ВВЕДЕНИЕ

§1. Предмет и задачи распознавания образов.

Определение. *Распознавание образов – это научная дисциплина, целью которой является классификация объектов называемые образами по нескольким категориям или классам.*

Классификация основывается на прецедентах.

Определение. *Прецедент – ранее классифицированный объект, принимаемый как образец при использовании данного классификатора.*

Будем считать, что все объекты или явления разбиты на конечное число классов. Для каждого класса известно и изучено конечное число объектов – прецедентов. Задача распознавания образов состоит в том, чтобы отнести новый распознаваемый объект к какому-либо классу.

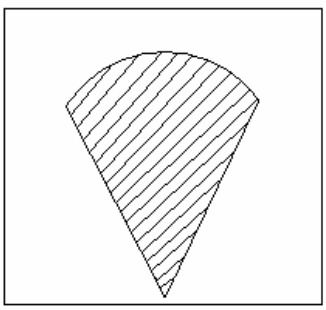
Задача распознавания образов ставится на первый план в большинстве интеллектуальных систем.

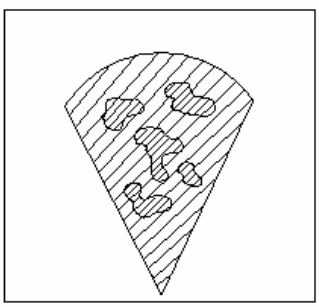
Рассмотрим примеры компьютерных интеллектуальных систем.

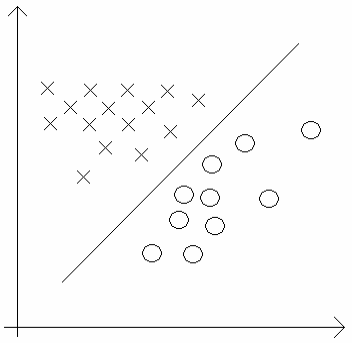
1) Машинное зрение. В данном случае – это получение изображения через камеру и составление его описания через какие либо параметры.
2) Символьное распознавание – это распознавание букв или цифр.
    a. Optical Character Recognition (OCR);
    b. Ввод и хранение документов;
    c. Pen Computer;
    d. Обработка чеков в банках;
    e. Обработка почты.
3) Диагностика в медицине.
    a. Маммография;
    b. Постановка диагноза по истории болезни;
    c. Электрокардиограмма.
4) Геология.
5) Распознавание речи.
6) Распознавание отпечаток пальцев, лица, подписи, жестов.

Любое распознавание происходит на основе некоторых признаков.

Пример. Рассмотрим диагностику печени. Если она здорова, то компьютер на мониторе будет выдавать равномерный и однородный цвет (рис. слева). В противном случае, на экране монитора на фоне этой однородности будут наблюдаться более темные пятна (рис. в центре). В этих двух ситуациях на координатной плоскости можно выделить два класса, разделенные линией решения (рис. справа).





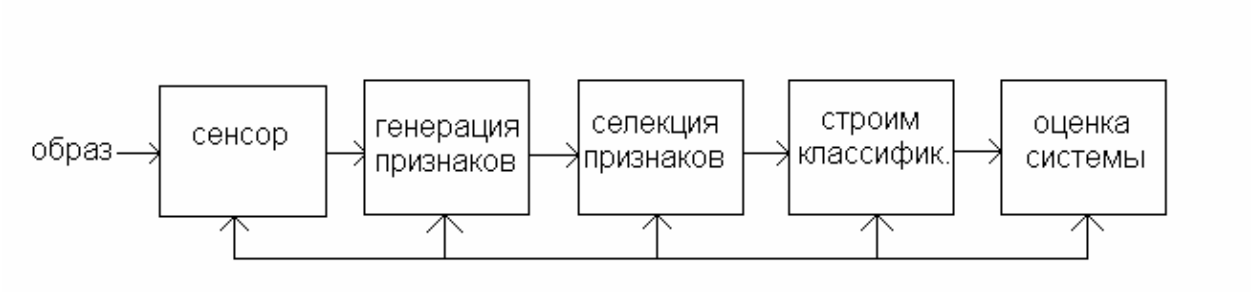


Определение. *Признак – это некоторое количественное измерение объекта произвольной природы.*

В процессе распознавания все признаки стараются разделить так называемой поверхностью решения.

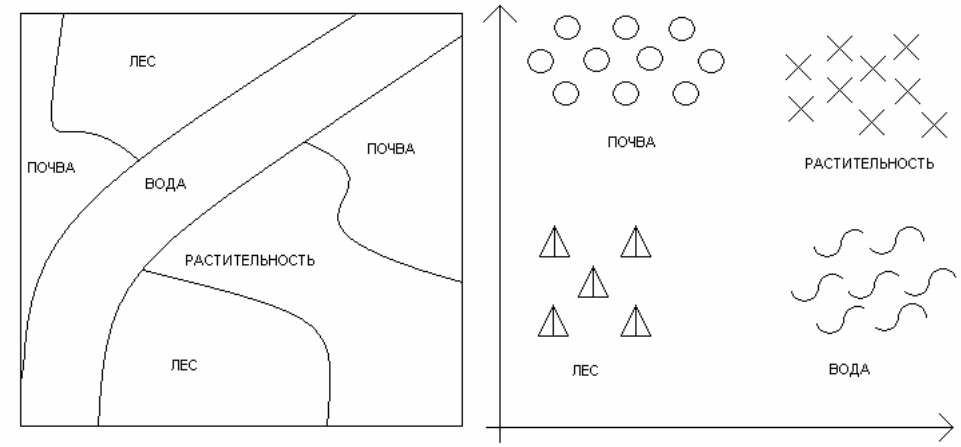
Рассмотрим практические вопросы, возникающие в задачах распознавания.

1) Как генерировать вектора признаков? Задача генерации признаков.
2) Какое наилучшее количество признаков? Задача селекции признаков.
3) Как построить классификатор? Задача построения классификатора.
4) Как оценить классификатор на предмет ошибок классификации? Задача оценки системы.



Рассмотрим две ситуации. Первая, когда обучающие прецеденты известны. В таком случае распознавание называется наблюдаемым или распознаванием с учителем. Вторая, когда обучающие прецеденты, помеченные метками классов, не заданы. Задача состоим в том, чтобы выявить сходство группы объектов. В таком случае распознавание называется ненаблюдаемым или без учителя. В литературе можно встретить название кластеризация.

Пример. Рассмотрим съему со спутника и классификацию поверхности по отраженной энергии. На рисунках изображены пример снимка из космоса (рисунок слева) и его классификация (рисунок справа).



Пример. Рассмотрим кластеризацию в общественных науках. Целью данной задачи является выявление закономерностей для выбора правильных действий государства. Например, связь между ВНП, уровнем грамотности и детской смертности. В данном случае страны можно представить трехмерными векторами, а задача заключается в построении меры сходства этих векторов, а потом построение схемы кластеризации (выбора групп) по этой мере.

§2. Постановка задачи распознавания образов

Ω – множество объектов распознавания.

$\omega: \varpi \in \Omega$ – объект распознавания.

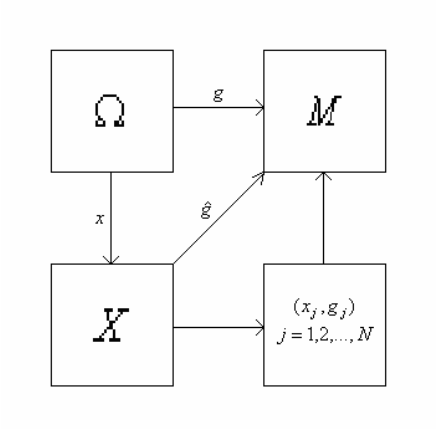
$g(\omega): \Omega \to \mathrm{M}$, $\mathrm{M} = \{1,2,...,m\}$ – индикаторная функция, разбивающая пространство объектов распознавания Ω на m непересекающихся классов $\Omega^1, \Omega^2, ..., \Omega^m$.

X – пространство признаков

$x(\omega): \Omega \to \mathrm{X}$ – образ объекта ω.

$\hat{g}(x): \mathrm{X} \to \mathrm{M}$ – решающее правило; оценка для $g(\omega)$ на основании $x(\omega)$, т.е. $\hat{g}(x) = \hat{g}(x(\omega))$.

Пусть $x_j = x(\omega_j)$, $j = 1,2...,N$ – доступная наблюдателю информация о функциях $g(\omega)$ и $x(\omega)$, но сами эти функции наблюдателю неизвестны. Тогда (g_j, x_j), $j = 1,2...,N$ – множество прецедентов.



Задача заключается в построении такого решающего правила $\hat{g}(x)$, чтобы распознавание проводилось с минимальным числом ошибок.

Пусть пространство признаков $X = R^l$; качество решающего правила измеряют частотой появления правильных решений. Обычно его оценивают, наделяя множество объектов Ω некоторой σ-алгеброй подмножеств и вероятностной мерой. Тогда задача записывается в виде $\min P\{\hat{y}(x(\omega)) \neq g(\omega)\}$.

Г Л А В А 1

КЛАССИФИКАЦИЯ НА ОСНОВЕ БАЙЕСОВСКОЙ ТЕОРИИ РЕШЕНИЙ

Байесовский подход исходит из статистической природы наблюдений. За основу берется предположение о существовании вероятностной меры на пространстве образов, которая либо известна, либо может быть оценена. Цель состоит в разработке такого классификатора, который будет правильно определять наиболее вероятный класс для пробного образа. Тогда задача состоит в определении "наиболее вероятного" класса.

§1. Постановка задачи

Задано M классов $\Omega_1,\Omega_2,...,\Omega_M$. $P(\Omega_i|x)$, $i=1,2,...,M$ – вероятность того, что неизвестный образ, представляемый вектором x, принадлежит классу Ω_i. $P(\Omega_i|x)$ задает распределение значению x после эксперимента, т.е. после того, как значение было наблюдаемо.

Определение. *$P(\Omega_i|x)$ называется апостериорной вероятностью.*

Рассмотрим случай двух классов Ω_1 и Ω_2. Пусть решающее правило таково: если $P(\Omega_1|x) > P(\Omega_2|x)$, то x классифицируется в Ω_1, иначе в Ω_2. Необходимо получить апостериорные вероятности $P(\Omega_i|x)$, $i=1,2$. Будем считать, что у нас достаточно данных для определения вероятности $P(\Omega_i)$, $i=1,2$ и функции правдоподобия x по отношению к Ω_i, $P(x|\Omega_i)$, $i=1,2$. $P(\Omega_i)\approx\frac{N_i}{N}$, где N_i – число прецедентов из Ω_i, $i=1,2$. N – общее число прецедентов. $P(x|\Omega_i)$ – распределение вектора признаков в класс Ω_i. Если распределения неизвестны, то их можно посчитать на множестве прецедентов.

Определение. *Байесовская формула – это формула, позволяющая вычислить апостериорные вероятности событий через априорные вероятности.*

Пусть $A_1,A_2,...,A_n$ – полная группа несовместных событий. $\bigcup_{i=1}^{n} A_i=\Omega$. $A_i\cap A_j=\varnothing$, при $i\neq j$. Тогда апостериорная вероятность имеет вид:

$$P(A_i|B)=\frac{P(A_i)P(B|A_i)}{\sum_{i=1}^{n}P(A_i)P(B|A_i)},$$

где $P(A_i)$ – априорная вероятность события A_i, $P(B|A_i)$ – условная вероятность события B при условии, что произошло событие A_i. Данная формула была получена и доказана Т. Бейесом в 1763 году.

Итак, Байесовский подход к статистическим задачам основывается на предположении о существовании некоторого распределения вероятностей для каждого параметра. Недостатком этого метода является необходимость постулирования существования как априорного распределения для неизвестного параметра, так и знание его формы.

Рассмотрим получение апостериорной вероятности $P(\Omega|x)$, зная $P(\Omega)$ и $P(x|\Omega)$.

$$P(AB)=P(A|B)P(B),\ P(AB)=P(B|A)P(A)$$

$$P(A|B)P(B)=P(B|A)P(A)$$

$$P(B|A)=\frac{P(A|B)P(B)}{P(A)}$$

Если $P(A)$ и $P(A|B)$ описываются плотностями $p(x)$ и $p(x|B)$, то

$$P(B|x) = \frac{p(x|B)P(B)}{p(x)} \Rightarrow P(\Omega_i|x) = \frac{p(x|\Omega_i)P(\Omega_i)}{p(x)}.$$

При проверке классификации сравнение $P(\Omega_1|x)$ и $P(\Omega_2|x)$ эквивалентно сравнению $p(x|\Omega_1)P(\Omega_1)$ и $p(x|\Omega_2)P(\Omega_2)$. В случае, когда $P(\Omega_1|x) = P(\Omega_2|x)$, считается, что мера множества x равна нулю.

