

# Информативность признаков

Сокращение числа признаков за счет удаления малоинформативных позволяет сократить время и повысить эффективность распознавания

## Критерий Фишера

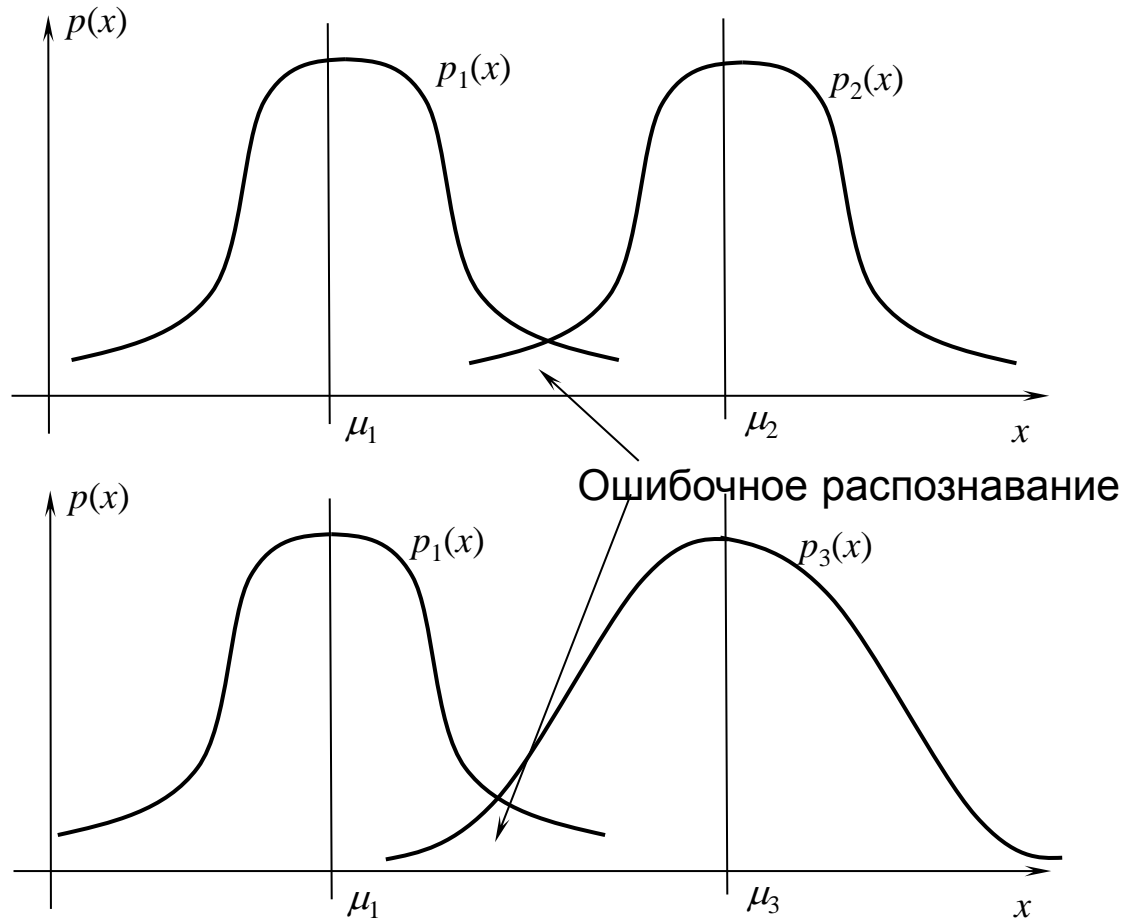
$$F^{(i)}(w_l, w_p) = \frac{(\mu_l - \mu_p)^2}{D_l + D_p}$$

где  $\mu_l$ ,  $\mu_p$  и  $D_l$ ,  $D_p$  – математические ожидания и дисперсии  $i$ -го признака для классов  $w_l$  и  $w_p$  соответственно. Определяет информативность  $i$ -го признака для пары классов  $w_l$  и  $w_p$ .

С увеличением критерия Фишера растет вероятность правильного распознавания по данному признаку. Информативность признака можно считать недостаточной, если критерий Фишера имеет слишком низкие значения для всех пар классов. Такой признак можно исключить.

**Замечание.** Возможно некоторый признак будет иметь низкие значения критерия Фишера для всех пар классов кроме одной пары, причем значения критерия Фишера для этой пары по всем другим признакам будут очень низкие. В таком случае данный признак **не следует исключать**, т.к. только он позволяет распознавать данную пару классов.

# Информативность признаков



**Сравнение** плотностей распределения вероятности значения  $p(x)$  по одному признаку для двух пар классов –  $w_1, w_2$  и  $w_1, w_3$  при одинаковых значениях математического ожидания ( $\mu_2 = \mu_3$ ) и разных дисперсиях ( $D_2 < D_3$ ). Вероятность правильного распознавания для пары классов  $w_1, w_2$  несколько больше чем для пары  $w_1, w_3$ . Критерии Фишера  $F(w_1, w_2) > F(w_1, w_3)$  для этих пар классов.

# Информативность признаков

## По энтропии признаков

**Энтропия множества событий**  $x = \{x_1, x_2, \dots, x_N\}$ , где  $x_i$  – событие, равна:

$$H(x) = - \sum_{i=1}^N p_i \log_2 p_i$$

где  $p_i$  – вероятность события  $x_i$ .

**Замечание.** Энтропию источника информации можно интерпретировать как среднюю неопределенность этого источника, тогда его информативность – величина устранимой источником энтропии.

**Информативность признака**  $x = \{x_1, x_2, \dots, x_N\}$  – средневзвешенное количество информации по всем значениям признака на множестве классов  $W = \{w_1, w_2, \dots, w_m\}$ , т.е.

$$I(x) = 1 + \sum_{i=1}^N \left( p(x_i) \sum_{j=1}^m p(w_j / x_i) \log_2 p(w_j / x_i) \right)$$

где  $p(w_j / x_i)$  – **условная** вероятность появления образа из класса  $w_j$  при значении признака  $x_i$ ,  $p(x_i)$  – вероятность появления значения  $x_i$  у признака  $x$  на всем множестве классов  $W$ ,  $N$  – число возможных значений признака  $x$ ,  $m$  – число классов

# Информативность признаков

Другой способ определения информативности признака  $x$  основан на суммировании уровней функциональной зависимости значения  $x_i$  и класса  $w_j$  по всем значениям признака на всем множестве классов с использованием  $p(w_j, x_i)$  – **совместной** вероятности появления образа из класса  $w_j$  со значением признака  $x_i$ , т.е.

$$I(x) = \sum_{i=1}^N \sum_{j=1}^m p(w_j, x_i) \log_2 \frac{p(w_j, x_i)}{p(x_i)p(w_j)}$$

где  $p(w_j)$  - вероятность появления образа из класса  $w_j$ .

Вероятности можно оценить по частоте соответствующих событий в обучающем множестве:

$$p(w_j / x_i) = \frac{S(w_j \cap x_i)}{S(x_i)}, p(w_j, x_i) = \frac{S(w_j \cap x_i)}{S}, p(x_i) = \frac{S(x_i)}{S}, p(w_j) = \frac{S(w_j)}{S},$$

где  $S(w_j \cap x_i)$  – число образов из класса  $w_j$  со значением признака  $x_i$ ,  $S(x_i)$  – число образов со значением признака  $x_i$  во всех классах,  $S(w_j)$  – число образов в классе  $w_j$ ,  $S$  – общее число образов в **обучающем множестве**.

**Замечание.** Оба подхода дают минимальное значение информативности признака, если практически отсутствует какая-либо связь между ним и распознаваемыми классами.

