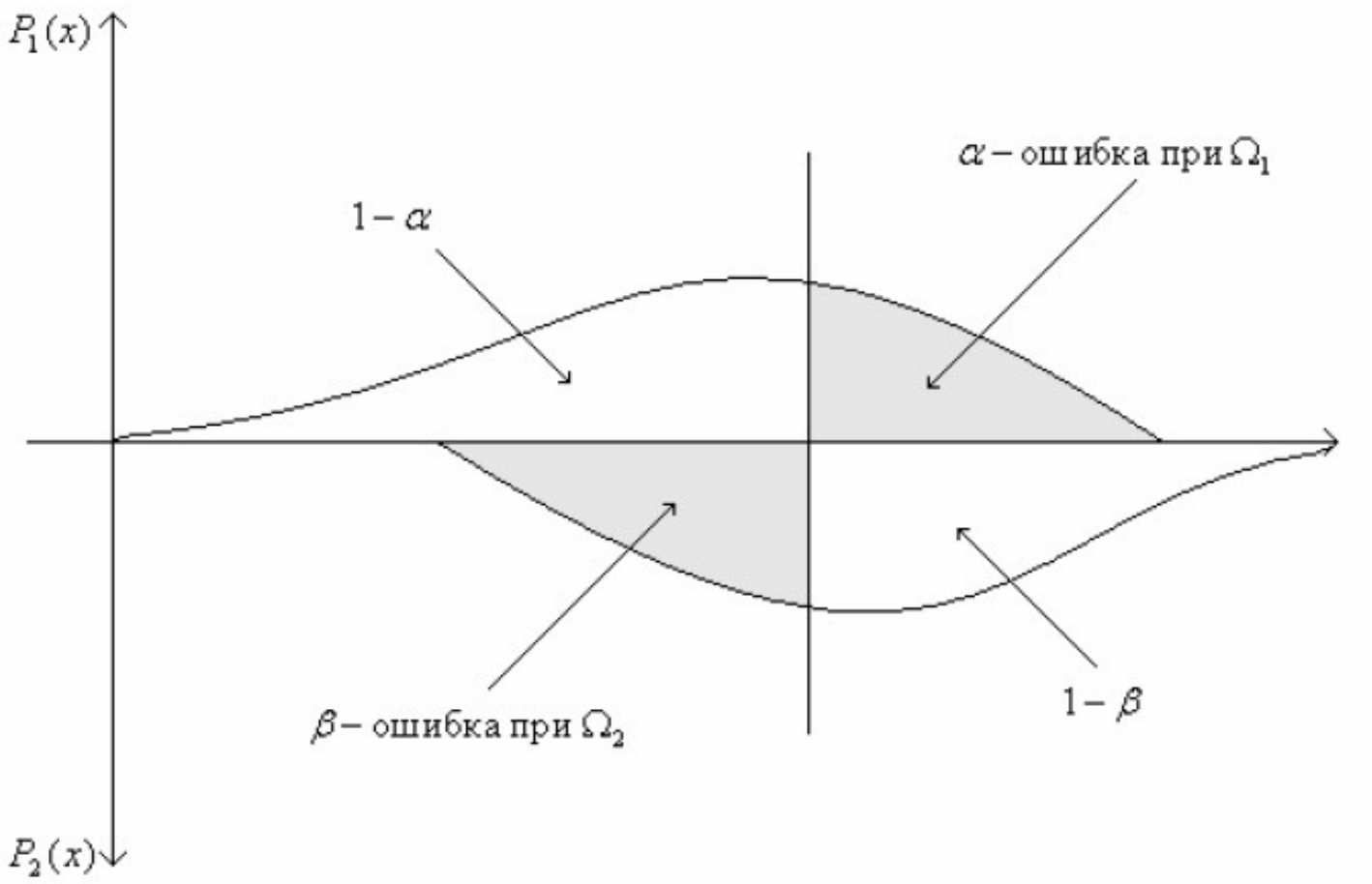
- Скалярная селекция
- Векторная селекция

Информативность независимых признаков

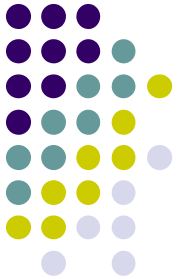


- Статистический анализ .
- Методы на основе вероятностной оценки информации о наборах данных .
- Алгоритмические методы вычисления информативности.

Распределение вероятности классификации по признаку x



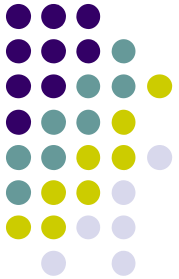
Статистический подход. Гипотезы



$X_1, \ldots, X_N$ – значение признака в первом классе со средним μ_1

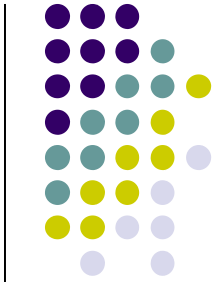
$Y_1, \ldots, Y_N$ - значение признака во втором классе со средним μ_2

Статистический подход. Гипотезы



- H_0– значения признаков отличаются существенно – нуль-гипотеза.
- H_1 - значения признаков отличаются несущественно – альтернативная гипотеза.

Статистический подход. Гипотезы

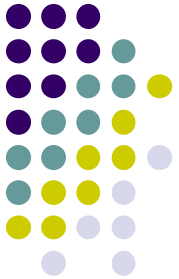


$$H_0 : \mu 1 - \mu 2 \neq 0$$

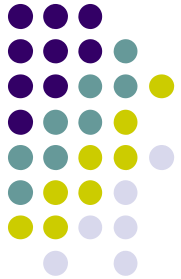
$$H_1 : \mu 1 - \mu 2 = 0$$

Методы определения

- Вычислить корреляцию



FisherIris



Обучение модели

	Количество признаков 70	
	Ошибка обобщ.%	Колич. эпох
Пирсона	28,2	58
Фехнера	25,1	53
Спирмена	30,3	55
Кендалла	22,3	57

Информативность по Шеннону

$$P(j) = \frac{1}{|V|} \sum_{j=1}^{V} P(i / j)$$

P(j) – вероятность различения образов по *j*-му признаку.