§2. Ошибка классификации

Определение. *Вероятность* $P_e = P(x \in R_2, \Omega_1) + P(x \in R_1, \Omega_2)$ *называется ошибкой классификации,* $R_1 = \{x : P(\Omega_1)p(x|\Omega_1) > P(\Omega_2)p(x|\Omega_2)\}$, $R_2 = \{x : P(\Omega_1)p(x|\Omega_1) < P(\Omega_2)p(x|\Omega_2)\}$ *– области решения (*$\Omega_1 \cap \Omega_2 = \varnothing$*).*

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

Доказательство. Рассмотрим ошибку классификации:

$$P_e = P(x \in R_2, \Omega_1) + P(x \in R_1, \Omega_2) =$$
$$= P(\Omega_1)\int_{R_2} p(x|\Omega_1)dx + P(\Omega_2)\int_{R_1} p(x|\Omega_2)dx =$$
$$= P(\Omega_1)\left(1 - \int_{R_1} p(x|\Omega_1)dx\right) + P(\Omega_2)\int_{R_1} p(x|\Omega_2)dx =$$
$$= P(\Omega_1) - P(\Omega_1)\int_{R_1} p(x|\Omega_1)dx + P(\Omega_2)\int_{R_1} p(x|\Omega_2)dx =$$

Учитывая формулу Байеса: $p(x|\Omega_i) = \frac{P(\Omega_i|x)p(x)}{P(\Omega_i)}$, $i = 1,2$ получим:

$$= P(\Omega_1) - P(\Omega_1)\int_{R_1} \frac{P(\Omega_1|x)p(x)}{P(\Omega_1)}dx + P(\Omega_2)\int_{R_1} \frac{P(\Omega_2|x)p(x)}{P(\Omega_2)}dx =$$
$$= P(\Omega_1) - \int_{R_1} P(\Omega_1|x)p(x)dx + \int_{R_1} P(\Omega_2|x)p(x)dx =$$
$$= P(\Omega_1) - \int_{R_1} p(x)\left(P(\Omega_1|x) - P(\Omega_2|x)\right)dx$$

Таким образом, минимум достигается, когда $R_1 = \{x : P(\Omega_1|x) > P(\Omega_2|x)\}$. R_2 выбирается из остальных точек.

ч.т.д.

Данная теорема была доказана для двух классов Ω_1 и Ω_2. Обобщим ее на M классов.

Пусть вектор признаков x относится к классу Ω_i, если $P(\Omega_i|x) > P(\Omega_j|x)$, при $i \neq j$, $i = 1,2,...,M$, $j = 1,2,...,M$. Соответственно необходимо доказать, что данное правило минимизирует вероятность ошибки классификации. Для доказательства следует воспользоваться формулой правильной классификации $P_r = 1 - P_e$.

Доказательство. Воспользуемся формулой правильной классификации $P_r = 1 - P_e$.

$$P_r = P(x \in R_1, \Omega_1) + P(x \in R_2, \Omega_2) + ... + P(x \in R_l, \Omega_l) =$$
$$= \sum_{i=1}^{l} P(x \in R_i|\Omega_i)P(\Omega_i) =$$

$$= \sum_{i=1}^{l} P(\Omega_i) \int_{R_i} p(x|\Omega_i)dx =$$

$$= P(\Omega_1)\left(1 - \sum_{i=2}^{l} \int_{R_i} p(x|\Omega_1)dx\right) + \sum_{i=2}^{l} P(\Omega_i) \int_{R_i} p(x|\Omega_i)dx =$$

$$= P(\Omega_1) - \sum_{i=2}^{l}\left[P(\Omega_1)\int_{R_i} p(x|\Omega_1)dx - P(\Omega_i)\int_{R_i} p(x|\Omega_i)dx\right] =$$

Учитывая формулу Байеса: $p(x|\Omega_i) = \dfrac{P(\Omega_i|x)p(x)}{P(\Omega_i)}$, $i = 1,2,...,l$ получим:

$$= P(\Omega_1) - \sum_{i=2}^{l}\left[P(\Omega_1)\int_{R_i} \frac{P(\Omega_1|x)p(x)}{P(\Omega_1)}dx - P(\Omega_i)\int_{R_i} \frac{P(\Omega_i|x)p(x)}{P(\Omega_i)}dx\right] =$$

$$= P(\Omega_1) - \sum_{i=2}^{l}\left[\int_{R_i} P(\Omega_1|x)p(x)dx - \int_{R_i} P(\Omega_i|x)p(x)dx\right] =$$

$$= P(\Omega_1) - \sum_{i=2}^{l}\int_{R_i} p(x)\left[P(\Omega_1|x) - P(\Omega_i|x)\right]dx$$

Таким образом, максимум достигается, когда $P(\omega_1|x) < P(\omega_i|x)$. Аналогично для всех $j = 1,2,...,l$ максимум достигается, когда $R_j = \{x : P(\omega_j|x) < P(\omega_i|x)\}$.

ч.т.д.

§3. Минимизация среднего риска

Вероятность ошибки классификации – не всегда лучший критерий проверки классификатора. Рассмотрим задачу классификации по M классам. R_j, $j = 1,2,...,M$ – области предпочтения классов ϖ_j. Предположим, что вектор x из класса Ω_k лежит в R_i, $i \neq k$, т.е. классификация происходит с ошибкой. Свяжем с этой ошибкой штраф λ_{ki} называемый потерей, что объект из класса Ω_k был принят за объект из класса Ω_i. Обозначим через $L = \|\lambda_{ki}\|$ матрицу потерь.

Определение. *Выражение* $r_k = \sum_{i=1}^{M} \lambda_{ki} P\{x \in R_i|\Omega_k\} = \sum_{i=1}^{M} \lambda_{ki} \int_{R_i} p(x|\Omega_k)dx$ *называется риском при классификации объекта класса* ω_k.

Определение. *Выражение* $r = \sum_{i=1}^{M} r_k P(\Omega_k)$ *называется общим средним риском.*

Введя понятие риска, мы автоматически поставили задачу о минимизации этого риска. Преобразуем выражение общего среднего риска:

$$r = \sum_{i=1}^{M} r_k P(\Omega_k) = \sum_{k=1}^{M} P(\Omega_k) \sum_{i=1}^{M} \lambda_{ki} \int_{R_i} p(x|\Omega_k)dx =$$

$$= \sum_{i=1}^{M}\left(\sum_{k=1}^{M} P(\Omega_k)\lambda_{ki} \int_{R_i} p(x|\Omega_k)dx\right) =$$

$$= \sum_{i=1}^{M}\int_{R_i}\left(\sum_{k=1}^{M} \lambda_{ki} p(x|\Omega_k)P(\Omega_k)\right)dx$$

Из этого выражения видно, что риск минимален, когда каждый из интегралов в данной сумме минимален, т.е. $x \in R_i$, если $l_i < l_j$, при $i \neq j$, где $l_i = \sum_{k=1}^{M} \lambda_{ki} p(x|\Omega_k) P(\Omega_k)$, $l_j = \sum_{k=1}^{M} \lambda_{kj} p(x|\Omega_k) P(\Omega_k)$.

Пример. Рассмотрим ситуацию радиолокационной разведки. На экране радара отражаются не только цели, но и помехи. Такой помехой может служить стая птиц, которую можно принять за небольшой самолет. В данном случае это двух классовая задача.

Рассмотрим матрицу штрафов: $L = \|\lambda_{ki}\|$, $i = 1,2$, $k = 1,2$. λ_{ki} – это штраф за принятие объекта из класса k за объект класса i. Тогда

$$l_1 = \lambda_{11} p(x|\Omega_1) P(\Omega_1) + \lambda_{21} p(x|\Omega_2) P(\Omega_2)$$
$$l_2 = \lambda_{12} p(x|\Omega_1) P(\Omega_1) + \lambda_{22} p(x|\Omega_2) P(\Omega_2)$$

Пусть x относится у классу Ω_1, если $l_1 < l_2$, т.е.

$$\lambda_{11} p(x|\Omega_1) P(\Omega_1) + \lambda_{21} p(x|\Omega_2) P(\Omega_2) < \lambda_{12} p(x|\Omega_1) P(\Omega_1) + \lambda_{22} p(x|\Omega_2) P(\Omega_2)$$
$$(\lambda_{21} - \lambda_{22}) p(x|\Omega_2) P(\Omega_2) < (\lambda_{12} - \lambda_{11}) p(x|\Omega_1) P(\Omega_1)$$

Т.к. $\lambda_{21} > \lambda_{22}$ и $\lambda_{12} > \lambda_{11}$, то

$$\frac{p(x|\Omega_1)}{p(x|\Omega_2)} > \frac{\lambda_{21} - \lambda_{22}}{\lambda_{12} - \lambda_{11}} \cdot \frac{P(\Omega_2)}{P(\Omega_1)}$$

Таким образом, мы получили отношение правдоподобия, которое описывает предпочтение класса Ω_1 классу Ω_2.

Пример. Рассмотрим двух классовую задачу, в которой для единственного признака x известна плотность распределения:

$$p(x|\Omega_1) = \frac{1}{\sqrt{\pi}} \exp\left(-x^2\right)$$
$$p(x|\Omega_2) = \frac{1}{\sqrt{\pi}} \exp\left(-(x-1)^2\right)$$

Пусть, также, априорные вероятности $P(\Omega_1) = P(\Omega_2) = \frac{1}{2}$.

Задача – вычислить пороги для

a) минимальной вероятности ошибки

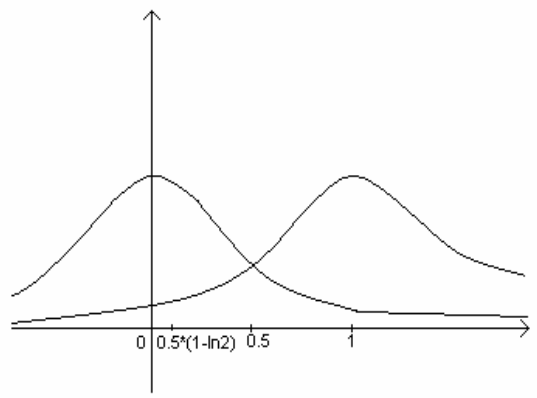
b) минимальной потере при матрице риска $L = \begin{pmatrix} 0 & 0.5 \\ 1 & 0 \end{pmatrix}$.

Решение задачи a):

$$p(x|\Omega_1) P(\Omega_1) = p(x|\Omega_2) P(\Omega_2)$$
$$\exp\left(-x^2\right) = \exp\left(-(x-1)^2\right)$$
$$-x^2 = -(x-1)^2$$
$$\hat{x} = \frac{1}{2}$$

Решение задачи b):

$$\frac{p(x|\Omega_1)}{p(x|\Omega_2)} = \frac{\lambda_{21} - \lambda_{22}}{\lambda_{12} - \lambda_{11}} \cdot \frac{P(\Omega_2)}{P(\Omega_1)}$$
$$\frac{\exp\left(-x^2\right)}{\exp\left(-(x-1)^2\right)} = \frac{1-0}{0.5-0} = \frac{1}{1/2} = 2$$



$$\exp\left(-x^2\right)=2\exp\left(-(x-1)^2\right)$$
$$-x^2=\ln 2-(x-1)^2$$
$$\widetilde{x}=\frac{1}{2}\left(1-\ln 2\right)$$

Пример. Рассмотрим двух классовую задачу с Гауссовскими плотностями распределения $p(x|\Omega_1)\cong N(0,\sigma^2)$ и $p(x|\Omega_2)\cong N(1,\sigma^2)$ и матрицей потерь $L=\begin{pmatrix}0 & \lambda_{12}\\ \lambda_{21} & 0\end{pmatrix}$.

Задача – вычислить порог для проверки отношения правдоподобия.