# Информативность признаков

## Пример

Определение информативности признаков  $\{x, y, z\}$ . Каждый может принимать значения  $\{1, 2, 3, 4\}$ , на множестве из 16-ти образов, относящихся к четырем классам  $W \{I, II, III, IV\}$

Эталон	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
x	1	1	1	1	2	2	2	2	3	3	3	3	4	4	4	4
y	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
z	1	1	2	4	2	2	3	1	3	3	4	2	4	4	1	3
W	I	I	I	I	II	II	II	II	III	III	III	III	IV	IV	IV	IV

Вероятности появления образов, относящихся к классам I, II, III и IV:

$$p(I) = p(II) = p(III) = p(IV) = 4/16 = 0.25.$$

У признака x значение 1 наблюдается на четырех эталонах  $\rightarrow p(x=1) = 4/16 = 0.25$ .

Аналогично,  $p(x=2) = p(x=3) = p(x=4) = 0.25$ .

Для y и z вероятности появления соответствующих значений совпадают с x.

### Для признака x

**Условная вероятность** появления образа из класса I при значении  $x = 1$ :

$$p(I/1) = 4/4 = 1,0. \text{ Аналогично, } p(II/2) = p(III/3) = p(IV/4) = 1,0$$

Остальные условные вероятности  $p(II/1) = p(III/1) = p(IV/1) = p(I/2) = p(III/2) = p(IV/2) = p(I/3) = p(II/3) = p(IV/3) = p(I/4) = p(II/4) = p(III/4) = 0/4 = 0$ .

**Информативность признака x по условной вероятности:  $I(x) = 1$**

# Информативность признаков

**Совместная вероятность** появления образа из класса I со значением  $x = 1$ :

$p(I,1) = 4/16 = 0,25$ . Аналогично  $p(II,2) = p(III,3) = p(IV,4) = 0,25$

Остальные совместные вероятности  $p(I,2) = p(I,3) = p(I,4) = p(II,1) = p(II,3) = p(II,4) = p(III,1) = p(III,2) = p(III,4) = p(IV,1) = p(IV,2) = p(IV,3) = 0/16 = 0$ .

**Информативность признака  $x$  по совместной вероятности:  $I(x) = 2$**

**Для признака  $y$**

**Условная вероятность** появления образа из класса I при значении признака  $y = 1$ :

$p(I/1) = 1/4 = 0,25$ . Аналогично,  $p(II/1) = p(III/1) = p(IV/1) = p(I/2) = p(II/2) = p(III/2) = p(IV/2) = p(I/3) = p(II/3) = p(III/3) = p(IV/3) = p(I/4) = p(II/4) = p(III/4) = p(IV/4) = 1/4 = 0,25$ .

**Информативность признака  $y$  по условной вероятности:  $I(y) = -1$**

**Совместная вероятность** появления образа из класса I, со значением признака  $y = 1$ :

$p(I,1) = 1/16 = 0,0625$ . Аналогично  $p(I,2) = p(I,3) = p(I,4) = p(II,1) = p(II,2) = p(II,3) = p(II,4) = p(III,1) = p(III,2) = p(III,3) = p(III,4) = p(IV,1) = p(IV,2) = p(IV,3) = p(IV,4) = 1/16 = 0,0625$ .

**Информативность признака  $y$  по совместной вероятности:  $I(y) = 0$**

**Для признака  $z$**

**Условная вероятность**  $p(I/1) = 2/4 = 0,5$ . Аналогично  $p(II/2) = p(III/3) = p(IV/4) = 0,5$ .

$p(I/2) = p(I/4) = p(II/1) = p(II/3) = p(I/3) = p(III/4) = p(III/2) = p(IV/1) = p(IV/3) = 1/4 = 0,25$ ;  $p(I/1) = p(II/4) = p(III/1) = p(IV/2) = 0/4 = 0$ .  **$I(z) = -0,5$ .**

**Совместная вероятность**  $p(I,1) = 2/16 = 0,125$ . Аналогично  $p(II,2) = p(III,3) = p(IV,4) = 0,125$ ,  $p(I,2) = p(I,4) = p(II,1) = p(II,3) = p(III,2) = p(III,4) = p(IV,1) = p(IV,3) = 1/16 = 0,0625$ ;  $p(I,3) = p(II,4) = p(III,1) = p(IV,2) = 0$ .  **$I(z) = 0,5$ .**

Соотношения информативности признаков совпадают

# Информативность бинарных признаков

Для бинарных признаков можно найти взаимную информацию каждого признака и каждого класса, потом выбрать из всех бинарных признаков наиболее информативные по всем классам (порог подбирается).

Взаимная информация  $I(x,y)$  описывает количество информации, содержащееся в одной случайной величине –  $x$  относительно другой –  $y$ .  $I(x,y) = H(x) - H(x|y)$

Признак  $x$  может принимать значения 0 или 1. Известно:  $N_{11}$  – число образов с  $x=1$  в кл.1,  $N_{01}$  – число образов с  $x=0$  в классе 1,  $N_{10}$  – число образов с  $x=1$  во всех других классах,  $N_{00}$  – число образов с  $x=0$  во всех других классах.

Общее число образов во всех классах  $N = N_{00} + N_{01} + N_{10} + N_{11}$ .

**Информативность признака  $x$  для класса 1:**

$$I(x,1) = p_{11} \log_2(p_{11} / p_1) + p_{01} \log_2(p_{01} / p_0) + p_{10} \log_2(p_{10} / p_1) + p_{00} \log_2(p_{00} / p_0)$$

$p_{11}=N_{11}/N$ ,  $p_{1|1}=N_{11}/(N_{01}+N_{11})$  – совместные и условные вероятности  $x=1$  и кл.1;

$p_{01}=N_{01}/N$ ,  $p_{0|1}=N_{01}/(N_{01}+N_{11})$  – совместные и условные вероятности  $x=0$  и кл.1;

$p_{10}=N_{10}/N$ ,  $p_{1|0}=N_{10}/(N_{00}+N_{10})$  – совместные и условные вероятности  $x=1$  и др.классов;

$p_{00}=N_{00}/N$ ,  $p_{0|0}=N_{00}/(N_{00}+N_{10})$  – совместные и условные вероятности  $x=0$  и др.классов;

$p_1=(N_{11}+N_{10})/N$ ,  $p_0=(N_{00}+N_{01})/N$  – вероятности  $x=1$  и  $x=0$  по всем классам.

# Информативность бинарных признаков

**Пример.** Информативность бинарных признаков  $x$  и  $y$  для класса 1.

Для признака  $x$  известно:  $N_{11} = 80$ ,  $N_{01} = 20$ ,  $N_{10} = 300$ ,  $N_{00} = 600$ .  $N = 1000$ .

Тогда:  $p_{11}=0,08$ ;  $p_{1|1}=0,8$ ;  $p_{01}=0,02$ ;  $p_{0|1}=0,2$ ;  $p_{10}=0,3$ ;  $p_{1|0}=3/9=0,333$ ;  $p_{00}=0,6$ ;  $p_{0|0}=6/9=0,667$ ;  
 $p_1 = 0,38$ ;  $p_0=0,62$ .

$$I(x,1) = 0,08 \cdot \log_2(0,8/0,38) + 0,02 \cdot \log_2(0,2/0,62) + 0,3 \cdot \log_2(0,333/0,38) + 0,6 \cdot \log_2(0,667/0,62) =$$

**0,0594**

Для признака  $y$  известно:  $N_{11} = 80$ ,  $N_{01} = 20$ ,  $N_{10} = 400$ ,  $N_{00} = 500$ .