P(i|j) – вероятность различения образа *i* по *j*-му признаку.

V – число образов выборки.

Информативность по Шеннону

- Вклад образа *i* в информативность признака *j*.

$$r_i = \frac{P(i/\ j)}{P(j)}$$

Информативность по Шеннону

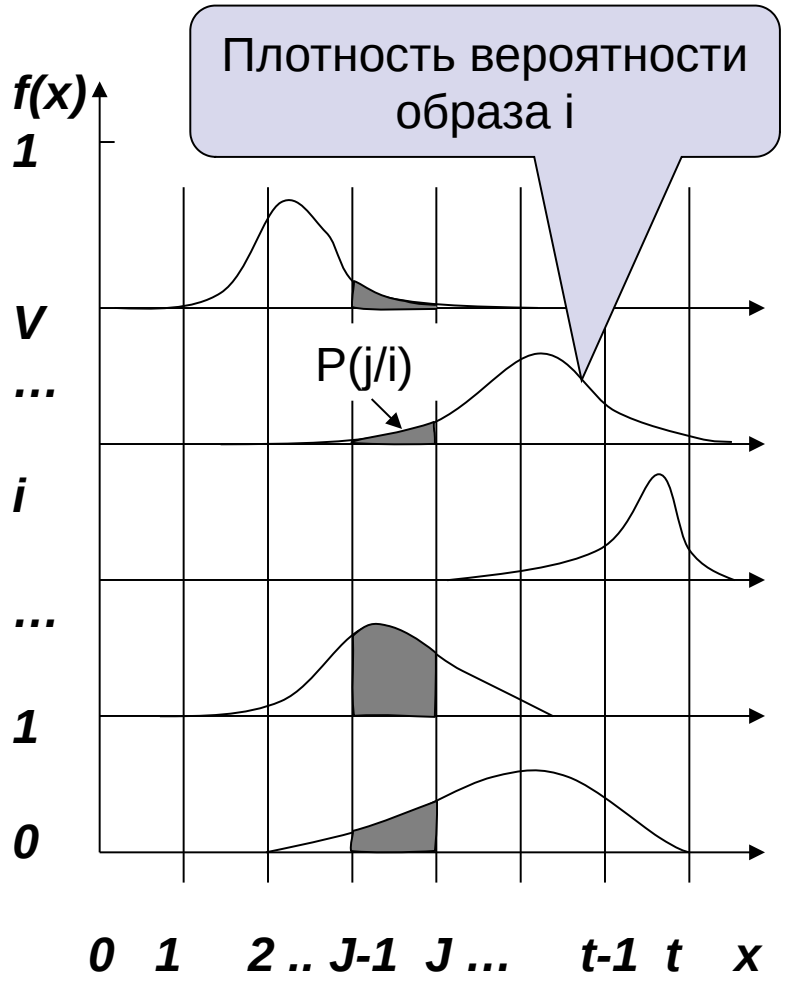
- По всему множеству образов

$$H_j = -(r_1 \log r_1 + r_2 \log r_2 + \cdot + r_i \log r_i + \cdot + r_k \log r_k)$$

Общая неопределенность по признаку x имеет вид

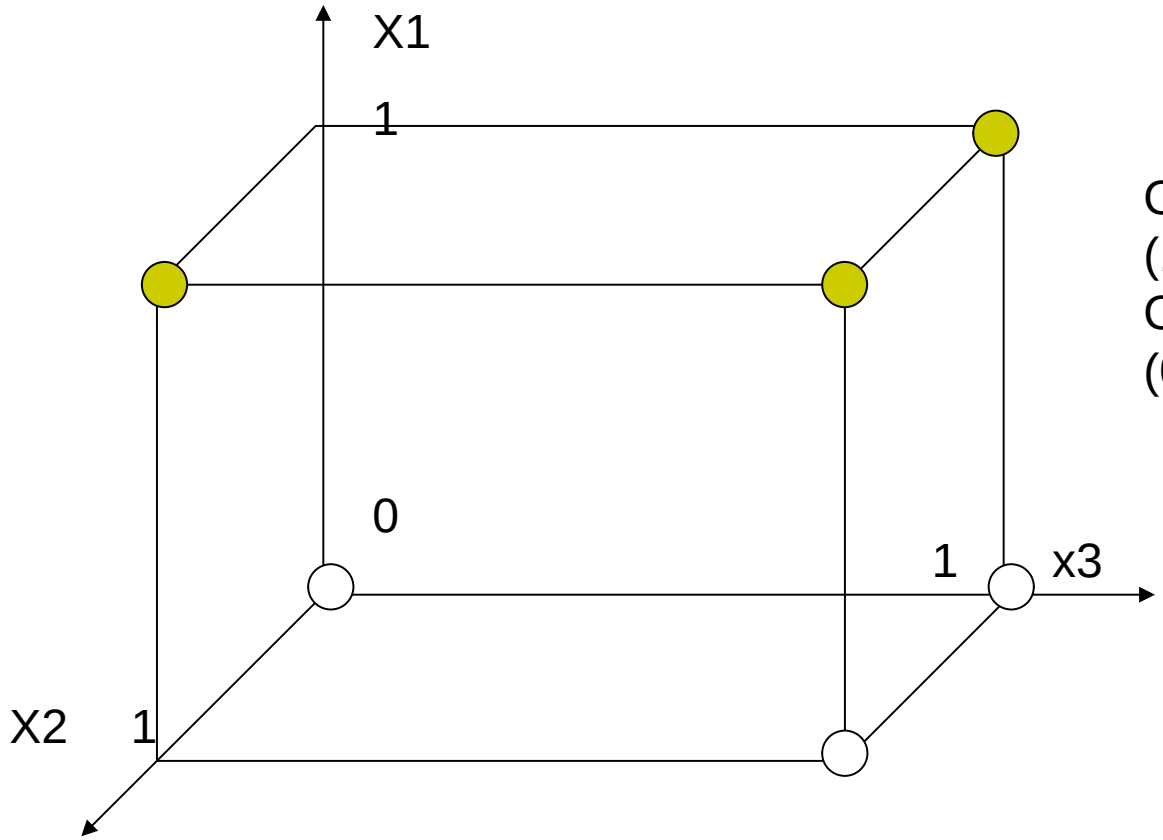
$$H_x = \sum_{j=1}^{t} H_j P_j$$

Информативность по Шеннону



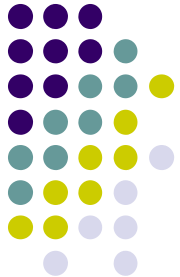
- H(j)=0 xj – абсолютно разделяющий.
- H(j)<H(m) xj информативней xm
- V – число образов
- t – число признаков