Решение. С учетом матрицы потерь отношение правдоподобия

$$\frac{p(x|\Omega_1)}{p(x|\Omega_2)}=\frac{\lambda_{21}-\lambda_{22}}{\lambda_{12}-\lambda_{11}}\cdot\frac{P(\Omega_2)}{P(\Omega_1)}$$

запишется в виде

$$\frac{p(x|\Omega_1)}{p(x|\Omega_2)}=\frac{\lambda_{21}}{\lambda_{12}}\cdot\frac{P(\Omega_2)}{P(\Omega_1)}$$

Запишем плотности распределения

$$p(x|\Omega_1)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{x^2}{2\sigma^2}\right);\ p(x|\Omega_2)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-1)^2}{2\sigma^2}\right)$$
$$\frac{p(x|\Omega_1)}{p(x|\Omega_2)}=\frac{\lambda_{21}}{\lambda_{12}}\cdot\frac{P(\Omega_2)}{P(\Omega_1)}=\exp\left(\frac{(x-1)^2}{2\sigma^2}-\frac{x^2}{2\sigma^2}\right)$$
$$x^2-(x-1)^2=-2\sigma^2\ln\left(\frac{\lambda_{21}}{\lambda_{12}}\cdot\frac{P(\Omega_2)}{P(\Omega_1)}\right)$$
$$x=\frac{1}{2}-\sigma^2\ln\left(\frac{\lambda_{21}}{\lambda_{12}}\cdot\frac{P(\Omega_2)}{P(\Omega_1)}\right)$$

Пример. Рассмотрим двух классовую задачу с матрицей потерь $L=\|\lambda_{ki}\|$, $k=1,2$, $i=1,2$. Пусть ε_1 – вероятность ошибки, соответствующая вектору из класса Ω_1 и ε_2 – вероятность ошибки, соответствующая вектору из класса Ω_2. Задача – найти средний риск.

Решение.

$$r=\sum_{i=1}^{M}r_k P(\Omega_k)=$$
$$=\sum_{i=1}^{M}\left(\sum_{k=1}^{M}P(\Omega_k)\lambda_{ki}\int_{R_i}p(x|\Omega_k)dx\right)=$$
$$=\lambda_{11}(1-\varepsilon_1)P(\Omega_1)+\lambda_{12}\varepsilon_1 P(\Omega_1)+\lambda_{21}\varepsilon_2 P(\Omega_2)+\lambda_{22}(1-\varepsilon_2)P(\Omega_2)=$$
$$=\lambda_{11}P(\Omega_1)+(\lambda_{12}-\lambda_{11})\varepsilon_1 P(\Omega_1)+(\lambda_{21}-\lambda_{22})\varepsilon_2 P(\Omega_2)+\lambda_{22}P(\Omega_2)$$

Пример. Доказать, что в задаче классификации по M классам, вероятность ошибки классификации ограничена: $P_e=\frac{M-1}{M}$.

Указание: показать, что $\max\limits_{i=1,\ldots,M}P(\varpi_i|x)\geq\frac{1}{M}$.

§4. Дискриминантная функция и поверхность решения.

Минимизация риска и вероятности ошибки эквивалентны разделению пространства признаков на M областей. Если области R_i и R_j смежные, то они разделены поверхно-

стью решения в многомерном пространстве. Для случая минимизации вероятности ошибки поверхность решения задается уравнением:

$$P(\omega_i|x) - P(\omega_j|x) = 0$$

В данном уравнении приходится оперировать с вероятностями. Иногда вместо вероятностей удобнее работать с функцией от вероятности:

$$g_i(x) = f\left(P(\omega_i|x)\right),$$

где функция f монотонно возрастает.

Определение. *Функция* $g_i(x) = f\left(P(\omega_i|x)\right)$ *называется дискриминантной функцией.*

Таким образом, поверхность решения будет задаваться уравнением:

$$g_i(x) - g_j(x) = 0,\ i = 1,2,...,M,\ j = 1,2,...,M,\ i \neq j.$$

Для задачи классификации по вероятности ошибки или риску не всегда удается вычислить вероятности. В этом случае бывает более предпочтительно вычислить разделяющую поверхность на основе другой функции стоимости. Такие подходы дают решения, субоптимальные по отношению к Байесовской классификации.

§5. Байесовский классификатор для нормального распределения.

Распределение Гаусса очень широко используется по причине вычислительного удобства и адекватности во многих случаях. Рассмотрим же многомерную плотность нормального распределения $N(\mu_i, \Sigma_i)$:

$$p(x|\Omega_i) = \frac{1}{(2\pi)^{l/2}|\Sigma_i|^{1/2}} \cdot \exp\left(-\frac{1}{2}\frac{(x-\mu_i)(x-\mu_i)^{\mathrm{T}}}{\Sigma_i}\right),\ i = 1,2,...,M$$

где $\mu_i = \mathrm{E}[X]$ – математическое ожидание случайной величины X в классе Ω_i,

Σ_i – матрица ковариации размерности $l \times l$ для класса Ω_i, $\Sigma_i = \mathrm{E}\left[(x-\mu_i)(x-\mu_i)^{\mathrm{T}}\right]$,

$|\Sigma_i|$ – определитель матрицы ковариации.

5.1. Квадратичная поверхность решения. На основе этих данных необходимо построить байесовский классификатор. Рассмотрим логарифмическую дискриминантную функцию:

$$\begin{aligned}
g_i(x) &= \ln(P(\Omega_i|x)) = \\
&= \ln(p(x|\Omega_i)P(\Omega_i)) = \\
&= \ln p(x|\Omega_i) + \ln P(\Omega_i) = \\
&= -\frac{1}{2}\frac{(x-\mu_i)}{\Sigma_i}(x-\mu_i)^{\mathrm{T}} + \ln P(\Omega_i) + \ln\frac{1}{(2\pi)^{l/2}|\Sigma_i|^{1/2}} = \\
&= -\frac{1}{2}\frac{(x-\mu_i)}{\Sigma_i}(x-\mu_i)^{\mathrm{T}} + \ln P(\Omega_i) - \frac{l}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma_i| = \\
&= -\frac{1}{2}\frac{(x-\mu_i)}{\Sigma_i}(x-\mu_i)^{\mathrm{T}} + \ln P(\Omega_i) + C_i,\ \text{где}\ C_i = -\frac{l}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma_i|
\end{aligned}$$

Эта функция представляет собой квадратичную форму. Следовательно, разделяющая поверхность $g_i(x) - g_j(x) = 0$ является гиперповерхностью второго порядка. Поэтому Байесовский классификатор является квадратичным классификатором.

Пример. Пусть $l = 2$, $\Sigma_i = \begin{pmatrix} \sigma_i^2 & 0 \\ 0 & \sigma_i^2 \end{pmatrix}$. Тогда $\frac{1}{\Sigma_i} = \begin{pmatrix} \frac{1}{\sigma_i^2} & 0 \\ 0 & \frac{1}{\sigma_i^2} \end{pmatrix}$.

$$g_i(x) = -\frac{1}{2\sigma_i^2}(x_1^2 + x_2^2) + \frac{1}{\sigma_i^2}(\mu_{i1}x_1 + \mu_{i2}x_2) - \frac{1}{2\sigma_i^2}(\mu_{i1}^2 + \mu_{i2}^2) + \ln(P(\Omega_i)) + C_i$$

Разделяющей поверхностью является коническое сечение.

Пример. Пусть $P(\Omega_1) = P(\Omega_2)$, $\mu_1 = (0,0)$, $\mu_2 = (1,0)$, $\Sigma_1 = \begin{pmatrix} 0.1 & 0 \\ 0 & 0.15 \end{pmatrix}$, $\Sigma_2 = \begin{pmatrix} 0.2 & 0 \\ 0 & 0.25 \end{pmatrix}$. Тогда $\frac{1}{\Sigma_1} = \begin{pmatrix} 10 & 0 \\ 0 & 20/3 \end{pmatrix}$, $\frac{1}{\Sigma_2} = \begin{pmatrix} 5 & 0 \\ 0 & 4 \end{pmatrix}$. Найдем поверхность решения.

$$g_1(x) = -\frac{1}{2}(x_1, x_2)\begin{pmatrix} 10 & 0 \\ 0 & 20/3 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \ln P(\Omega_1) - \ln(2\pi) + \frac{1}{2}\ln\frac{200}{3} =$$

$$= -\frac{1}{2}\left(10x_1^2 + \frac{20}{3}x_2^2\right) + \ln P(\Omega_1) - \ln(2\pi) + \frac{1}{2}\ln\frac{200}{3}$$

$$g_2(x) = -\frac{1}{2}(x_1 - 1, x_2)\begin{pmatrix} 5 & 0 \\ 0 & 4 \end{pmatrix}\begin{pmatrix} x_1 - 1 \\ x_2 \end{pmatrix} + \ln P(\Omega_2) - \ln(2\pi) + \frac{1}{2}\ln 20 =$$

$$= -\frac{1}{2}\left(5(x_1 - 1)^2 + 4x_2^2\right) + \ln P(\Omega_2) - \ln(2\pi) + \frac{1}{2}\ln 20$$

$$g_1(x) - g_2(x) = -\frac{1}{2}\left(10x_1^2 + \frac{20}{3}x_2^2 - 5(x_1 - 1)^2 - 4x_2^2\right) + \frac{1}{2}\left(\ln\frac{200}{3} - \ln 20\right) =$$

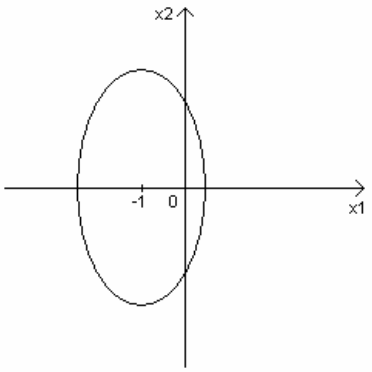
$$= -\frac{1}{2}\left(5(x_1 + 1)^2 + \frac{8}{3}x_2^2\right) + 5 + \frac{1}{2}\ln\frac{10}{3}$$

Т.к. $g_1(x) - g_2(x) = 0$, то $-\frac{1}{2}\left(5(x_1 + 1)^2 + \frac{8}{3}x_2^2\right) + 5 + \frac{1}{2}\ln\frac{10}{3} = 0$

$$5(x_1 + 1)^2 + \frac{8}{3}x_2^2 = 10 + \ln\frac{10}{3}$$

$$\frac{(x_1 + 1)^2}{8/3} + \frac{x_2^2}{5} = \frac{3}{40}\left(10 + \ln\frac{10}{3}\right)$$

$$\frac{(x_1 + 1)^2}{\left(2\sqrt{2/3}\right)^2} + \frac{x_2^2}{\left(\sqrt{5}\right)^2} = \frac{3}{40}\left(10 + \ln\frac{10}{3}\right)$$



– эллипс центром в точке $(-1,0)$.

Пример. Пусть $P(\Omega_1) = P(\Omega_2)$, $\mu_1 = (0,0)$, $\mu_2 = (1,0)$, $\Sigma_1 = \begin{pmatrix} 0.1 & 0 \\ 0 & 0.15 \end{pmatrix}$, $\Sigma_2 = \begin{pmatrix} 0.15 & 0 \\ 0 & 0.1 \end{pmatrix}$. Тогда $\frac{1}{\Sigma_1} = \begin{pmatrix} 10 & 0 \\ 0 & 20/3 \end{pmatrix}$, $\frac{1}{\Sigma_2} = \begin{pmatrix} 20/3 & 0 \\ 0 & 10 \end{pmatrix}$. Найдем поверхность решения.