Тогда:  $p_{11}=0,08$ ;  $p_{1|1}=0,8$ ;  $p_{01}=0,02$ ;  $p_{0|1}=0,2$ ;  $p_{10}=0,4$ ;  $p_{1|0}=4/9=0,444$ ;  $p_{00}=0,5$ ;  $p_{0|0}=5/9=0,556$ ;  
 $p_1 = 0,48$ ;  $p_0=0,52$ .

$$I(y,1) = 0,08 \cdot \log_2(0,8/0,48) + 0,02 \cdot \log_2(0,2/0,52) + 0,4 \cdot \log_2(0,444/0,48) + 0,5 \cdot \log_2(0,556/0,52) =$$

**0,0347**

Признаки  $x$  и  $y$  встречаются в образах класса 1 одинаковое число раз (соответственно, **не** встречаются тоже одинаковое число раз). Признак  $y$  встречается в образах всех других классов чаще чем признак  $x$ . Для класса 1 информативность признака  $x$  выше, чем  $y$ .

**Замечание.** Пример распознавания образов с большим числом бинарных признаков – классификация текстов по содержащимся в них словам. Признак – наличие слова, число признаков – число уникальных слов во всех текстах обучающего множества. Как правило, таких слов очень много, причем среди них большинство малоинформативные.



# Построение классификаторов

Два подхода:

1. Классификация по расстоянию в пространстве признаков
2. Классификация разбиением пространства признаков на области

## **Классификация по расстоянию в пространстве признаков**

Выбрать способ определения расстояния между точкой, соответствующей распознаваемому образу в пространстве признаков, и множеством точек, соответствующих одному из классов.

После определения расстояний от распознаваемого образа до всех классов образ относится к тому из них, расстояние до которого минимально. Если это расстояние превышает некоторый порог, то образ классифицируется как нераспознанный

**Замечание.** Подход дает хорошие результаты для классов, имеющих компактную эллипсоидную форму в пространстве признаков.

Требует значительных расходов времени на определения расстояний при большом числе классов, признаков или расстоянии – до ближайшего представителя класса.

Сложно учитывать априорную вероятность появления образов для каждого класса и стоимость ошибки распознавания.

# Построение классификаторов

## Классификация по K ближайшим соседям

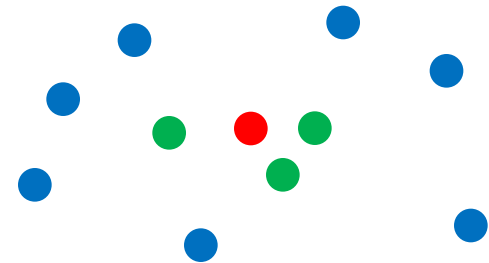
Метод работает со всеми образами из обучающего множества. Выбирается  $K$  – число ближайших соседей и метод вычисления расстояния между точками в пространстве признаков. Находятся расстояния от распознаваемого образа до всех образов обучающего множества, выбирается  $K$  ближайших и проводится их анализ.

### Возможные метода анализа:

1. Распознаваемый образ относится к классу с наибольшим числом соседей
  2. Ближайшие соседи из  $K$  получают больший вес. Каждому  $i$ -му образу присваивается вес, как  $1/d_i$ , где  $d_i$  – расстояние от  $i$ -го образа до распознаваемого. Образ относят к классу с максимальным суммарным весом
- ... Известны и другие, более сложные методы.

### Пример

$K=10$



По 1-му методу образ ● будет отнесен к классу ●

По 2-му – к классу ●

**Замечание.** Метод очень прост в реализации, его следует применять при значительном пересечении в пространстве признаков классов из образов обучающего множества.

**Недостаток** - при большом объеме обучающего множества может потребоваться значительное время на обработку и большой объем памяти для хранения обучающего множества. Результат во многом зависит от выбора  $K$  и метода вычисления расстояния

# Построение классификаторов

## Разбиение пространства признаков на области

Разбиение пространства признаков на области, соответствующие классам  $(D_1, D_2, \dots, D_m)$  должно обеспечить минимальную вероятность ошибки распознавания.

Границы областей описываются функциями  $S_g(\mathbf{x})$ , причем если образ, описываемый вектором признаков  $\mathbf{x}$ , фактически относится к  $g$ -му классу, то  $S_g(\mathbf{x}) > S_j(\mathbf{x})$  для всех  $j \neq g$ . Отсюда, граница между областями  $D_g$  и  $D_k$ , соответствующих  $g$ -му и  $k$ -му классам:  $S_g(\mathbf{x}) - S_k(\mathbf{x}) = 0$ .

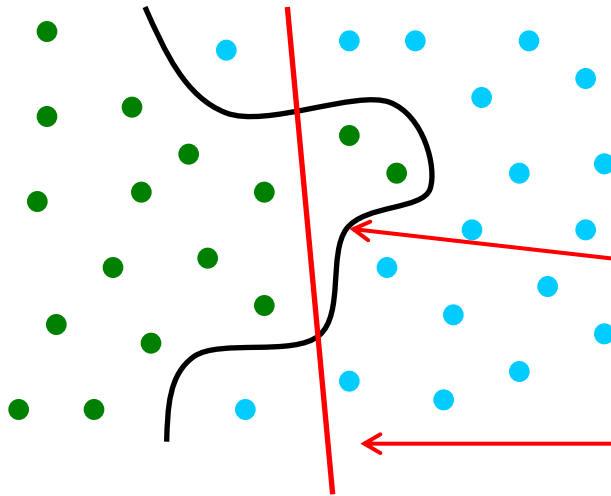
**Замечание.** Форма границы зависит от вида разделяющей функции (полиномиальной или линейной). В последнем случае граница – гиперплоскость в многомерном пространстве признаков или линия в двумерном.

Если классы не пересекаются в пространстве признаков, то границы могут быть проведены так, что распознавание, по крайней мере, образов обучающего множества будет выполняться без ошибок. В случае пересекающихся классов от системы распознавания можно добиться только минимизации среднего риска при ошибочном распознавании.

Выяснение вопроса – пересекаются ли классы в пространстве признаков трудная задача. Сложный вопрос и определение **вида** разделяющей функции. Пример – представление границ каждого класса как совокупность гиперсфер вокруг образов данного класса. Фактически, получаем метод распознавания до ближайшего представителя класса.

# Построение классификаторов

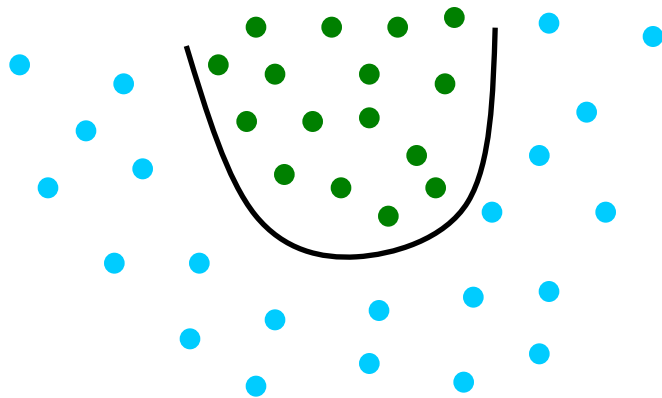
## Проблема «истинной» границы между классами



Можно считать, что два класса пересекаются, т.к. крайне сложно провести разделяющую линию, описываемую полиномом.

Линия может считаться границей, но ее «истинность» вызывает сомнения.

Прямая линия, разделяющая классы с минимизацией ошибки распознавания.