Информативность по Шеннону



Образ 1 ((1,1,1), (1,1,0),(1,0,1))
Образ 2 ((0,0,1), (0,1,1),(0,0,0))

Информативность по Шеннону



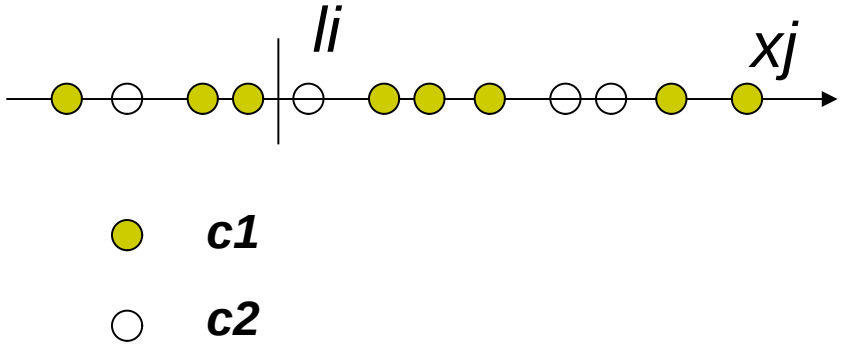
- Образ 1 ((1,1,1),(1,1,0),(1,0,1))
- Образ 2 ((0,0,1),(0,1,1),(0,0,0))
- Признак 2 для образа 1 P1(2/1)=2/3, P0(2/1)=1/3
- Признак 2 для образа 2 P1(2/2)=1/3, P0(2/2)=2/3
- V=2
- P(2)= (P1(2/1)+P1(2/2)+P0(2/1)+P0(2/2)) / 2 = 1;
- R11= P1(2/1)/P(2)=2/3, R10=1/3
- R20=2/3, R21=1/3 H=-(2* 2/3*log(2/3) + 2*1/3*log(1/3)) = 0.55

- Признак 1 для образа 1 P1(1/1)=3/3=1, P0(1/1)=0/3=0
- Признак 1 для образа 2 P1(1/2)=0/3=0, P0(1/2)=3/3=1
- V=2
- P(1)= (P1(1/1)+P1(1/2)+P0(1/1)+P0(1/2)) / 2 = 1;
- R11=1, r20=1, R10=r21=0 H=-(1*log1 + 0 + 1*log 1 +0) = 0
- Признак 1 разделяющий.

Дихотомия выборки по признаку

- Информативность *I*

$$I = \min_{l_i} R$$



$$l_i \Big| \qquad \begin{aligned} R_1 &= m_{1c1} \cdot m_{1c2} \\ R_2 &= m_{2c1} \cdot m_{2c2} \end{aligned} \qquad R = R_1 + R_2$$

Дихотомия выборки по признаку

l_1

$$R_1 = m_{1\,черн.} \cdot m_{1\,бел.} = 0 \cdot 3 = 0$$

$$R_2 = m_{2\,черн.} \cdot m_{2\,бел.} = 3 \cdot 1 = 3$$

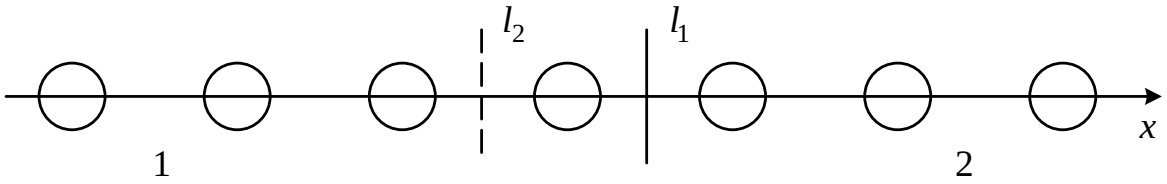
$$R = R_1 + R_2 = 3$$

l_2

$$R_1 = m_{1\,черн.} \cdot m_{1\,бел.} = 2 \cdot 1 = 2$$

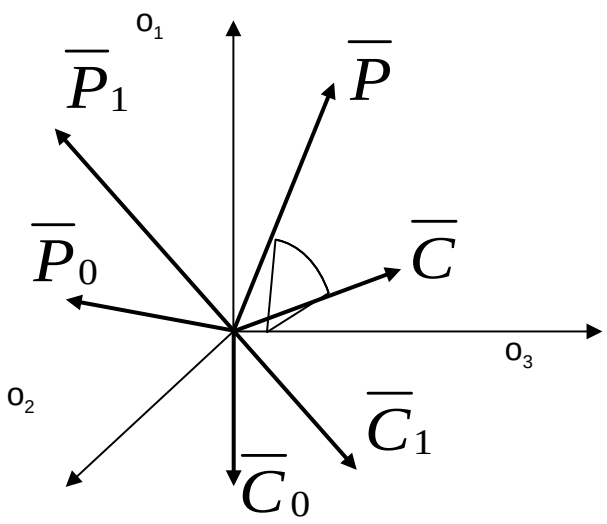
$$R_2 = m_{2\,черн.} \cdot m_{2\,бел.} = 3$$