Из предыдущего примера:

$$g_1(x) = -\frac{1}{2}\left(5(x_1 - 1)^2 + 4x_2^2\right) + \ln P(\Omega_2) - \ln(2\pi) + \frac{1}{2}\ln 20$$

$$g_2(x) = -\frac{1}{2}(x_1 - 1, x_2)\begin{pmatrix} 20/3 & 0 \\ 0 & 10 \end{pmatrix}\begin{pmatrix} x_1 - 1 \\ x_2 \end{pmatrix} + \ln P(\Omega_2) + \frac{1}{2}\ln\frac{200}{3}$$

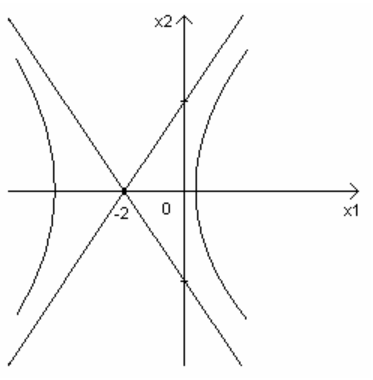
$$g_1(x) - g_2(x) = -\frac{1}{2}\left(10x_1^2 + \frac{20}{3}x_2^2 - \frac{20}{3}(x_1 - 1)^2 - 10x_2^2\right) =$$

$$= -\frac{1}{2}\left(\frac{10}{3}x_1^2 - \frac{10}{3}x_2^2 + \frac{40}{3}x_1 - \frac{20}{3}\right) =$$

$$= -\frac{1}{2}\cdot\frac{10}{3}\left(x_1^2 - x_2^2 + 4x_1 - 2\right) = -\frac{5}{3}\left((x_1+2)^2 - x_2^2 - 6\right)$$

Т.к. $g_1(x) - g_2(x) = 0$, то $-\frac{5}{3}\left((x_1+2)^2 - x_2^2 - 6\right) = 0$

$(x_1+2)^2 - x_2^2 = 6$ – гипербола с центром в точке $(-2,0)$



5.2. Линейная поверхность решения. Условие остается тем же: $p(x|\Omega_i) = \frac{1}{(2\pi)^{1/2}\cdot|\Sigma_i|^{1/2}}\cdot\exp\left(-\frac{1}{2}\frac{x-\mu_i}{\Sigma_i}(x-\mu_i)^{\mathrm{T}}\right)$, $i = 1,2,...,M$.

В предыдущем пункте мы получили квадратичную форму:

$$h_i(x) = \ln\left(p(x|\Omega_i)P(\Omega_i)\right) =$$

$$= \ln p(x|\Omega_i) + \ln P(\Omega_i) =$$

$$= -\frac{1}{2}\frac{x-\mu_i}{\Sigma_i}(x-\mu_i)^{\mathrm{T}} + \ln P(\Omega_i) + C_i, \text{ где } C_i = \ln\frac{1}{(2\pi)^{1/2}|\Sigma_i|^{1/2}}.$$

Пусть $\Sigma_i = \Sigma_j$, тогда

$$h_i(x) = -\frac{1}{2}\left[\frac{x}{\Sigma_i}x^{\mathrm{T}} - \frac{\mu_i}{\Sigma_i}x^{\mathrm{T}} - \frac{x}{\Sigma_i}\mu_i^{\mathrm{T}} + \frac{\mu_i}{\Sigma_i}\mu_i^{\mathrm{T}}\right] + \ln P(\Omega_i) + C_i =$$

$$= -\frac{1}{2}\left[\frac{x}{\Sigma_i}x^{\mathrm{T}} - 2\frac{\mu_i}{\Sigma_i}x^{\mathrm{T}} + \frac{\mu_i}{\Sigma_i}\mu_i^{\mathrm{T}}\right] + \ln P(\Omega_i) + C_i =$$

$$= -\frac{1}{2}\left[K_i(x) - 2W_i x^{\mathrm{T}} + W_i\mu_i^{\mathrm{T}}\right] + \ln P(\Omega_i) + C_i =$$

$$= -\frac{1}{2}K_i(x) + L_i(x) + C_i, \text{ где } L_i(x) = W_i x^{\mathrm{T}} + W_{i0};\ W_i = \frac{\mu_i}{\Sigma_i};$$

$$W_{i0} = \ln P(\Omega_i) - \frac{1}{2}W_i\mu_i^{\mathrm{T}}$$

При $\Sigma_i = \Sigma_j$ можно сравнивать только $L_i(x)$ и $L_j(x)$. Таким образом, при $\Sigma_i = \Sigma_j$ мы получили линейную поверхность решения.

5.2.1. Линейная поверхность решения с диагональной матрицей ковариации. Рассмотрим случай, когда матрица Σ диагональная с одинаковыми элементами: $\Sigma = \begin{pmatrix}\sigma^2 & 0\\ 0 & \sigma^2\end{pmatrix}$. Тогда $L_i(x)$ имеет вид: $L_i(x) = \frac{1}{\sigma^2}\mu_i^{\mathrm{T}}x + W_{i0}$;

$$L_{ij}(x) = L_i(x) - L_j(x) = W^{\mathrm{T}}(x - x_0) = 0,$$

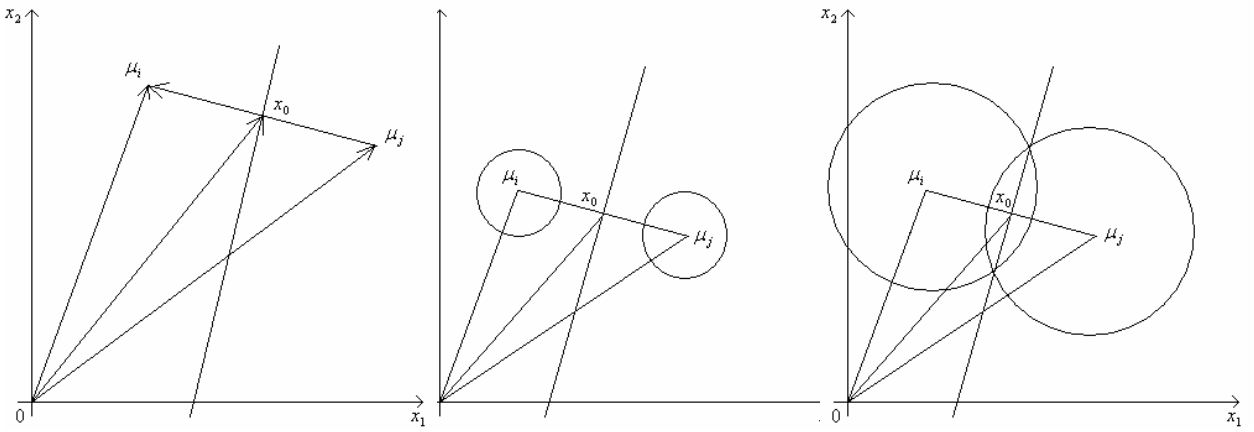
где $W = \mu_i - \mu_j$, $x_0 = \frac{1}{2}(\mu_i + \mu_j) - \sigma^2\frac{\mu_i - \mu_j}{\left\|\mu_i - \mu_j\right\|^2}\ln\frac{P(\Omega_i)}{P(\Omega_j)}$ В данном случае под нормой понимается евклидова норма. Поверхностью решения является гиперплоскость, проходящая через точку x_0.

Если $P(\Omega_i) = P(\Omega_j)$, то x_0 – это середина вектора $\overline{\mu_i\mu_j}$.

Т.к. $L_{ij}(x) = 0$, то $W^{\mathrm{T}}(x - x_0) = (\mu_i - \mu_j)^{\mathrm{T}}(x - x_0) = 0$. Следовательно, поверхность решения ортогональна $\overline{\mu_i\mu_j}$.

Пример. Рассмотрим пример разделяющей поверхности решения для двух классовой задачи с нормальным распределением. Поверхность решения лежит ближе к μ_i, если

$P(\Omega_i) < P(\Omega_j)$. Соответственно, поверхность решения лежит ближе к μ_j, если $P(\Omega_i) > P(\Omega_j)$. Также, если σ^2 мало по отношению к $\|\mu_i - \mu_j\|$, то положение поверхности решения не очень чувствительно к изменению $P(\Omega_i)$ и $P(\Omega_j)$. Последнее справедливо, т.к. вектора лежат в малых окрестностях μ_i и μ_j, поэтому изменение гиперплоскости



их затрагивает не сильно. В центре изображен случай малого, а справа случай большого σ^2.

5.2.2. Линейная поверхность решения с недиагональной матрицей ковариации. В этом случае уравнение:

$$L_{ij}(x) = L_i(x) - L_j(x) = W^{\mathrm{T}}(x - x_0) = 0$$

будет иметь несколько иные параметры:

$$W = \frac{\mu_i - \mu_j}{\Sigma} \text{ и } x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\mu_i - \mu_j}{\|\mu_i - \mu_j\|^2_{\Sigma^{-1}}}$$

В данном случае под нормой понимается так называемая Σ^{-1} норма x, которая имеет вид: $\|x\|_{\Sigma^{-1}} = \left(x^{\mathrm{T}}\Sigma^{-1}x\right)^{1/2}$. Для такой нормы поверхность решения не ортогональна вектору $\overline{\mu_i\mu_j}$, Но она ортогональна его образу при преобразовании $\Sigma^{-1}\left(\mu_i - \mu_j\right)$.

§6. Классификаторы по минимуму расстояния.

Будем рассматривать равновероятные классы с одинаковой матрицей ковариации. Тогда $\Sigma_1 = \Sigma_2 = ... = \Sigma_n = \Sigma$ и выражение

$$L_i(x) = -\frac{1}{2}(x - \mu_i)\Sigma_i^{-1}(x - \mu_i)^{\mathrm{T}} + \ln P(\Omega_i) + C_i$$

примет вид

$$L_i(x) - L_j(x) = -\frac{1}{2}(x - \mu_i)\Sigma^{-1}(x - \mu_i)^{\mathrm{T}}$$

(т.к. логарифм и константа сократятся).

6.1. Классификатор по минимуму расстояния с диагональной матрицей ковариации. Рассмотрим случай, когда матрица Σ диагональная с одинаковыми элементами: $\Sigma = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}$. Тогда максимизация $L_i(x)$ влечет минимизацию евклидового расстояния, определяемое выражением $d_E = \|x - \mu_i\|$. В данном случае будет считаться, что объект относится к данному классу, если он близок в смысле евклидового расстояния.

6.2. Классификатор по минимуму расстояния с недиагональной матрицей ковариации. В этом случае максимизация $L_i(x)$ влечет минимизацию расстояния Махалонобиса, определяемое выражением $d_M = \left((x-\mu_i)^{\mathrm{T}}\Sigma^{-1}(x-\mu_i)\right)^{1/2}$.

Т.к. матрица ковариации является симметрической, ее можно представить в виде:

$$\Sigma = \Phi \cdot \Lambda \cdot \Phi^{\mathrm{T}},$$

где $\Phi^{\mathrm{T}} = \Phi^{-1}$ и Λ в – диагональная матрица с собственными значениями матрицы Σ на диагонали. Матрица Φ имеет столбцы, соответствующие собственным векторам матрицы Σ:

$$\Phi = (\nu_1, \nu_2, ..., \nu_l)$$

Таким образом, получаем линию равноудаленных точек x:

$$(x-\mu_i)^{\mathrm{T}} \cdot \Phi \cdot \Lambda^{-1} \cdot \Phi^{\mathrm{T}} (x-\mu_i) = C^2$$

Пусть $x' = \Phi^{\mathrm{T}} x$. Тогда координатами x' являются $\nu_k^{\mathrm{T}} x$, $k = 1,2,...,l$, т.е. проекции x на собственные вектора. Другими словами, мы получили координаты в новой системе, у которой оси определяются собственными векторами $\nu_k x$, $k = 1,2,...,l$. Тогда последнее уравнение преобразуется в уравнение эллипсоида в новой системе координат:

$$\frac{\left(x_1^{/} - \mu_{i1}^{/}\right)^2}{\lambda_1} + \frac{\left(x_2^{/} - \mu_{i2}^{/}\right)^2}{\lambda_2} + ... + \frac{\left(x_l^{/} - \mu_{il}^{/}\right)^2}{\lambda_l} = C^2$$

При $l = 2$ центр эллипса находится в точке $\mu_i = (\mu_{i1}, \mu_{i2})$, а главные оси лежат по собственным векторам и имеют длины $2\sqrt{\lambda_1}C$ и $2\sqrt{\lambda_2}C$ соответственно.

Пример. Рассмотрим двумерный двух классовый случай классификации двух нормально распределенных векторов с ковариационной матрицей $\Sigma = \begin{pmatrix} 1.1 & 0.3 \\ 0.3 & 1.9 \end{pmatrix}$ и средними значениями $\mu_1 = (0,0)^{\mathrm{T}}$ и $\mu_2 = (3,3)^{\mathrm{T}}$.