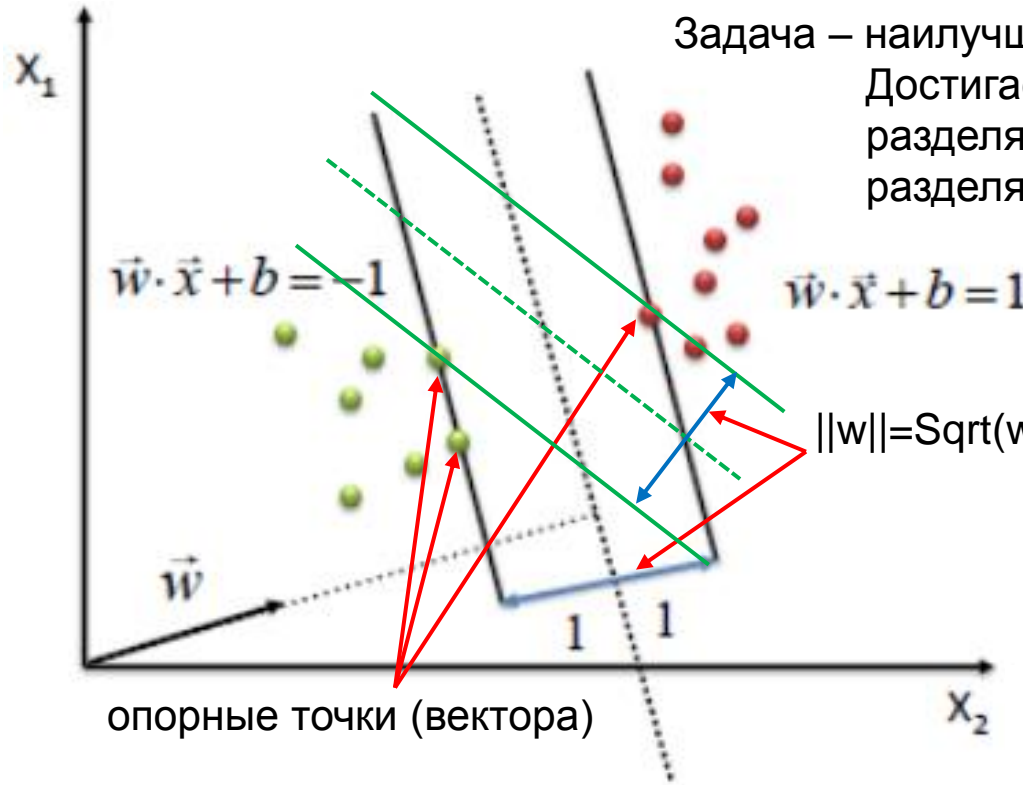


Два класса неразделимы прямой, но кривая, описываемая уравнением параболы, вполне может быть принята за границу между ними.

**Замечание.** В многомерном ( $n > 3$ ) пространстве признаков возможно только аналитическое представление разделяющих поверхностей и определить их «истинность» довольно сложно.

# Метод опорных векторов

## Машина опорных векторов - Support Vector Machine



Задача – наилучшее линейное разделение двух классов  
 Достигается при максимальной ширине полосы, разделяющей классы. Для плоскости уравнение разделяющей линии:  $w_1x_1 + w_2x_2 + b = 0$

$$\max \frac{2}{\|w\|}$$

$$\|w\| = \sqrt{w_1^2 + w_2^2}$$

при условии (x – образ)

$$(w \cdot x + b) \geq 1, \forall x \text{ of class 1}$$

$$(w \cdot x + b) \leq -1, \forall x \text{ of class 2}$$

При отсутствии линейной разделимости возможна **линеаризация пространства признаков** с использованием спец. ядра

**Достоинства и недостатки:**  $w \cdot \bar{x} + b = 0$

- наиболее быстрый метод нахождения разделяющей линии сводится к решению задачи квадратичного программирования в выпуклой области, всегда имеет одно решение
- метод чувствителен к шумам (выбросам образов в обучающем множестве X)
- нет общего подхода к автоматическому выбору ядра (и построению спрямляющего подпространства в целом) в случае линейной неразделимости классов

# Построение классификаторов

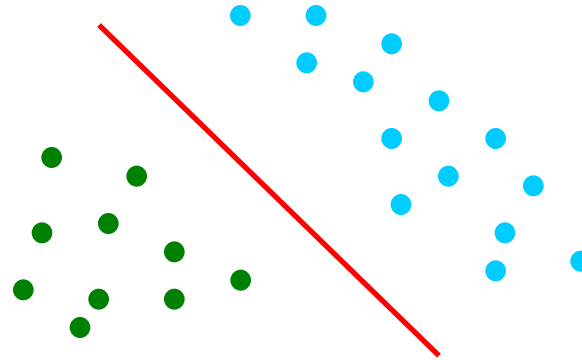
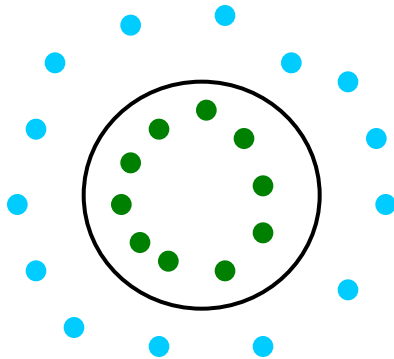
## Линеаризация разделяющей функции

Нелинейную границу между классами в пространстве признаков можно линеаризовать, вводя дополнительные признаки.

**Пример1.** Граница между классами на плоскости описывается уравнением окружности:

$$x^2 + y^2 = R$$

Введем новые признаки  $x' = x^2$  и  $y' = y^2$ . Образы в измененном пространстве признаков могут быть разделены прямой линией, описываемой уравнением  $x' + y' = R$ .



**Пример 2.** Граница между классами на плоскости в двумерном пространстве признаков описывается полиномом:  $ax_1^2 + bx_2^2 + cx_1x_2 + dx_1 + ex_2 + f = 0$ . Введем еще три признака:  $x_3 = x_1^2$ ,  $x_4 = x_2^2$ ,  $x_5 = x_1x_2$ . Уравнение границы примет вид:  $ax_3 + bx_4 + cx_5 + dx_1 + ex_2 + f = 0$ . Тогда парабола в 2-х мерном пространстве признаков станет в 5-ти мерном пространстве гиперплоскостью

# Построение классификаторов

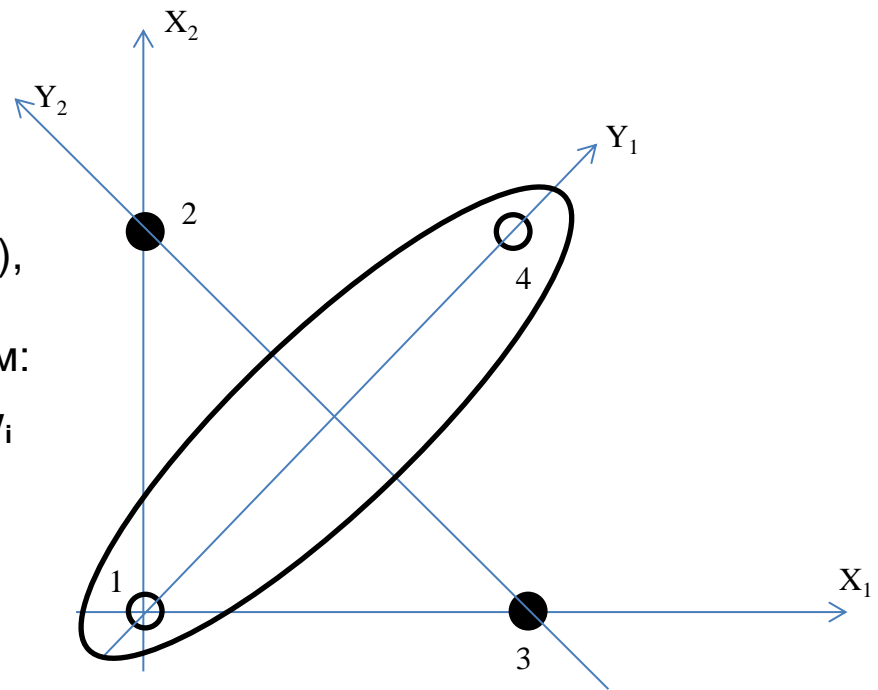
## Линеаризация разделяющей функции

**Пример3.** Представленные в таблице образы линейно неразделимы (результат логической операции «сложение по модулю два» или «исключающее или»)