$$R = R_1 + R_2 = 5$$



Геометрическая мера информативности

$$I(P) = \cos(P \wedge C) = \frac{\langle P \cdot C \rangle}{|P| \cdot |C|}$$



- Признаки похожи на цели

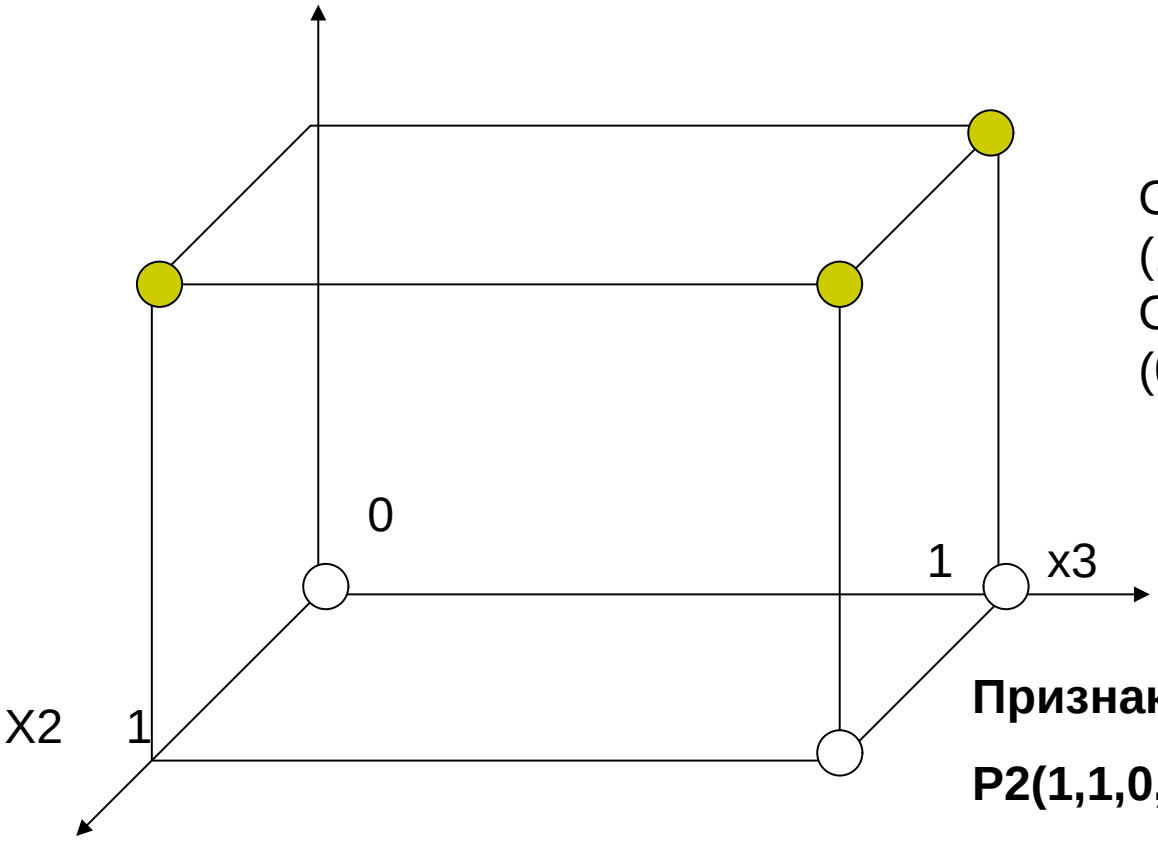
  $\cos(\overline{P}_1\overline{C}_1) = 1$

- Признаки не похожи на цели

  $\cos(\overline{P}_0\overline{C}_0) = 0$

- P –вектор признака, C – вектор целей, их длина равна числу объектов.

Геометрическая мера информативности



Образ 1 ((1,1,1), (1,1,0),(1,0,1))
Образ 2 ((0,0,1), (0,1,1),(0,0,0))

Признак 1 признак 2

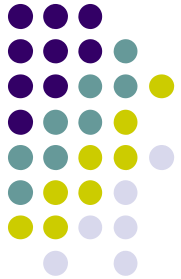
P2(1,1,0,0,1,0) P1(1,1,1,0,0,0)

C(1,1,1,2,2,2) C(1,1,1,2,2,2)

cos(P1 C)<1 cos(P1 C)=1

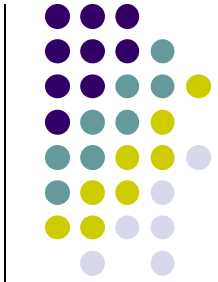
информативный

Критерий независимости признаков



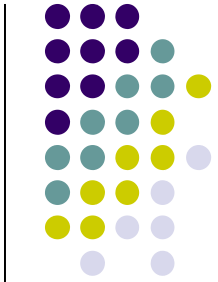
- Взаимная информация пары признаков с позиции любой из мер информативности.
- Если взаимная информация близка к 0, то признаки независимы.

Системы зависимых признаков



- Как отбирать?
- Как оценивать отобранное?