Найдем Σ^{-1}:

$$|\Sigma| = 1.1 \cdot 1.9 - 0.3^2 = 2.09 - 0.09 = 2$$

$$\Sigma^{-1} = \frac{1}{2}\begin{pmatrix} 1.9 & -0.3 \\ -0.3 & 1.1 \end{pmatrix} = \begin{pmatrix} 0.95 & -0.15 \\ -0.15 & 0.55 \end{pmatrix}$$

Классифицируем вектор $(1.0,2.2)$. Для этого посчитаем расстояние Махалонобиса:

$$d_m^2(\mu_1, x) = (x-\mu_1)^{\mathrm{T}}\Sigma^{-1}(x-\mu_1) =$$

$$= (1,2.2)\begin{pmatrix} 0.95 & -0.15 \\ -0.15 & 0.55 \end{pmatrix}\begin{pmatrix} 1 \\ 2.2 \end{pmatrix} =$$

$$= (0.95 - 0.33) + (-0.15 + 1.21) \cdot 2.2 =$$

$$= 0.57 + 1 \cdot 0.6 \cdot 2.2 = 0.57 + 2.332 = 2.952$$

$$d_m^2(\mu_2, x) = (x-\mu_2)^{\mathrm{T}}\Sigma^{-1}(x-\mu_2) =$$

$$= (-1,-0.8)\begin{pmatrix} 0.95 & -0.15 \\ -0.15 & 0.55 \end{pmatrix}\begin{pmatrix} -2 \\ -0.8 \end{pmatrix} =$$

$$= -(-1.9 + 0.12) - (0.3 - 0.44) \cdot 0.8 =$$

$$= 3.56 + 0.112 = 3.672$$

Таким образом, хотя сам вектор $(1.0,2.2)$ по евклидову расстоянию близок к $(0,0)$, но по расстоянию Махалонобиса от близок к $(3,3)$.

Теперь вычислим главные оси эллипса с центром в точке $(0,0)$. Для этого найдем собственные значения:

$$\begin{vmatrix} 1.1-\lambda & 0.3 \\ 0.3 & 1.9-\lambda \end{vmatrix} = 2.09 - 3\lambda + \lambda^2 - 0.09 = \lambda^2 - 3\lambda + 2 = 0$$

$$\lambda_1 = 1,\ \lambda_2 = 2.$$

Тогда собственные вектора (и направление главных осей эллипса) будут иметь вид:

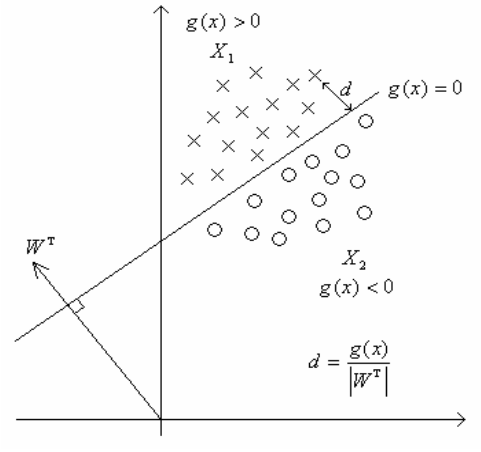
$$V_1 = \left(\frac{3}{\sqrt{10}}, \frac{-1}{\sqrt{10}}\right)^{\mathrm{T}},\ V_2 = \left(\frac{1}{\sqrt{10}}, \frac{3}{\sqrt{10}}\right).$$

§7. Гипотеза компактности.

Гипотеза компактности: объекты, близкие по свойствам, расположены компактно друг к другу.

Из гипотезы компактности следует, что для решения задачи классификации необходимо:

1) установить разделимость множеств объектов по свойствам;
2) найти разделяющую гиперплоскость.



Рассмотрим линейную дискримантную функцию: $g(x) = W^{\mathrm{T}}x + W_0$, где $W^{\mathrm{T}} = (W_1, W_2, ..., W_l)^{\mathrm{T}}$ – весовой вектор, W_0 – порог. Поведение решения задается уравнением $g(x) = 0$. Пусть X_1 и X_2 – два вектора признаков, относящихся к классу Ω_1 и Ω_2 соответственно, т.е X_1 принадлежит классу Ω_1 при $g(x) > 0$, а X_2 принадлежит классу Ω_2 при $g(x) < 0$.

Определение. *Множество, содержащее отрезок, соединяющий две произвольные внутренние точки, называется выпуклым.*

Определение. *Выпуклая оболочка – это минимальное выпуклое множество, содержащее данное.*

Утверждение. *Два множества на плоскости линейно разделимы тогда и только тогда, когда их выпуклые оболочки не пересекаются.*

Доказательство. Пусть множества X_1 и X_2 линейно разделимы. Тогда их линейные комбинации лежат по разные стороны от разделяющей их прямой $g(x) = 0$. Следовательно, их выпуклые оболочки не пересекаются.

Пусть выпуклые оболочки множеств X_1 и X_2 не пересекаются. Тогда их подмножества X_1 и X_2 не пересекаются.

ч. т. д.

Из этого утверждения вытекает алгоритм построения разделяющей прямой:

1) Построить выпуклые оболочки.
2) Проверить пересечение выпуклых оболочек. Если они пересекаются, то множества не разделимы.
3) Найти ближайшую пару элементов из каждого множества.
4) Построить срединный перпендикуляр к данной ближайшей паре, являющийся разделяющей прямой.

Рассмотрим $(W')^{\mathrm{T}} = (W^{\mathrm{T}}, W_0)$ – пополненный весовой вектор, $(X')^{\mathrm{T}} = (X^{\mathrm{T}}, 1)$ – пополненный вектор признаков. Тогда $g(x) = ((W')^{\mathrm{T}}, (X')^{\mathrm{T}})$ – дискриминантная функция в $(l+1)$-мерном пространстве.

Определение. *Множество $\overline{X} = -X$ называется симметричным множеством к множеству X.*

Утверждение. *Два замкнутых множества* X_1 *и* X_2 *разделимы тогда и только тогда, когда выпуклая оболочка множества* $X_1 \bigcup \overline{X}_2$ *не содержит начала координат.*

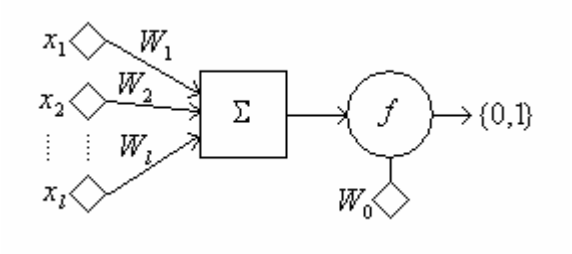
Доказательство. Пусть множества X_1 и X_2 разделимы. Тогда $g(x) > 0$ при $x \in X_1$ и $g(x) < 0$ при $x \in X_2$. Рассмотрим множество $X = X_1 \bigcup \overline{X}_2$, тогда $g(x) > 0$ при $x \in X$. Следовательно, $g(x) > 0$ для выпуклой линейной комбинации из X; а это означает, что $O \notin convX$, т.к. X – замкнутое.

Пусть $O \notin convX$, и пусть $\widetilde{x}$ а – ближайшая к началу координат O точка из $convX$. Плоскость с вектором $W = \widetilde{x}$ не пересекает $convX$, а, значит, $(W, x) > 0$ на $x \in X$. Следовательно, $(W, x) < 0$ на $x \in X_2$.

ч. т. д.

§8. Алгоритм Персептрона.

8.1. Математическая модель нейрона. В алгоритме Персептрона в основу положен принцип действия нейрона. Обобщенная схема нейрона представлена на рисунке. $x_1, x_2, ..., x_l$ – признаки; Σ – сумматор; $W_1, W_2, ..., W_l$ – синоптические веса (синопсы); f – функция активации; W_0 – порог. В нейроне на функцию активации приходит выражение



$\sum_{i=1}^{l} W_i x_i$, которое необходимо сравнивать с порогом W_0. Таким образом, дискриминантная функция имеет вид:

$$g(x) = \sum_{i=1}^{l} W_i x_i + W_0$$

Тогда задача построения линейного классификатора сводится к задаче обучения нейрона, т.к. при $g(x) > 0$ – класс Ω_1, при $g(x) < 0$ – класс Ω_2. Обучение состоит в коррекции синоптических весов и порога.

8.2. Алгоритм коррекции весов (Персептрона). Коррекция весов W_i происходит путем проверки классификации очередного прецедента x_{i+1}.

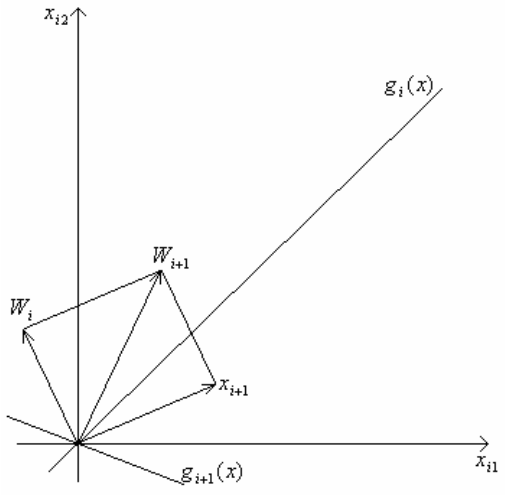
Если $x_{i+1} \in \Omega_1$ и $W_i x_{i+1} > 0$, то $W_{i+1} = W_i$.

Если $x_{i+1} \in \Omega_1$ и $W_i x_{i+1} \le 0$, то $W_{i+1} = W_i + x_{i+1}$.

Если $x_{i+1} \in \Omega_2$ и $W_i x_{i+1} < 0$, то $W_{i+1} = W_i$.

Если $x_{i+1} \in \Omega_2$ и $W_i x_{i+1} \ge 0$, то $W_{i+1} = W_i - x_{i+1}$.

Данную процедуру коррекции необходимо повторять до тех пор, пока не получится верный результат.



На данном рисунке $g_i(x)$ – дискриминантная функция после i-ого шага алгоритма Персептрона; W_i – весовой вектор после i-ого шага алгоритма Персептрона.

8.3. Сходимость алгоритма Персептрона.

Теорема Новикова. *Пусть* $\{x_i\}$ *– бесконечная последовательность векторов из двух непересекающихся замкнутых множеств* X_1 *и* X_2*; и пусть существует гиперплоскость, проходящая через начало координат и разделяющая* X_1 *и* X_2 *(не имеет с ними*

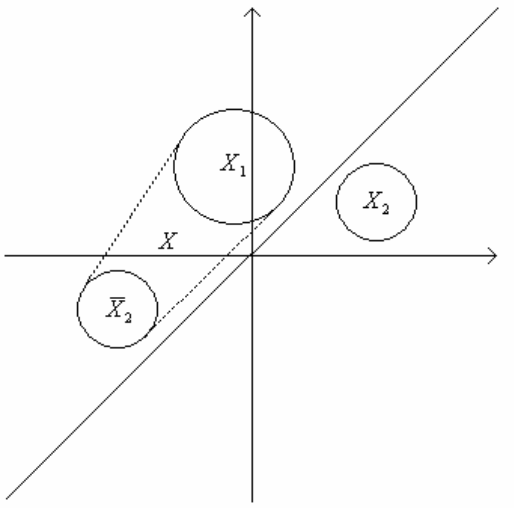
общих точек). Тогда при использовании алгоритма Персептрона число коррекций весового вектора конечно.

Доказательство. Пусть $X = conv\left(X_1 \bigcup \overline{X}_2\right)$, $\overline{X}_2$ в – симметричное к X_2 множество; $\rho_0 = \rho(0, X)$, где ρ – евклидово расстояние, $\rho_0 > 0$.

По условию $\left(W^*, X\right) \geq \rho_0 \ \ \forall x \in X$. Оценим $\left(W_i, W^*\right)$.

Пусть W^* – еденичный вектор нормали, разделяющий X_1 и X_2.

$$\left(W^*, X\right) \geq \rho_0 \text{ при } x \in X_1$$
$$\left(W^*, X\right) \leq -\rho_0 \text{ при } x \in X_2$$



Пусть W_i – весовой вектор после предъявления вектора x_i; $W_0 = 0$ – начальная итерация весового вектора ($\left|W^*\right| = 1$). Тогда, если $\left(W_i, x_{i+1}\right) > 0$, то коррекции не происходит. Иначе, если $\left(W_i, x_{i+1}\right) \leq 0$, то коррекция:

$$W_{i+1} = W_i + x_{i+1},$$
$$\left|W_{i+1}\right|^2 = \left|W_i\right|^2 + 2\left(x_{i+1}, W_i\right) + \left|x_{i+1}\right|^2 \leq \left|W_i\right|^2 + D^2,$$

т.к. $\left(x_{i+1}, W_i\right) \leq 0$ и $\left|x_{i+1}\right| \leq \sup\limits_{x \in X} |x| = D$.