№ образа	Признак $x_1$	Признак $x_2$	Класс $w_i$
1	0	0	0
2	0	1	1
3	1	0	1
4	1	1	0

Сместим начало координат в точку  $(0.5, 0.5)$ , повернем новые оси  $y_1, y_2$  на 45 градусов и выполним масштабирование по новым осям:

№ образа	Признак $y_1$	Признак $y_2$	Класс $w_i$
1	-1	0	0
2	0	1	1
3	0	-1	1
4	1	1	0



Разделяющая классы граница – эллипс:  $dy_1^2 + ey_2^2 + f = 0$ .

Замена переменных  $z_1 = y_1^2$ ,  $z_2 = y_2^2$  приводит линейной границе:  $dz_1 + ez_2 + f = 0$ .

Практически любую разделяющую поверхность в пространстве признаков можно привести к линейному виду заменой переменных и введением дополнительных признаков.

# Построение классификаторов

## Минимизация ошибки классификации

**Байесовская** стратегия позволяет минимизировать стоимости потерь

Пусть образы разделяются на два класса  $w_1$  и  $w_2$  по одному вероятностному признаку  $x$ . Вектор признаков состоит из одного элемента, а описания классов представляют собой функции условной плотности распределения вероятности значений признака  $p_1(x)$ ,  $p_2(x)$ . Пусть известны априорные вероятности появления образов разных классов  $P_1$ ,  $P_2$  и

матрица стоимости потерь  $\mathbf{C} = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix}$ , где  $c_{11}$ ,  $c_{22}$  – потери при правильных решениях,

$c_{12}$  и  $c_{21}$  – потери при ошибочной классификации (образ, относящийся к первому классу, ошибочно распознан как образ, относящийся ко второму классу, и наоборот).

Первый случай называется **ошибкой первого рода**, а второй – **ошибкой второго рода**.

Условная вероятность ошибки первого рода  $Q_1 = \int_{x_0}^{\infty} p_1(x) dx$

Условная вероятность ошибки второго рода  $Q_2 = \int_{-\infty}^{x_0} p_2(x) dx$

где  $x_0$  – пороговое значение признака.

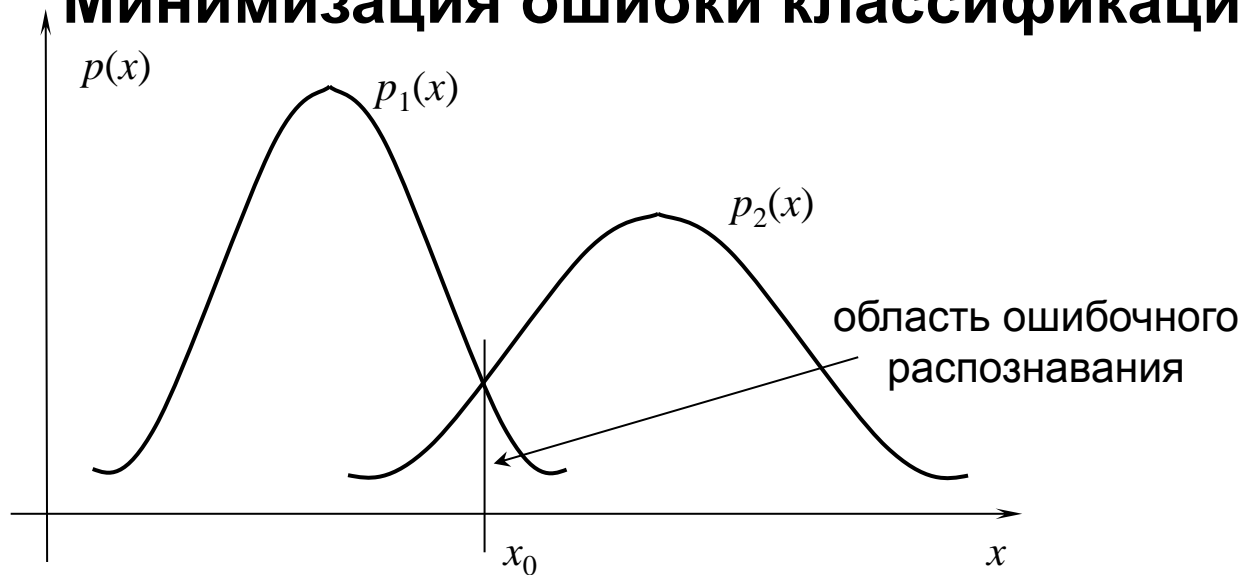
Образ распознается как относящийся к первому классу, если для него  $x < x_0$ ,

Образ распознается как относящийся ко второму классу, если для него  $x > x_0$ .



# Построение классификаторов

## Минимизация ошибки классификации



Потери при распознавании:  $R = P_1 c_{11}(1 - Q_1) + P_1 c_{12} Q_1 + P_2 c_{22}(1 - Q_2) + P_2 c_{21} Q_2$

Найдем такое  $x_0$ , при котором  $R$  минимально  $\left| \frac{dR}{dx} \right|_{x=x_0} = P_1 [c_{11} p_1(x_0) - c_{12} p_1(x_0)] + P_2 [c_{21} p_2(x_0) - c_{22} p_2(x_0)] = 0$

Откуда  $\frac{p_2(x_0)}{p_1(x_0)} = \frac{P_1(c_{12} - c_{11})}{P_2(c_{21} - c_{22})} = \lambda_0$  – пороговый коэффициент правдоподобия

Для  $P_1 = P_2$ ,  $c_{11} = c_{22}$ , и  $c_{12} = c_{21}$   $\lambda_0 = 1$ .

Тогда **пороговое значение**  $x_0$  – абсцисса точки пересечения функций  $p_1(x)$  и  $p_2(x)$ .

# Построение классификаторов

## Минимизация ошибки классификации

В случае множества классов риск решения о принадлежности объекта к классу  $w_j$  ( $j=1, 2, \dots, m$ ) описывается **платежной матрицей С** следующего вида (признак все еще один –  $x$ ):

$$C = \begin{vmatrix} c_{11} & c_{12} & \dots & c_{1m} \\ c_{21} & c_{22} & \dots & c_{2m} \\ \dots & \dots & \dots & \dots \\ c_{m1} & c_{m2} & \dots & c_{mm} \end{vmatrix}$$

$C_{jj}$  – потери при правильных решениях. Обычно  $C_{jj} = 0$  или  $C_{jj} < 0$ .

По обеим сторонам от главной диагонали стоят потери при ошибочных решениях

$C_{jg}$  – потери от принятия решения о принадлежности распознаваемого образа к классу  $w_g$ , когда на самом деле он принадлежит классу  $w_j$ .

**Замечание.** Платежная матрица в общем случае несимметрична.

Риск принятия решения о принадлежности образа к классу  $w_g$ , когда на самом деле он может принадлежать любому другому классу  $w_j$  – среднее значение потерь, стоящих в  $g$ -м столбце платежной матрицы.