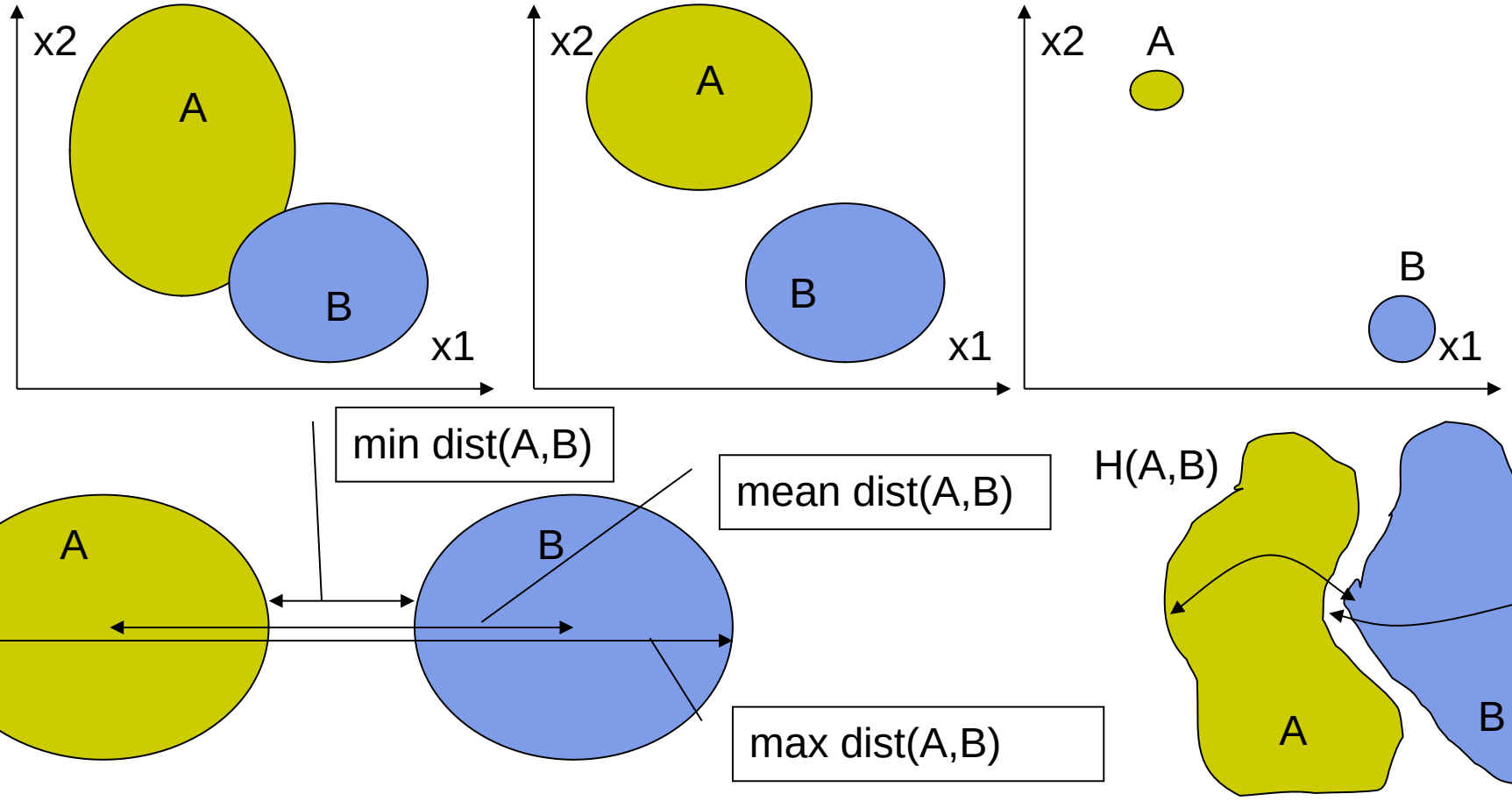
Оценка качества совокупности признаков



- Как оценить информативность признаков в группе?
  - Построить классификатор и сравнить ошибки
  - Ввести критерий и сравнить