Таким образом, к моменту t происходит k коррекций, то

$$\left|W_t\right|^2 \leq k \cdot D^2, \text{ т.к. } \left|W_0\right| = 0 \quad (*)$$

В начальный момент времени $\left(W_0, W^*\right) = 0$. Если в момент $i+1$ произошла коррекция, то

$$\left(W_{i+1}, W^*\right) = \left(W_0, W^*\right) + \left(x_{i+1}, W^*\right) \geq \left(W_i, W^*\right) + \rho_0$$

Если коррекция не происходит, то

$$\left(W_{i+1}, W^*\right) = \left(W_i, W^*\right)$$

Если к моменту t произошло k коррекций, то

$$\left(W_t, W^*\right) \geq k\rho_0$$

С другой стороны

$$\left(W_t, W^*\right) \leq \left|W_t\right| \cdot \left|W^*\right| = \left|W_t\right|$$

Поэтому $\left|W_t\right| \geq k\rho_0 \quad (**)$

Из неравенств $(*)$ и $(**)$ следует:

$$k^2\rho_0 \leq \left|W_t\right|^2 \leq kD^2 \ \Rightarrow \ k\rho_0 \leq D^2 \ \Rightarrow \ k \leq \frac{D^2}{\rho_0}$$

Таким образом, число коррекций k не превосходит $\left\lfloor \frac{D^2}{\rho_0} \right\rfloor$.

ч. т. д.

8.4. Оптимизационная интерпретация. Рассмотрим непрерывную кусочно-линейную функцию $J(W)$:

$$J(W) = \sum_{x \in Y} \delta_x\left(W, x\right), \text{ где } \delta_x = \begin{cases} -1, \text{при } x \in X_1 \\ 1, \text{при } x \in X_2 \end{cases};$$

Y – множество векторов неправильно классифицированных гиперплоскостью W. Тогда $J(W) \geq 0$ и $J(W) = 0 \ \Leftrightarrow \ Y = \varnothing$. Задача состоит в минимизации этой функции:

$$J(W)=\sum_{x\in Y}\delta_x(W,x)\to\min$$

Построим минимизацию по схеме градиентного спуска:

$$W_{t+1}=W_t-\rho_t\frac{dJ(W)}{dW}$$

Т.к. $\frac{dJ(W)}{dW}=\sum_{x\in Y}\delta_x x$, то $W_{t+1}=W_t-\rho_t\sum_{x\in Y}\delta_x x$

Таким образом, для сходимости алгоритма необходимо, чтобы:

$$\sum_{t=0}^{\infty}|\rho_t|>\infty \text{ и } \sum_{t=0}^{\infty}\rho_t^2<\infty$$

8.5. Схема Кеслера. Рассмотрим задачу классификации по M классам. Для каждого класса необходимо определить линейную дискримринантную функцию W_i, $i=1,2,...,M$. Пусть x – $(l+1)$-мерный вектор в расширенном пространстве. Вектор x относится к классу Ω_i, если

$$W_i x>W_j x,\ \forall i\neq j$$

Для каждого вектора-прецедента из Ω_i строим $(M-1)$ векторов x_{ij} размерности $(l+1)M$:

$$x_{ij}=(\underbrace{0,...,0}_{1},\underbrace{0,...,0}_{2},...,\underbrace{x_1,...,x_l}_{i},\underbrace{0,...,0}_{i+1},...,\underbrace{-x_1,...,-x_l}_{j},...,\underbrace{0,...,0}_{M})^{\mathrm{T}}$$

и вектор $W=(W_1,W_2,...,W_M)^{\mathrm{T}}$, где W_i – весовой вектор i-ой дискриминантной функции.

Пусть $x=(x_1,x_2,...,x_M)$, тогда вектор x_{ij} можно записать в виде:

$$x_{ij}=(\underset{1}{0},\underset{2}{0},...,\underset{i-1}{0},\underset{i}{x},\underset{i+1}{0},...,\underset{j-1}{0},\underset{j}{-x},\underset{j+1}{0},...,\underset{M}{0})$$

Если x относится к Ω_i, то $Wx_{ij}>0\ \ \forall j=1,2,...,M$, $i\neq j$, т.к.

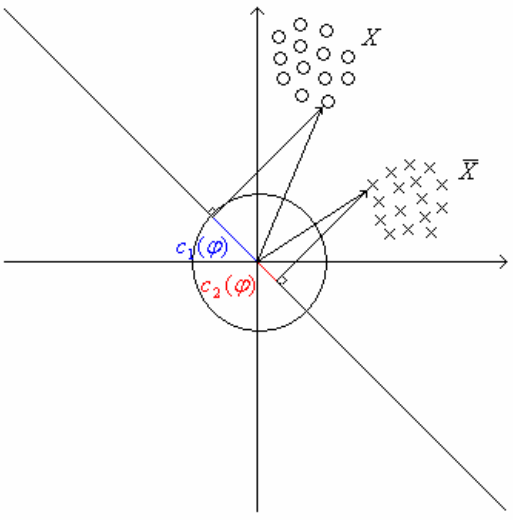
$$W_i x>W_j x \text{ и } Wx_{ij}=W_i x-W_j x>0.$$

Таким образом, задача заключается в построении линейного классификатора в $(l+1)M$-мерном пространстве так, чтобы каждый из $(M-1)N$ векторов-прецедентов лежал в положительном полупространстве. Если вектора в исходной задаче разделимы, то их можно разделить по алгоритму Персептрона.

§9. Оптимальная разделяющая гиперплоскость.

Пусть X и $\overline{X}$ - конечные множества точек в евклидовом пространстве R^l.

Определение. *X и $\overline{X}$ разделимы гиперплоскостью, если существует единичный вектор φ и число c, что $(x,\varphi)>c$ при $x\in X$, $(\bar{x},\varphi)<c$ при $\bar{x}\in\overline{X}$.*



Обозначим $c_1(\varphi)=\min_{x\in X}(x,\varphi)$, $c_2(\varphi)=\max_{\bar{x}\in\overline{X}}(\bar{x},\varphi)$. Тогда $(x,\varphi)>c_1(\varphi)$ при $x\in X$, $(\bar{x},\varphi)<c_2(\varphi)$ при $\bar{x}\in\overline{X}$. Если $c_1(\varphi)\geq c_2(\varphi)$, то гиперплоскость

$$(x,\varphi)=\frac{c_1(\varphi)+c_2(\varphi)}{2}\quad(*)$$

разделяет X и $\overline{X}$.

В силу непрерывности $c_1(\varphi)$ и $c_2(\varphi)$ существует множество разделяющих гиперпло-

скостей, если существует $(*)$.

Определение. *Оптимальной называется разделяющая гиперплоскость* $(*)$*, соответствующая вектору* φ^**, при котором достигается максимум* $\Pi(\varphi) = c_1(\varphi) - c_2(\varphi)$.

Теорема. *Если два множества* X *и* $\overline{X}$ *разделимы гиперплоскостью, то оптимальная разделяющая гиперплоскость существует и единственна.*

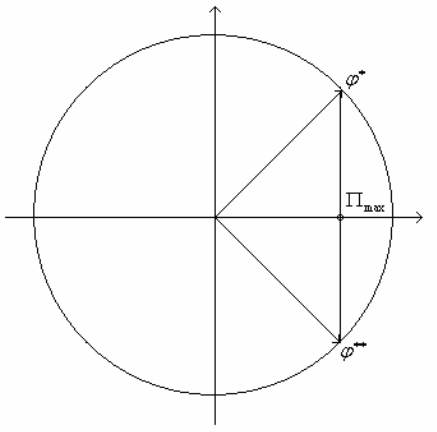
Доказательство. Функция $\Pi(\varphi)$ непрерывна на сфере $|\varphi| \le 1$. Значит, $\max\limits_{|\varphi|\le 1} \Pi(\varphi)$ существует и достигается при некотором значении φ^*. Предположим, что он достигается внутри сферы, т.е. $|\varphi^*| < 1$. Тогда для $\varphi^{**} = \dfrac{\varphi^*}{|\varphi^*|}$ получаем

$$\Pi(\varphi^{**}) = c_1(\varphi^{**}) - c_2(\varphi^{**}) =$$
$$= \min_{x\in X}(x, \varphi^{**}) - \max_{\bar{x}\in\overline{X}}(\bar{x}, \varphi^{**}) =$$
$$= \frac{1}{|\varphi^*|}\Pi(\varphi^*) > \Pi(\varphi^*),$$

что противоречит предположению о том, что φ^* - точка максимума $\Pi(\varphi)$. Следовательно, максимум достигается на границе сферы, т.е. $|\varphi^*| = 1$.

Докажем единственность максимума. Предположим, что это не так и существуют различные φ^* и φ^{**} такие, что $\Pi(\varphi^*) = \Pi(\varphi^{**}) = \Pi_{\max}$. Рассмотрим значение $\varphi = \alpha\varphi^* + \beta\varphi^{**}$, $\alpha + \beta = 1, \alpha > 0, \beta > 0$, не совпадающее ни с φ^*, ни с φ^{**}.

$$c_1(\varphi) = \min_{x\in X}(x, \alpha\varphi^* + \beta\varphi^{**}) =$$
$$= \min_{x\in X}[\alpha(x, \varphi^*) + \beta(x, \varphi^{**})] \ge$$
$$\ge \alpha\min_{x\in X}(x, \varphi^*) + \beta\min_{x\in X}(x, \varphi^{**}) =$$
$$= \alpha\cdot c_1(\varphi^*) + \beta\cdot c_1(\varphi^{**}).$$



Аналогично $c_2(\varphi) \le \alpha\cdot c_2(\varphi^*) + \beta\cdot c_2(\varphi^{**})$.
Тогда

$$\Pi(\varphi) = c_1(\varphi) - c_2(\varphi) \ge$$
$$\ge \alpha\cdot c_1(\varphi^*) + \beta\cdot c_1(\varphi^{**}) - \alpha\cdot c_2(\varphi^*) + \beta\cdot c_2(\varphi^{**}) =$$
$$= \alpha\cdot\Pi(\varphi^*) + \beta\cdot\Pi(\varphi^{**}) =$$
$$= \alpha\cdot\Pi_{\max} + \beta\cdot\Pi_{\max} = \Pi_{\max}$$

и φ - тоже значение, на котором достигается максимум.

$$|\varphi|^2 = |\alpha\varphi^* + \beta\varphi^{**}|^2 = \alpha^2|\varphi^*|^2 + 2\alpha\beta(\varphi^*, \varphi^{**}) + \beta^2|\varphi^{**}|^2 < 1,$$

т.к. $|\varphi^*|^2 = 1$, $|\varphi^{**}|^2 = 1$ и $(\varphi^*, \varphi^{**}) < 1$ при $\alpha + \beta = 1, \alpha > 0, \beta > 0$.

Но φ лежит внутри сферы $|\varphi| \le 1$ и поэтому не может быть точкой максимума. Следовательно, предположение о существовании двух максимумов неверно и максимум единственный.

ч.т.д.

Таким образом, если максимум функции $\Pi(\varphi)$ достигается при значении $\varphi = \varphi_{опт}$, то гиперплоскость $(x, \varphi_{опт}) = \dfrac{c_1(\varphi_{опт}) + c_2(\varphi_{опт})}{2}$ максимально удалена от X и $\overline{X}$ и разделяет их.

Теорема. *Если два множества* X *и* $\overline{X}$ *разделимы гиперплоскостью,* $Conv(X)$ *и* $Conv(\overline{X})$ *– выпуклые оболочки этих множеств, а* $x^* \in Conv(X)$ *и* $\overline{x}^* \in Conv(\overline{X})$ *– пара ближайших точек в выпуклых оболочках, то*

$$\max_{|\varphi|=1} \Pi(\varphi) = \left|x^* - \overline{x}^*\right|,$$

где $\left|x^* - \overline{x}^*\right|$ *– обозначает евклидово расстояние между точками* x^* *и* $\overline{x}^*$.