Тогда этот **средний риск** можно представить:

$$R(w_g / x) = \sum_{j=1}^m p(w_j / x) c_{jg}$$

где  $p(w_j / x)$  – условная вероятность того, что образ со значением признака  $x$  принадлежит классу  $w_j$ , может быть вычислена по **формуле Байеса**

$$p(w_j / x) = \frac{p(x / w_j) P_j}{p(x)}$$

$P_j$  – априорная вероятность появления образа из класса  $w_j$

# Построение классификаторов

## Пример Байесовского классификатора

Цель - определение для каждого возможного значения признака класса, соответствующего минимальному риску ошибочного распознавания.

Один признак  $x$  – вес человека, разделенный на четыре диапазона (А – недостаточный вес, В – норма, С – превышение, D – избыточный вес).

Два класса:  $w_1$  – у человека нет гипертонии (здоров),  $w_2$  – есть (болен).

Результаты для 100 человек (**обучающее множество**) приведены в таблице.

	A	B	C	D	Всего
<b>x (вес)</b>	20	30	40	10	<b>100</b>
<b><math>w_1</math> (здоров)</b>	19	27	30	2	<b>78</b>
<b><math>w_2</math> (болен)</b>	1	3	10	8	<b>22</b>

Вероятность появления образа из класса  $w_1$ :  $P_1 = 78/100$ ; из класса  $w_2$ :  $P_2 = 22/100$ .

Вероятность наличия у образа значения признака  $x = A$ :  $p(A) = 20/100$ .

Вероятность наличия у образа из класса  $w_1$ , значения признака  $x = A$ :  $p(A/w_1) = 19/78$ .

Вероятность принадлежности образа классу  $w_1$ , при наличии у образа значения  $x = A$ :  
 $p(w_1/A) = (19/78 * 78/100)/(20/100) = 19/20$  (соответствует данным таблицы)

# Построение классификаторов

## Пример Байесовского классификатора

Результаты вычислений условных вероятностей для всех значений признака:

x	A	B	C	D
$p_x$	20/100	30/100	40/100	10/100
$p(w_1/x)$	19/20	27/30	30/40	2/10
$p(w_2/x)$	1/20	3/30	10/40	8/10

Риск принятия решения о принадлежности образа со значением признака  $x = A$  к классу  $w_1$  (здоров), когда на самом деле образ относится к классу  $w_2$  (болен):  $R(w_1/A) = p(w_2/A) c_{21}$ .

Риск принятия решения о принадлежности образа со значением признака  $x = A$  к классу  $w_2$  (болен), когда на самом деле образ относится к классу  $w_1$  (здоров):  $R(w_2/A) = p(w_1/A) c_{12}$ .

Результаты вычисления рисков при разных значениях  $c_{12}$  и  $c_{21}$ :

x	A	B	C	D
$R(w_1/x)$	<b>0,025</b>	<b>0,05</b>	<b>0,125</b>	0,4
$R(w_2/x)$	0,475	0,45	0,375	<b>0,1</b>
$R(w_1/x)$	<b>0,04</b>	<b>0,08</b>	0,2	0,64
$R(w_2/x)$	0,019	0,18	<b>0,15</b>	<b>0,04</b>

Если  $c_{12} = c_{21} = 0,5$ :

Если  $c_{21} = 0,8$ ;  $c_{12} = 0,2$ :

# Построение классификаторов

## Байесовский классификатор для вектора признаков

Принципиальных отличий при описании образа **вектором** признаков нет

В формуле вычисления среднего риска при отнесении распознаваемого образа к классу  $w_g$  значения признака  $x$  заменяется комбинацией значений **вектора** признаков  $\mathbf{x}$

$$R(w_g / \mathbf{x}) = \sum_{j=1}^m p(w_j / \mathbf{x}) c_{jg} \quad p(w_j / \mathbf{x}) = \frac{p(\mathbf{x} / w_j) P_j}{p(\mathbf{x})}$$

**Основная проблема** - определение условных вероятностей для всех возможных комбинаций значений **вектора** признаков  $\mathbf{x}$  в классе  $w_j$ , что может потребовать получения гигантского объема данных. Пусть мощность вектора признаков  $n = 8$ , каждый признак может принимать только десять различных значений ( $k = 10$ ), число классов  $m = 5$ . Тогда число возможных комбинаций значений признаков:  $N = k^n = 10^8$ , а число условных вероятностей  $p(\mathbf{x}/w_j)$  по всем классам, которые необходимо найти  $= mN = 500\,000\,000$ .

Учитывая необходимость обеспечения достоверности каждого значения  $p(\mathbf{x}/w_j)$ , может потребоваться сбор и последующая обработка информации о таком огромном количестве образов, что это будет далеко не всегда выполнимо.

**Замечание.** Разработкой алгоритма классификации заканчивается создание системы распознавания. На практике процесс разработки является итерационным приближением к минимально возможной ошибке распознавания.

# Построение классификаторов

## «Наивный» Байесовский классификатор для вектора признаков

Сократить число условных вероятностей  $p(\mathbf{x}/w_j)$  можно:

- исключив признаки с низкой информативностью (вектор признаков уменьшится);
- сократив число возможных значений каждого признака (у бинарного признака  $k = 2$ );
- используя только статистически независимые признаки ( $p(\mathbf{x}/w_j)$  вычисляется так...).

**Замечание.** События A и B статистически независимы, если  $p(A,B) = p(A)*p(B)$ . 

На практике признаки редко статистически независимы.

Если для каждого класса признаки статистически независимы, то для каждого класса необходимо найти всего  $N = \sum_{i=1}^n k_i$  значений признаков, где  $k_i$  – число возможных значений  $i$ -го признака ( $i = 1, n$ ). Для класса  $w_j$  условная вероятность  $p(\mathbf{x} / w_j) = \prod_{i=1}^n p(x_i^{(k)} / w_j)$ , где  $p(x_i^{(k)} / w_j)$  – вероятность появления  $k$ -го значения признака  $x_i$  у образа из класса  $w_j$ .

При отсутствии в классе  $w_j$  образов со значением  $x_k^{(k)}$   $p(x_k^{(k)} / w_j) = 0$ . Тогда произведение вероятностей тоже будет = 0. Для предотвращения этого число образов с каждым  $x_k$  увеличивается на 1, и тогда любая комбинация значений признаков  $\mathbf{x}$  будет встречаться в множестве образов хотя бы один раз. Такое изменение приводит к заметному увеличению вероятностей значений признаков, встречающихся в множестве образов малое число раз.

**Замечание.** Часто  $p(x_k / w_j)$  – очень маленькие числа и  $N$  велико. Для уменьшения ошибки при их многократном умножении вычисляется **сумма логарифмов** и при необходимости

находится экспонента:  $\log p(\mathbf{x} / w_j) = \sum_{k=1}^N \log(p(x_k / w_j)) = a \quad p(\mathbf{x} / w_j) = \exp(a)$

# Построение классификаторов

## «Наивный» Байесовский классификатор. Пример реализации

Два признака  $x_1$  – бинарный,  $x_2$  – может принимать три значения. Три класса  $w_1, w_2, w_3$ .