Критерии качества системы признаков

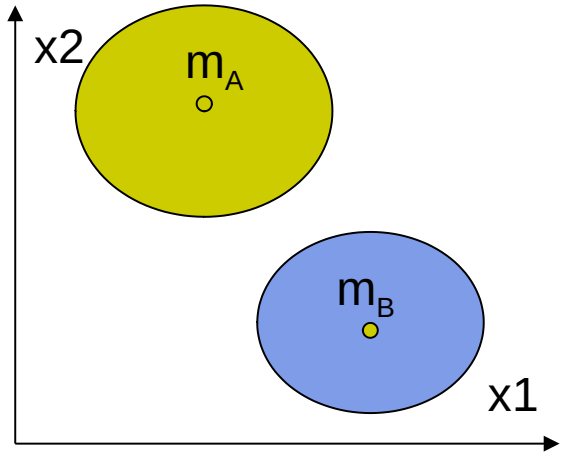
- Разделимость классов



Разделимость классов

- Сумма квадратов ошибок

- $Je = \sum_{i=1,V} \sum_{x \in V_i} ||x - m_i||^2$



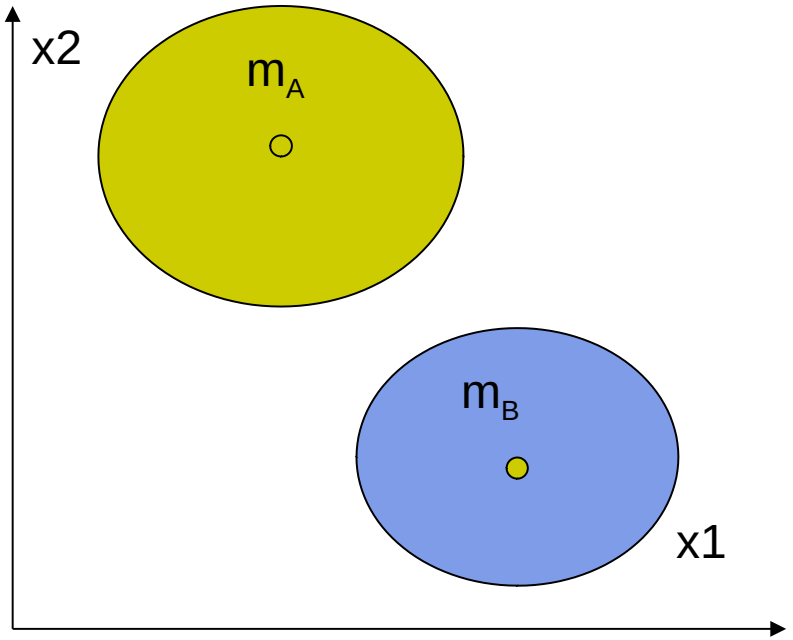
m_i – средний вектор образа i

Разделимость классов

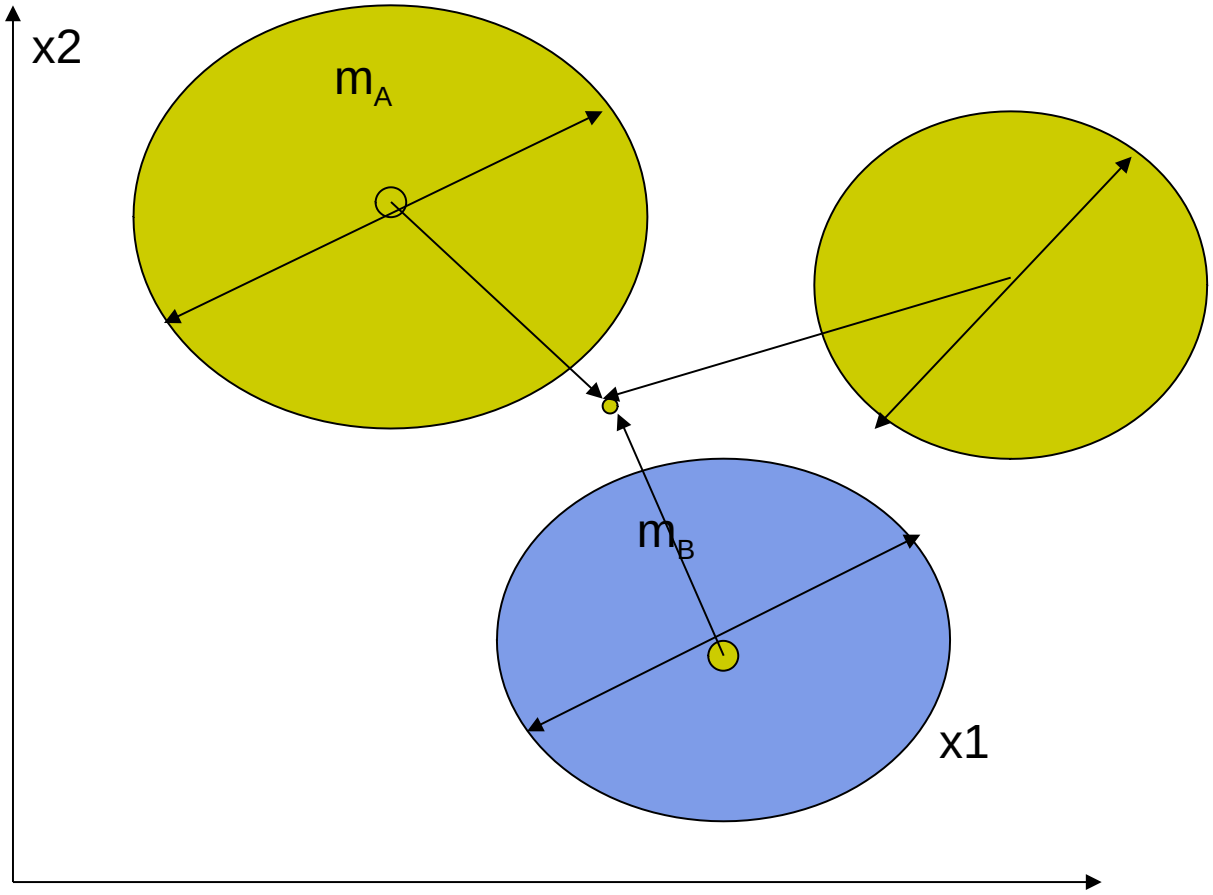
- Критерий минимума дисперсии

$$J_e = \frac{1}{2} \sum_{i=1}^{c} n_i \bar{s}_i,$$

$$\bar{s}_i = \frac{1}{n_i^2} \sum_{x \in V_i} \sum_{x' \in V_i} \| x - x' \|^2.$$



Критерий разделимости



Разделимость классов

- Критерии разделимости по рассеянию

$$J_1 = tr\left(SW^{-1}SB\right), J_4 = \frac{tr\left(SB\right)}{tr\left(SW\right)}$$

- матрица рассеяния внутри класса (SW) (расстояния между элементами одного класса)
- матрица рассеяния между классами (SB) (расстояния между элементами разных классов)

Разделимость классов

- Матрица рассеяния для i-й группы

$$S_i = \sum_{x' \in V_i} (x - m_i)(x - m_i)^t$$

- Матрица рассеяния внутри группы

$$SW = \sum_{j=1..|V|} P_j S_j$$

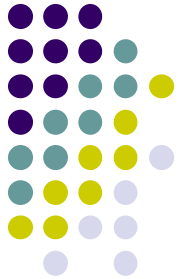
P_i – априорная вероятность класса

- Матрица рассеяния между группами

$$SB = \sum_{i=1..|V|} P_i (m_i - m)(m_i - m)^t$$

След матриц рассеяния измеряет квадрат радиуса рассеяния

Байесов классификатор



- ***Теорема.*** *Байесовский классификатор является оптимальным по отношению к минимизации вероятности ошибки классификации.*

Байесов классификатор

- $fx(X|C1)=1/(2*\pi* D_1^2) * exp(-1/(2*D_1^2)*||X-M_1||^2)$
- $fx(X|C2)=1/(2*\pi* D_2^2) * exp(-1/(2*D_2^2)*||X-M_2||^2),$
- где

fx(X|C1)- функция плотности условной вероятности для класса C1,

fx(X|C2)- функция плотности условной вероятности для класса C2,

D_1^2 – дисперсия класса C1, M_1 – вектор средних значений по всем признакам класса C1, ||.|| - оператор вычисления расстояния по Евклиду,

D_2^2 – дисперсия класса C2, M_2 – вектор средних значений по всем признакам класса C2.