Доказательство. Положим $\varphi^* = \dfrac{\left(x^* - \overline{x}^*\right)}{\left|x^* - \overline{x}^*\right|}$. Из условий $c_1(\varphi) = \min_{x \in X} (x, \varphi)$, $c_2(\varphi) = \max_{\overline{x} \in \overline{X}} (\overline{x}, \varphi)$ следует, что $c_1(\varphi^*) \le (x^*, \varphi^*)$, $c_2(\varphi^*) = (\overline{x}^*, \varphi^*)$ и, следовательно,

$$\Pi(\varphi) = c_1(\varphi) - c_2(\varphi) \le (x^*, \varphi^*) - (\overline{x}^* \varphi^*) = (x^* - \overline{x}^*, \varphi^*) = \left|x^* - \overline{x}^*\right| \quad (*)$$

Следовательно $\max_{|\varphi|=1} \Pi(\varphi) \le \left|x^* - \overline{x}^*\right|$ и для доказательства теоремы нужно показать, что справедливо неравенство

$$\Pi(\varphi^*) \ge \left|x^* - \overline{x}^*\right| \quad (**)$$

Пусть точки $y \in X$ и $\overline{y} \in \overline{X}$ такие, что $c_1(\varphi^*) = (y, \varphi^*)$ и $c_2(\varphi^*) = (\overline{y}, \varphi^*)$.
Тогда

$$\begin{aligned}
\Pi(\varphi^*) &= c_1(\varphi^*) - c_2(\varphi^*) = (y - \overline{y}, \varphi^*) = \\
&= \left(x^* + (y - x^*) - \overline{x}^* - (\overline{y} - \overline{x}^*), \varphi^*\right) = \\
&= (x^* - \overline{x}^*, \varphi^*) + (y - x^*, \varphi^*) - (\overline{y} - \overline{x}^*, \varphi^*) = \\
&= \left|x^* - \overline{x}^*\right| + (y - x^*, \varphi^*) - (\overline{y} - \overline{x}^*, \varphi^*).
\end{aligned}$$

Теперь покажем, что $(y - x^*, \varphi^*) \ge 0$, а $(\overline{y} - \overline{x}^*, \varphi^*) \le 0$, или, что то же самое:

$$(y - x^*, x^* - \overline{x}^*) \ge 0, \ (\overline{y} - \overline{x}^*, x^* - \overline{x}^*) \le 0 \quad (***)$$

Пусть $z = \lambda y + (1 - \lambda) x^*$, $0 < \lambda < 1$ – точка в R^l. Очевидно, что она лежит в выпуклой оболочке X, т.е. $z \in Conv(X)$. Тогда имеем

$$\begin{aligned}
\left|z - \overline{x}^*\right|^2 &= \left|\lambda(y - \overline{x}^*) + (1 - \lambda)(x^* - \overline{x}^*)\right|^2 = \\
&= \left|\lambda(y - x^*) + (x^* - \overline{x}^*)\right|^2 = \\
&= \left|x^* - \overline{x}^*\right|^2 + 2\lambda(x^* - \overline{x}^*, y - \overline{x}^*) + \lambda^2 \left|y - x^*\right|^2 \quad (****)
\end{aligned}$$

Поскольку точки x^* и $\overline{x}^*$ – ближайшие в выпуклых оболочках $Conv(X)$ и $Conv(\overline{X})$, получаем, что $\left|z - \overline{x}^*\right|^2 \ge \left|x^* - \overline{x}^*\right|^2$.
Тогда из $(****)$ следует, что

$$2\lambda(x^* - \overline{x}^*, y - x^*) + \lambda^2 \left|y - x^*\right|^2 \ge 0,$$

или $2(x^* - \overline{x}^*, y - x^*) + \lambda \left|y - x^*\right|^2 \ge 0 \quad \forall \lambda > 0$, что возможно лишь при $(x^* - \overline{x}^*, \ y - x^*) \ge 0$. Тем самым первое из неравенств $(***)$ доказано. Второе неравенство $(***)$ доказывается аналогично. Тем самым доказано неравенство $(**)$, а из него $(*)$ и утверждение теоремы.

ч.т.д.

Оптимальная разделяющая гиперплоскость ортогональна отрезку, соединяющему ближайшие точки выпуклых оболочек множеств X и $\overline{X}$, и проходит через середину этого отрезка. Задача поиска пары ближайших точек сводится к задаче квадратичного программирования следующим образом.

Каждая точка y, лежащая в выпуклой оболочке $Conv(X)$, представима в виде $y=\sum_{x\in X}\alpha_x x$, $\sum_{x\in X}\alpha_x=1$, $\alpha_x\ge 0$. Аналогично, точка $\bar{y}\in Conv(\bar{X})$ представима в виде $\bar{y}=\sum_{\bar{x}\in\bar{X}}\beta_{\bar{x}}\bar{x}$, $\sum_{\bar{x}\in\bar{X}}\beta_{\bar{x}}=1$, $\beta_{\bar{x}}\ge 0$. Нужно найти пару точек y и $\bar{y}$, обеспечивающих минимум выражения:

$$\left|y-\bar{y}\right|^2=\left(\sum_{x\in X}\alpha_x x-\sum_{\bar{x}\in\bar{X}}\beta_{\bar{x}}\bar{x},\sum_{x\in X}\alpha_x x-\sum_{\bar{x}\in\bar{X}}\beta_{\bar{x}}\bar{x}\right) \quad (1)$$

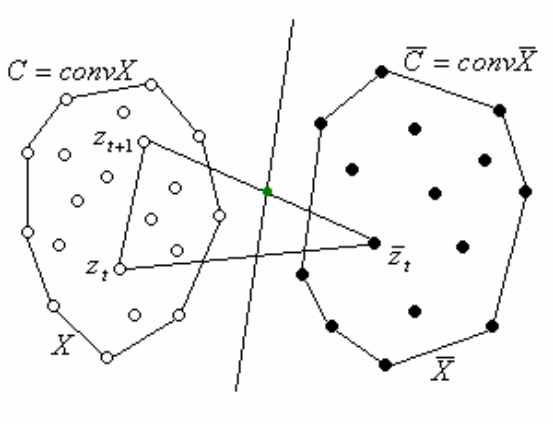
при условиях:

$$\sum_{x\in X}\alpha_x=1,\ \alpha_x\ge 0, \quad (2)$$

$$\sum_{\bar{x}\in\bar{X}}\beta_{\bar{x}}=1,\ \beta_{\bar{x}}\ge 0. \quad (3)$$

Задача математического программирования $(1)-(3)$ имеет два ограничения и квадратичную целевую функцию.

§10. Алгоритм Гаусса-Зейделя.



Задача состоит в нахождении наименьшего расстояния между множествами X и $\bar{X}$.

1. В качестве начальных значений берем произвольную пару x_0 и $\bar{x}_0$. Другими словами в начальный момент $t=0$ $z_t=x_0\in X$ и $\bar{z}_t=\bar{x}_0\in\bar{X}$.

2. Необходимо найти точку z_{t+1} ближайшую к $\bar{z}_t$ на отрезке $[z_t,x_t]$. Обозначаем $\bar{z}_{t+1}=\bar{z}_t$. Напишем условие ортогональности векторов $(z_{t+1}-\bar{z}_t)$ и (z_t-x_k):

$$(z_{t+1}-\bar{z}_t,z_t-x_k)=0.$$

Т.к. $z_{t+1}=\lambda z_t+(1-\lambda)x_k=x_k+\lambda(z_t-x_k)$, то

$$(z_{t+1}-\bar{z}_t,z_t-x_k)=(x_k+\lambda(z_t-x_k)-\bar{z}_t,z_t-x_k)=$$
$$=\lambda(z_t-x_k,z_t-x_k)+(x_k-\bar{z}_t,z_t-x_k)=0$$

Следовательно, $\lambda=\dfrac{(\bar{z}_t-x_k,z_t-x_k)}{|z_t-x_k|^2}$.

Если $\lambda\le 0$, то $z_{t+1}=x_k$. Если $\lambda\ge 1$, то $z_{t+1}=z_t$. Если $0<\lambda<1$, то $z_{t+1}=\lambda z_t+(1-\lambda)x_k$.

3. Далее необходимо найти точку $\bar{z}_{t+1}$ ближайшую к z_t на отрезке $[\bar{z}_t,x_r]$. Обозначаем $z_{t+1}=z_t$.

Данную процедуру необходимо повторять, пока не найдутся две ближайшие точки множеств X и $\bar{X}$.

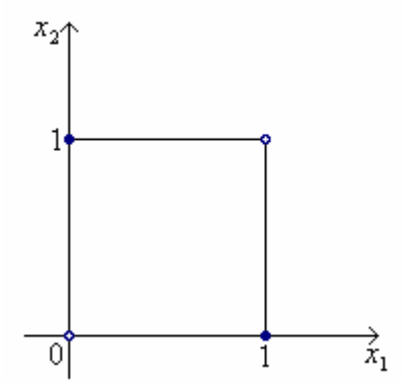
Г Л А В А 2

НЕЛИНЕЙНЫЕ КЛАССИФИКАТОРЫ

§1. Задача исключающего ИЛИ

Рассмотрим булеву функцию $xor(x_1, x_2)$ как некий классификатор. В данном случае имеется четыре прецедента и два класса. Напомним таблицу значений функции $xor(x_1, x_2)$.

№ прецедента	x_1	x_2	$xor(x_1, x_2)$	Класс
1	0	0	0	Ω_1
2	0	1	1	Ω_0
3	1	0	1	Ω_0
4	1	1	0	Ω_1



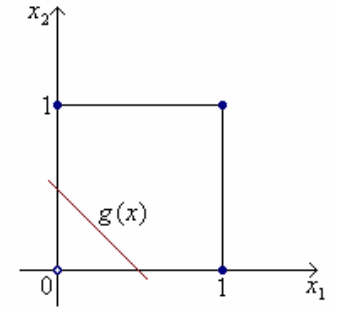
Как видно из рисунка тут нельзя построить разделяющую прямую, а, следовательно, и линейный классификатор. Попытаемся построить необходимый нелинейный классификатор через несколько линейных.

Рассмотрим две вспомогательные булевы функции $or(x_1, x_2)$ и $and(x_1, x_2)$. Напомним таблицы значений этих функций:

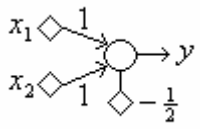
№ прецедента	x_1	x_2	$and(x_1, x_2)$	$or(x_1, x_2)$
1	0	0	0	0
2	0	1	0	1
3	1	0	0	1
4	1	1	1	1

1.1. Построение линейного классификатора функции $or(x_1, x_2)$. Очевидно, что линейный классификатор наиболее оптимально задает следующая функция:

$$x_1 + x_2 = \frac{1}{2}$$

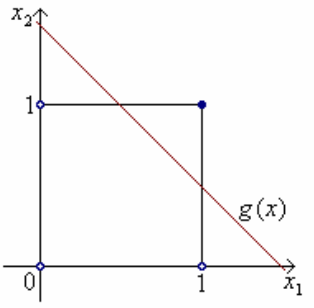


Соответствующий персептрон имеет вид:

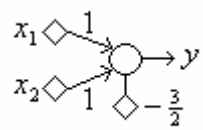


1.2. Построение линейного классификатора функции $and(x_1, x_2)$. Очевидно, что линейный классификатор наиболее оптимально задает следующая функция:

$$x_1 + x_2 = \frac{3}{2}$$



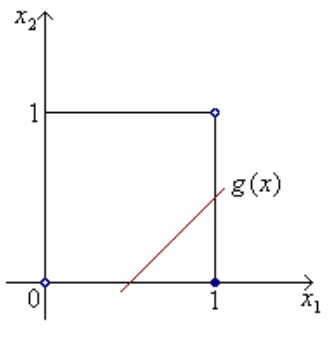
Соответствующий персептрон имеет вид:



1.3. Построение нелинейного классификатора функции $xor(x_1, x_2)$**.** Пусть на выходе персептрона для функции $or(x_1, x_2)$ – y_1, а на выходе персептрона для функции $and(x_1, x_2)$ – y_2. Посмотрим, какие значения принимает вектор (y_1, y_2).