Призн.	$w_1$	$w_2$	$w_3$	Всего
$x_1=1$	1247	128	247	1622
$x_1=0$	253	872	253	1378
$x_2=A$	3	354	0	652
$x_2=B$	456	211	342	1009
$x_2=C$	1041	435	158	1339
Всего	1500	1000	500	3000
$p_1$	0,8313	0,128	0,494	0,541
$P_2$	0,1687	0,872	0,506	0,459
$P_3$	0,0020	0,354	0	0,217
$P_4$	0,3040	0,211	0,684	0,336
$p_5$	0,6940	0,435	0,316	0,446
$P_w$	0,5	0,3333	0,1667	1

Призн.	$w_1$	$w_2$	$w_3$	Всего
$x_1=1$	1248	129	248	1625
$x_1=0$	254	873	254	1381
$x_2=A$	4	355	1	655
$x_2=B$	457	212	343	1012
$x_2=C$	1042	436	159	1342
Всего	1502/1503	1002/1003	502/503	3006/3009
$p_1$	0,8309	0,1287	0,494	0,541
$P_2$	0,1691	0,8713	0,506	0,459
$P_3$	0,0027	0,3539	0,002	0,218
$P_4$	0,3041	0,2114	0,6819	0,336
$p_5$	0,6933	0,4347	0,3161	0,446
$P_w$	0,5	0,3333	0,1667	1

Неизвестный образ  $x$ :  $\{x_1=1, x_2=B\}$ . Цена ошибки одинаковая для всех классов. Тогда образ  $x$  следует отнести к классу  $w_j$  с максимальной  $p(w_j/x)$ . Если независимость признаков для каждого класса существует, то:  $p(x/w_1) = p(x_1=1/w_1) * p(x_2=B/w_1) = 0,8309 * 0,3041 = 0,2527$ . Найти  $p(x)$  из этой таблицы нельзя, т.к. из статистической независимости признаков для каждого класса не следует их статистическая независимость. Но на соотношение  $p(w_j/x)$   $p(x)$  не влияет, т.к. не зависит от номера класса, т.е. для классификации  $p(x)$  можно не находить. Примем  $p(x) = 0,541 * 0,336 = 0,182$ . Тогда  $p(w_1/x) = P_1 * p(x/w_1) / p(x) = 0,6318$ . Аналогично:  $p(w_2/x) = 0,0498$ ;  $p(w_3/x) = 0,3083$ . Для образа  $x$  наиболее вероятен класс  $w_1$ .

# Разделение пространства

**Пример.** Байесовский классификатор для 2-х признаков и 2-х классов. Классифицируем человека по двум признакам  $x, y$  (вес, рост), которые могут принимать по пять значений каждый, на два класса  $w_1, w_2$  (болен, здоров). Для каждой комбинации значений признаков надо найти наиболее вероятный класс.

Статистические данные для класса «Болен».

2671	210-230	1413	785	387	69	17
982	190-209	726	121	35	13	87
993	170-189	411	24	5	42	511
1824	150-169	67	18	37	193	1509
3530	130-149	36	94	436	1361	1603
10000	Рост/Вес	30-49	50-69	70-89	90-109	110-130
Сумма	10000	2653	1042	900	1678	3727

Расчетные вероятности сочетаний призна.

210-230	7,09	2,78	2,40	4,48	9,95
190-209	2,61	1,02	0,88	1,65	3,66
170-189	2,63	1,03	0,89	1,67	3,70
150-169	4,84	1,90	1,64	3,06	6,80
130-149	9,37	3,68	3,18	5,92	13,16
Рост/Вес	30-49	50-69	70-89	90-109	110-130

Все исходные данные условные!

Наиболее часто больные встречаются среди людей с недостаточным или избыточным весом. Реальная вероятность сочетания значений признаков в % – отношение значения ячейки, соответствующей данному сочетанию значений признаков к сумме образцов\*100. Например, вероятность в % того, что человек болен, если его признаки соответствуют сочетанию: вес в диапазоне 30-49, рост 170-189, равна  $100 \cdot (411/10000) = 4,11$ .

Расчетные вероятности даны в предположении статистической независимости признаков. Для сочетания 30-49 и 170-189 расчетная вероятность равна  $100 \cdot 0,2653 \cdot 0,0993 = 2,63$

**Замечание.** Вероятность дана в % для сокращения записи. Сумма 10000 для наглядности.



# Разделение пространства

**Пример.** Байесовский классификатор для 2-х признаков и 2-х классов  
Статистические данные для класса «Здоров». Расчетные вероятности сочетаний призна.

446	210-230	19	42	65	123	197
3099	190-209	43	345	586	1911	214
3857	170-189	71	552	2670	466	98
2193	150-169	125	1356	389	281	42
405	130-149	176	143	62	13	11
10000	Рост/Вес	30-49	50-69	70-89	90-109	110-130
Сумма	10000	434	2438	3772	2794	562

210-230	0,19	1,0873	1,6823	1,2461	0,2507
190-209	1,345	7,5554	11,689	8,6586	1,7416
170-189	1,6739	9,4034	14,549	10,776	2,1676
150-169	0,9518	5,3465	8,272	6,1272	1,2325
130-149	0,1758	0,9874	1,5277	1,1316	0,2276
Рост/Вес	30-49	50-69	70-89	90-109	110-130

Все исходные данные условные!

Наиболее часто здоровые встречаются среди людей с нормальным телосложением.

Все расчетные вероятности сочетания значений признаков в предположении о статистической независимости признаков в классе «Болен» не совпадают с реальными вероятностями. В классе «Здоров» они совпадают только для одного сочетания значений признаков – вес 30-49 и рост 210-230. Для подтверждения статистической независимости признаков необходимо такое совпадение для всех возможных сочетаний значений признаков в каждом классе.

**Замечание.** Суммы образов в классах совпадают, т.е. вероятность появления образа из одного или другого класса одинаковая. Реально эта вероятность зависит от возраста. Среди молодых чаще встречаются здоровые, среди пожилых – больные.

# Разделение пространства

**Пример.** Байесовский классификатор для 2-х признаков и 2-х классов

Вероятности появления образов из класса «Болен» и «Здоров» совпадают. Тогда образ с некоторым сочетанием значений признаков  $x$  следует отнести к классу, для которого больше условная вероятность данного сочетания значений признаков. Найдя такой класс для всех возможных сочетаний значений признаков и пометив им соответствующую точку в пространстве признаков, можно получить разделение пространства признаков.

210-230	<u>Б</u>	<u>Б</u>	Б	<u>З</u>	<u>З</u>
190-209	Б	<u>З</u>	<u>З</u>	<u>З</u>	<u>З</u>
170-189	Б	<u>З</u>	<u>З</u>	<u>З</u>	Б
150-169	<u>З</u>	<u>З</u>	<u>З</u>	З	Б
130-149	<u>З</u>	<u>З</u>	Б	<u>Б</u>	<u>Б</u>
Рост/Вес	30-49	50-69	70-89	90-109	110-130

210-230	<u>Б</u>	Б	Б	Б	<u>Б</u>
190-209	Б	<u>З</u>	<u>З</u>	<u>З</u>	Б
170-189	Б	<u>З</u>	<u>З</u>	<u>З</u>	Б
150-169	Б	<u>З</u>	<u>З</u>	<u>З</u>	Б
130-149	<u>Б</u>	Б	Б	Б	<u>Б</u>
Рост/Вес	30-49	50-69	70-89	90-109	110-130

Разделение пространства на основе реальных значений условных вероятностей

Разделение пространства на основе расчетных значений условных вероятностей

Условные обозначения: желтые клетки – классы не совпадают;

- жирно и подчеркнуто – вероятность одного класса значительно выше чем другого;
- жирно – вероятность одного класса заметно выше чем другого;
- курсив – вероятность одного класса незначительно выше чем другого.

# Классификатор по энтропии

Основан на методе линейной логистической регрессии и энтропии множества событий, учитывает зависимость между признаками.