Байесов классификатор

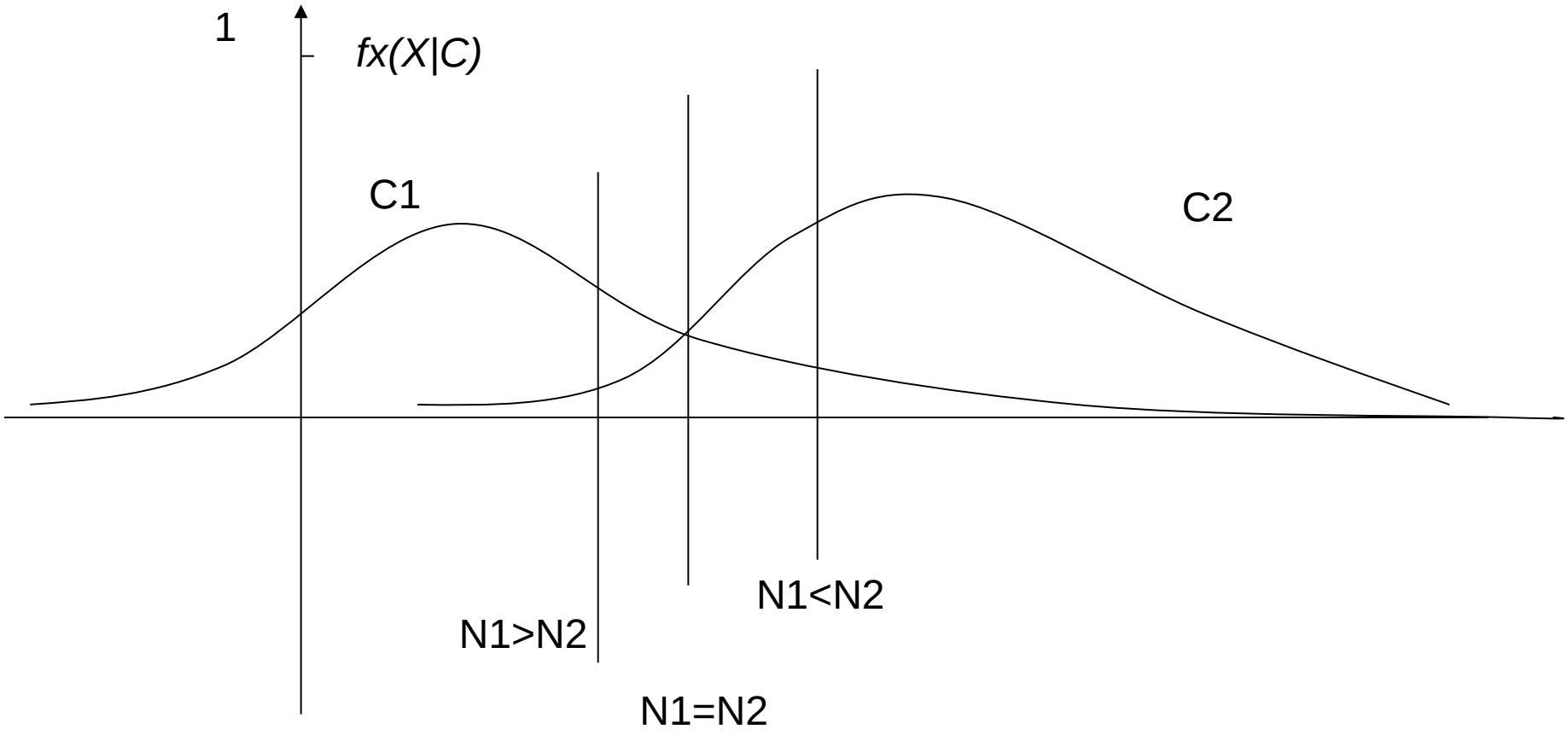
- отношение правдоподобия

(X)> θ для класса C1, где

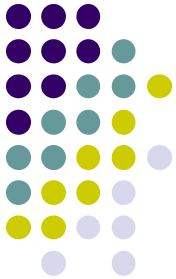
(X)= fx(X|C1)/fx(X|C2), θ*=p1/p2,*

pi – априорная вероятность класса Ci,

Байесов классификатор



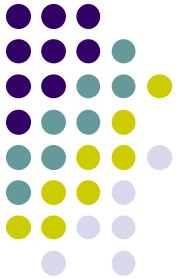
Байесов классификатор



- вероятность ошибки классификатора

- Pe = p1*P(e|C1) + p2*P(e|C2), где
- P(e|Ci) – условная вероятность ошибки для входного вектора класса i (установлена по фактическому отнесению примера к классу i байесовым классификатором),
- e - множество результатов некорректной классификации

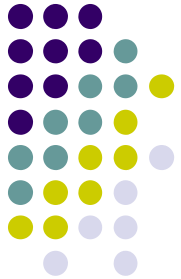
Построение информативных наборов признаков для зависимых характеристик



- Решение лежит в поиске комбинаций удовлетворяющих гипотезе компактности.

- Для отбора из 20 исходных признаков пяти наиболее информативных приходится иметь дело примерно с $15{,}5\ 10^3$ вариантами.

Жадные алгоритмы



Алгоритм ADD

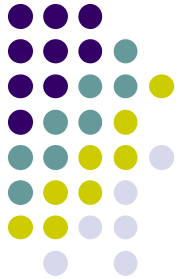
Цель: выделение k более или менее информативных признаков из всех доступных n.

- Оцениваем информативность каждого признака в отдельности,
- Ищем наиболее информативный признак среди независимо оцененных, и последовательно проверяем информативность пары из наиболее информативного признака и всех остальных.
- Наиболее информативная пара фиксируется. На этом этапе осуществляется n-1 проверка.

Алгоритм ADD

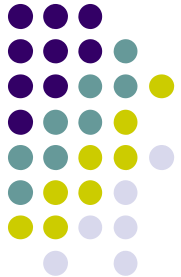
- Следующий шаг добавить к двум признакам третий и т.д.
- Процесс продолжаем до тех пор пока не получаем нужное количество признаков в множестве отобранных. На последнем шаге будет осуществлено n-k проверок.

Особенности ADD



- На последнем шаге сохраняется система признаков, которая дает наименьшую ошибку.
- Не обязательно получена идеальная система признаков.
- Однако, практически будет получено хорошее приближение к идеальному случаю
- Экономия времени.
- Для отбора 5 признаков из 20 при данном подходе требуется просмотреть 90 вариантов

Поиск



- Поиск В ширину
- Поиск в глубину

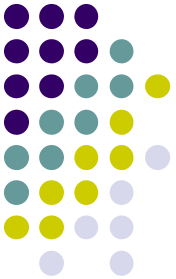
- Метод ветвей и границ
- МГУА

Алгоритм случайного поиска с адаптацией.