Пусть в обучающем множестве  $N$  образов, число классов -  $m$ , число признаков -  $n$ ,  $i$ -й признак может принимать  $k_i$  значений. Назовем событием наличие одного из возможных значений одного из признаков у одного из образов из одного из классов обучающего мно-

жества. Число возможных событий в классе –  $D = \sum_{i=1}^n k_i$ . Бинарная функция  $f_{wx} = 1$ , если

у образа из класса  $w_j$   $i$ -й признак имеет значение  $x_k$ , иначе  $f_{wx} = 0$ .  $f_{wx}$  – **индикатор события**, т.е. определяет существование **каждого** возможного события (число значений  $f_{wx} = D$ ). В каждом классе каждый индикатор  $f_{wx}$  имеет свой вес –  $\lambda_{wx}$ , т.е. общее число весов –  $m \cdot D$ . Вероятность отнесения образа с вектором признаков  $\mathbf{x}$  к классу  $w_j$ :

$$p(w_j / \mathbf{x}) = \exp \sum_{w=j, x=1}^{x=D} \lambda_{wx} f_{wx} \Bigg/ \left( \sum_{w=1}^m \exp \sum_{w=j, x=1}^{x=D} \lambda_{wx} f_{wx} \right)$$

В числителе – экспонента суммы весов имеющихся значений признаков вектора  $\mathbf{x}$  для  $w_j$ .  
В знаменателе – сумма аналогичных экспонент для всех классов, включая и класс  $w_j$ .

Набор весов  $\lambda_{wx}$  для всех классов определяется по обучающему множеству.

# Классификатор по энтропии

Набор весов должен соответствовать максимуму энтропии событий.

Энтропия множества событий максимальна, если события равновероятны. Необходимо найти набор весов, обеспечивающий максимум функции правдоподобия – произведение вероятностей всех образов обучающего множества. Логарифм произведения – сумма логарифмов сомножителей, т.е. для функции правдоподобия  $P(\lambda_{wx})$ :

$$\log(P(\lambda)) = \sum_{i=1}^N \log \left( \exp \sum_{w=j, x=1}^{x=D} \lambda_{wx} f_{wx} / \left( \sum_{w=1}^m \exp \sum_{w=j, x=1}^{x=D} \lambda_{wx} f_{wx} \right) \right)$$

Набор весов  $\lambda_{wx}$  находится методом **линейной логистической регрессии** (численная оптимизация). Например, методом градиентного спуска, позволяющего найти набор весов при котором градиент функции максимально близок к нулю.

Оптимальный набор весов  $\lambda_{wx}$  – это набор, с помощью которого классификация **любого** образа из **обучающего** множества выполняется правильно.

## Преимущества метода:

- оптимальное распределение и максимум функции правдоподобия всегда существуют;
- у функции правдоподобия один глобальным максимум;
- найденный данным методом набор весов соответствуют распределению вероятностей событий с максимальной возможной энтропией.

# Композиция классификаторов

Сильный классификатор в результате обучения позволяет получить произвольно малую ошибку распознавания

Слабый классификатор в результате обучения дает ошибку больше 0,5, но ее величина недостаточна для распознавания

**Идея** – композиция (объединение) слабых классификаторов и принятие решения после анализа результатов их работы. Пример – группа экспертов оценивает что-то (качество фото, выступление фигуриста)

## Бэггинг (Bagging - bootstrap aggregation)

предложил L. Breiman

Дано: – множество  $X$  из  $n$  образов (классификация их известна)

– множество  $s$  слабых классификаторов (возможно только один)

Алгоритм:

1. Из  $X$  случайным образом формируется  $s$  обучающих выборок (возможно пересечение выборок по образам)
2. Каждый из  $s$  классификаторов обучается на своей выборке
3. Получают результаты распознавания неизвестного образа  $x$  каждым классификатором
4. Решение – простое голосование (к какому классу образ  $x$  отнесен большинством классификаторов) или среднее из их решений

# Композиция классификаторов

## Преимущества бэггинга:

- Из-за различности классификаторов (обучающих выборок) их ошибки взаимно компенсируются при голосовании
- Образы-выбросы из  $X$  могут не попасть в некоторые из обучающих выборок
- Дисперсия ошибок (нестабильность) композиции классификаторов меньше, чем нестабильность каждого классификатора в отдельности

## Бустинг (*boosting*)

предложил R.Schapire

Итерационный метод (развитие бэггинга) – использование весовой версии одних обучающих данных и одного слабого классификатора

## Отличия бустинга и бэггинга

В бэггинге каждый образ из  $X$  имеет одинаковые шансы попасть в каждую обучающую выборку. В бустинге обучающая выборка на каждой итерации определяется, исходя из ошибок классификации на предыдущих итерациях. Большие веса назначаются ошибочно распознанным образам, что позволяет сосредоточиться на них на следующей итерации. Классификаторы образуются последовательно, различаясь только весами обучающих данных

# Композиция классификаторов

## AdaBoost (Adaptive Boosting)

Дано – обучающее множество  $X$  и один слабый классификатор

Алгоритм:

1. В начале обучения веса всех образов в  $X$  принимаются одинаковыми
2. Обучение – на каждом шаге классификатор обучается на всем  $X$ , по результатам распознавания образов из  $X$  (ошибки классификации) вычисляются веса образов и вес данного классификатора
3. После  $s$  шагов получают  $s$  отличающихся весами классификаторов
4. Распознавание неизвестного образа  $x$  каждым из  $s$  классификаторов (результат – голосование с учетом весов классификаторов)

Преимущества метода: простота реализации и модификации; композиция алгоритмов превосходит по качеству базовый алгоритм; возможность идентифицировать выбросы (образы в  $X$ , веса которых в процессе обучения принимают наибольшие значения)

Недостатки метода: очень слабый классификатор дает малые изменения весов образов; требуется достаточно большое обучающее множество; при сильном шуме возникает переобучение (слишком большой вес назначается «плохим» образам  $X$  – выбросам, а «хорошие» образы практически не участвуют в обучении)

Возможные решения: ограничение числа шагов обучения; ограничение весов образов в  $X$

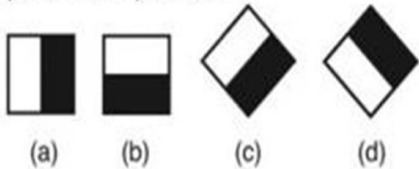
# Композиция классификаторов

## Метод П.Виолы и М.Джонса (2001г) для распознавания объектов на изображении

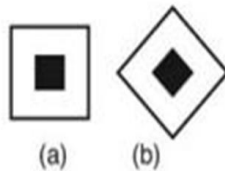
Метод комбинирует четыре концепции:

- Использование для вычисления признаков объектов функций Хаара
- Интегральное представление изображения по этим признакам
- Классификаторы на основе алгоритма AdaBoost
- Комбинирование классификаторов в каскадную структуру

1. Граничные признаки

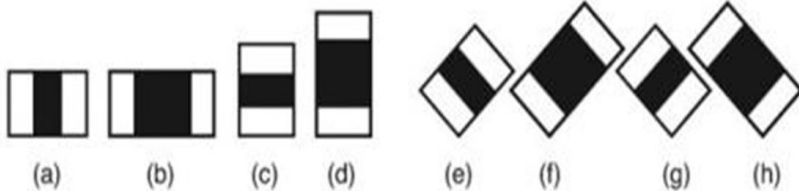


3. «Центральные» признаки



Функции (признаки) Хаара

2. Линейные признаки



Метод является наиболее часто применяемым. Качество сопоставимо с более медленными и сложными одноэтапными классификаторами. Позволяет распознавать объекты в реальном времени.

Информация в Интернете: [https://levutkin.github.io/files/Machine\\_Learning\\_LTU\\_7.pdf](https://levutkin.github.io/files/Machine_Learning_LTU_7.pdf)  
[https://vuzlit.ru/789252/kaskadnyy\\_klassifikator](https://vuzlit.ru/789252/kaskadnyy_klassifikator)