выбор приблизительно наилучшего подмножества признаков из доступных признаков

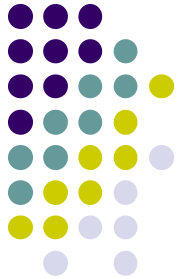
- Разбиваем отрезок [0, 1] на *g* частей. Каждая i-я часть отрезка сопоставлена с вероятностью p_i выбора i-го признака в состав информативного подмножества.
- Выбираем нужное число признаков (k) из всех возможных. Эта процедура осуществляется за счет размещения случайным образом на основе равномерного распределения точек вдоль отрезка [0, 1].

Алгоритм случайного поиска с адаптацией



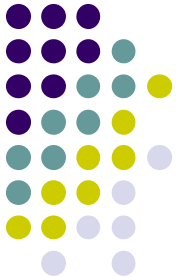
- Такую выборку повторяем *m* раз, a_j - ошибка обобщения при *j*-том опыте.
- Находим $min(\alpha_i)$
- Поощряем признак *i*, увеличивая соответствующий ему отрезок на фиксированное *h*
- Находим $max(\alpha_j)$
- Наказываем признак j на фиксированное *h*

Алгоритм случайного поиска с адаптацией



- Проводим следующую серию опытов, и снова поощряем лучший признак и наказываем худший. Повторяем эксперимент R раз.
- После R серий опытов длины некоторых отрезков сократились до нуля. Отрезки соответствующие информативным признакам увеличатся.

Алгоритм случайного поиска с адаптацией



- Адаптация состоит в изменении вектора вероятностей $\bar{p} = \{ p_1, p_2, ..., p_j, ..., p_n \}$

который сопоставлен с отрезком [0, 1].

- Выбор признаков на последующих этапах поиска осуществляется в зависимости от результатов предыдущих этапов.
- Скорость сходимости и качество решения зависят от *h.*

  Малое h – мягкая стратегия, большой перебор.

  Большое h – высокая скорость сходимости и грубое решение.

Таксономия признаков

- Производится таксономия множества признаков на k таксонов.
- Выбирается по одному типичному признаку из каждого таксона.
- Делается перебор C_k^n признаков, эти сочетания сравниваются по качеству распознавания, и выбирается такое сочетание, которое приводит к наименьшему числу ошибок.

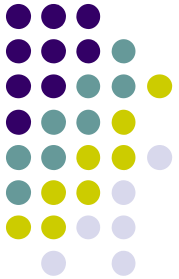
Генетические алгоритмы

- В качестве объекта G популяции можно рассматривать подмножество признаков, а в качестве функции полезности

$$h(G) = Er(G)^{-1}$$

(для $Er(G)$ <>0), где $Er(G)$ - ошибка распознавания тестового набора данных на основе подмножества признаков G.

Предобработка признаков



- Удаление выбросов
- Нормализация
- Заполнение пробелов

константа Липшица

- константа Липшица для выборки $\{x_i, y_i\}, i = 1, N$

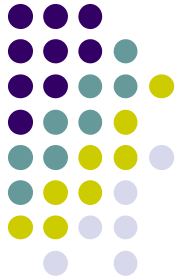
$$L = \max_{\substack{x_i \neq x_j \\ i \neq j}} \frac{\|y_i - y_j\|}{\|x_i - x_j\|}$$

где N-число объектов выборки,

x_i - значения входа, y_i - значение выхода

Построение набора признаков минимизирующих константу Липшица *часто* приводит к повышению качества обучения нейросети .

критерий оптимальности предобработки признака

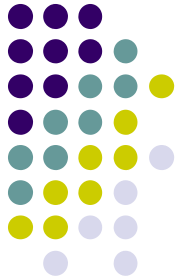


- Образ 1 ((1,1,1),(1,1,0),(1,0,1))
- Образ 2 ((0,0,1),(0,1,1),(0,0,0))
- Для признака 2

  x2(1,1,0,0,1,0)

  y(1,1,1,2,2,2)
- L=max (|y[i]-y[j]|)/|x2[i]-x2[j]+0.01|) = max {0,0,1,100, … } = 100
- |y[1]-y[5]|/|x2[1]-x2[5]+0.01| = |1-2|/|1-1+0.01| =100.

Переинтерпретация системы признаков



- провести линейное преобразование и определить систему из K новых признаков следующим образом

$$y_j = \sum_{k=1`}^{K} x_k a_{jk}$$

- $\left\| a_{jk} \right\|$ – диагональная матрица, причём её элементы равны либо 0, либо 1

Задача

№	X1	X2	X3	класс
1 (1)	1	2.1	5.6	+
2 (2)	1.9	2.3	4.5	+
3 (3)	2.4	1.9	4.8	+
4 (4)	2.1	3.0	4.9	+
5 (5)	1.3	2.4	4.7	+
6 (1)	3.4	4.1	5.1	-
7 (2)	3.6	5.0	6.2	-
8 (3)	4.1	4.6	6.0	-
9 (4)	3.9	4.7	5.6	-
10 (5)	4.3	4.8	5.7	-
11	3.7	4.7	5.5	-

- 1) определить по Байесу принадлежность X=(2.0, 3.5, 5.1)
- 2) определить по Байесу принадлежность X=(3.0, 4.5, 5.5)
- 3) найти информативный признак по вероятностной мере (каждый признак разбить на 4 интервала).
- 4) найти информативный признак по вероятностной мере (каждый признак разбить на 3 интервала).

Литература

- Саймон Хайкин. Нейронные сети полный курс. – М.: ООО «И.Д. Вильямс», 2006.
- Методы современной и классической теории управления. Т5. - 2004
- Математические методы распознавания образов. Курс лекций. МГУ, ВМиК, кафедра «Математические методы прогнозирования», Местецкий Л.М., 2002–2004.

Мера различимости