Первый слой персептрона				Второй слой персептрона
x_1	x_2	y_1	y_2	Класс
0	0	0	0	Ω_1
0	1	1	0	Ω_0
1	0	1	0	Ω_0
1	1	1	1	Ω_1

Обозначив классы как показано в таблице, получаем разделяющую прямую изображенную на рисунке и соответствующий линейный классификатор:

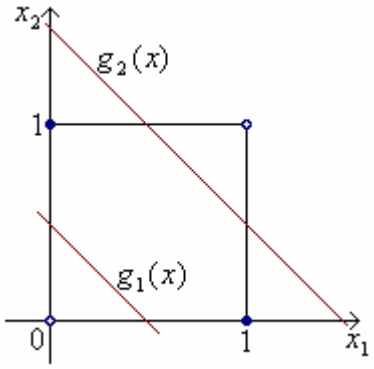


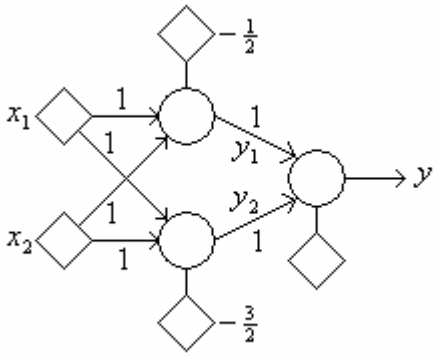
$$y_1 - y_2 = \frac{1}{2}$$

Учитывая вышеизложенное, получаем нелинейный классификатор, который задается через два линейных классификатора, как показано на рисунке слева:

$$x_1 + x_2 = \frac{1}{2} \text{ и } x_1 + x_2 = \frac{3}{2}$$

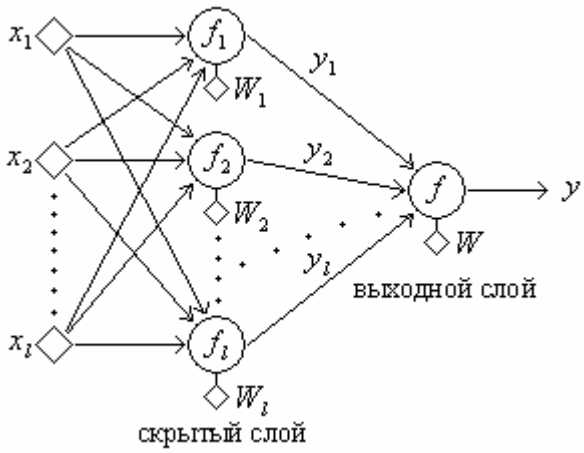
Соответствующий двухслойный персептрон изображен на рисунке справа.



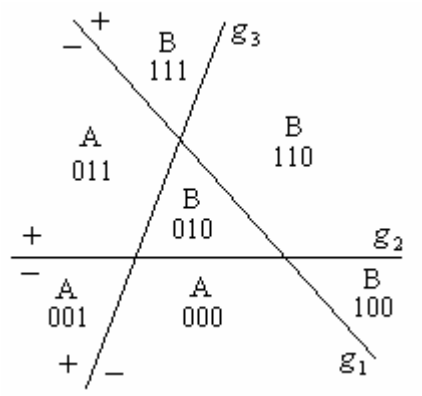


§2. Классификационные способности двухслойного персептрона.

Рассмотрим общий случай двухслойного персептрона. Пусть $x \in R^l$ и в скрытом слое p нейронов. Скрытый слой нейронов отображает R^l в $H_p \in R^p$, где $H_p = \{(y_1, y_2, ..., y_p) \in R^p, y_i \in [0,1], 1 \le i \le p\}$ – гиперкуб. Другими словами каждый нейрон задает гиперплоскость, которая разделяет пространство пополам, т.е. скрытый слой нейронов делит пространство R^l на полиэдры. Все вектора из каждого полиэдра отображаются в вершину p-мерного единичного куба. Выходной нейрон разделяет вектора в классах, описанных полиэдрами, т.е. производит сечение гиперкуба, полученного в скрытом слое.



Пример. Рассмотрим с двумя входами ($l=2$) и тремя нейронами ($k=3$). Тогда пространство $R^l = R^2$. Пусть первый слой нейронов задан разбиением как на рисунке.



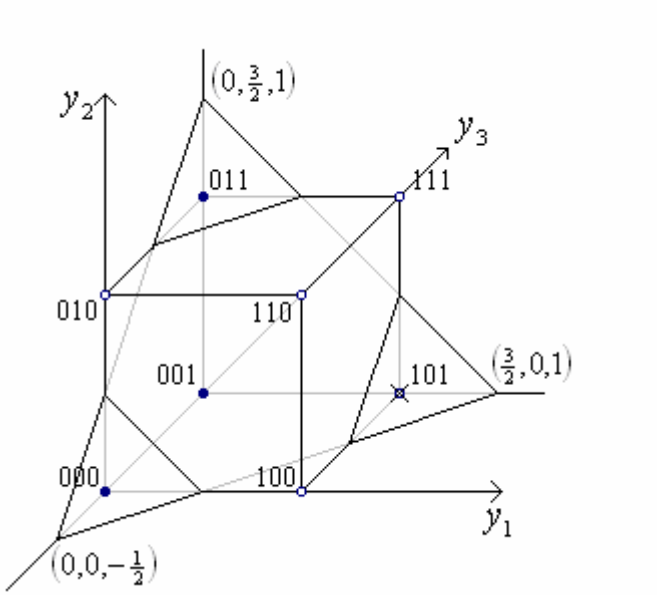
Разделим все пространство на два класса A и B. Пусть при $g(x)<0$ объект принадлежит классу A, а при $g(x)>0$ объект принадлежит классу B. В пространстве $R^k = R^3$ получим единичный куб H^3, у которого закрашенные вершины классу A, а не закрашенные – классу B.

Рассмотрим куб H^3. Построим в нем сечение, которое задается уравнением:

$$y_1 + y_2 - y_3 = \frac{1}{2}$$

Это сечение действительно разобьет куб на два класса, т.к. вершина (1,0,1) не загружена.

Определение. *Полиэдр, которому соответствует не загруженная вершина единичного гиперкуба называется виртуальным.*



Рассмотрим построение сечения p-мерного единичного куба. Диагональ куба имеет длину $\sqrt{p}$. Длины диагоналей $(p-1)$-мерных единичных кубов, являющиеся боковыми сторонами p-мерного единичного куба, равны $\sqrt{p-1}$. Центр куба находится в точке $\left(\frac{1}{2},\frac{1}{2},...,\frac{1}{2}\right)$. Расстояние от центра куба до любой вершины равно $\sqrt{p}/2$. Плоскость проводим так, чтобы расстояние от вершины, которую надо отсечь было, до секущей плоскости было равно $1-\dfrac{\sqrt{p}+\sqrt{p-1}}{2}$, причем данная точка должна находиться на диагонали куба, проведенной к отсекаемой вершине.

Пусть V – отделяемая вершина, $\overline{V}$ – диагонально противоположная вершина, следовательно, $W = V - \overline{V}$ – направляющий вектор. Тогда гиперплоскость проходит через точку:

$$U = \overline{V} + \left(V - \overline{V}\right) \cdot \frac{\sqrt{p}+\sqrt{p-1}}{2\sqrt{p}}$$

Обозначим:

$$\gamma = \frac{\sqrt{p}+\sqrt{p-1}}{2\sqrt{p}} = \frac{1}{2}\cdot\left(1+\sqrt{1-\frac{1}{p}}\right)$$

Тогда

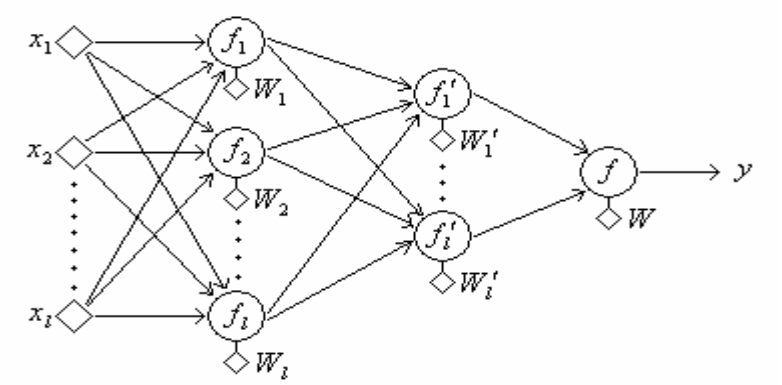
$$U = \overline{V} + \left(V - \overline{V}\right) \cdot \gamma$$

и уравнение гиперплоскости запишется в виде:

$$\left((z-U),W\right) > 0\,.$$

§3. Трехслойный персептрон.

Внешний (выходной) нейрон реализует лишь одну гиперплоскость. Поэтому двухслойная сеть не всегда может обеспечить желаемое разделение. С аналогичной ситуацией мы уже сталкивались в задаче *исключающего или*. Попробуем ввести еще один слой нейронов.



Утверждение. *Трехслойная нейронная сеть позволяет описать любые разделения объединений полиэдров.*

Доказательство. Рассмотрим первый слой нейронов. На первом формируются гиперплоскости, т.к. существует полиэдральное разбиение пространства гиперплоскостями такое, что ни в каком полиэдре не окажется пары точек из разных классов.

Во втором слое нейронов происходит сечение гиперкуба – выделение областей классификации. Каждая вершина V гиперкуба может быть отделена гиперплоскостью:

$$(t - U, W) > 0,$$

где $t = (t_1, t_2, ..., t_p)$ – точка в R^p.

Третий слой определяет классы. С каждой вершиной связан свой класс. Значит третий слой – это *задача or* для вершин входящих в один класс. Таким образом разделяющая гиперплоскость задана уравнением:

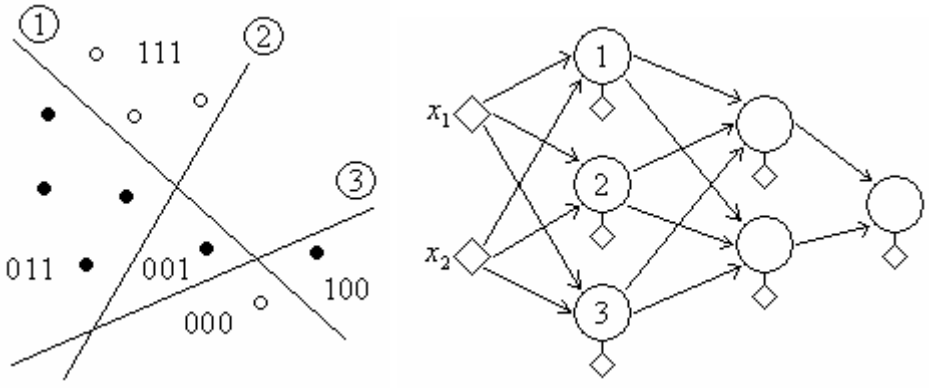
$$z_1 + z_2 + ... + z_p = \frac{1}{2}.$$

ч.т.д.

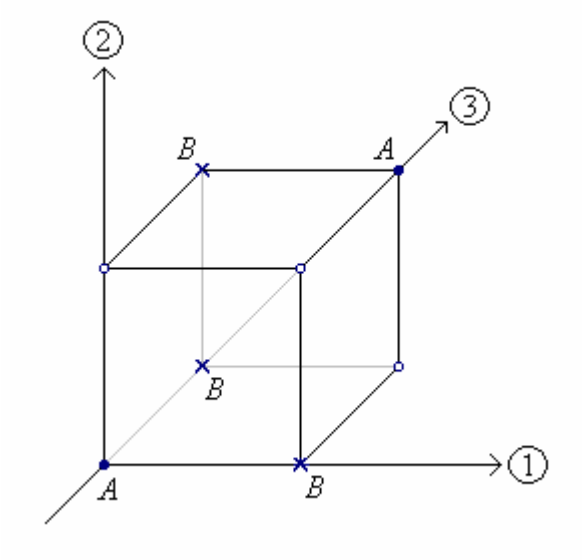
§4. Построение классификатора.

Существует два подхода к задаче построения классификатора. Первый подход заключается в построении сети, варьируя архитектуру. Данный метод основан на точной классификации прецедентов. Второй подход состоит в подборке параметров (весов и порогов) для сети с заданной архитектурой.

4.1 Алгоритм, основанные на точной классификации множества прецедентов. Опишем общую идею метода. За основу берется один нейрон. Далее наращиваем нейрон, пока не получим правильную классификацию всех прецедентов.



Рассмотрим более подробно алгоритм. Начинаем с одного нейрона $n(X)$, называемого мастером. После его тренировки получаем разделение множества X на X^+ и X^-.



Если X^+ содержит вектора из двух классов, то вводим новый узел $n(X^+)$, называемый последователем.

Таким образом, на первом слое нейронов находится один мастер и несколько последователей. Никакие вектора из разных классов не имеют одинакового выхода из первого слоя.

$$X_1 = \{y : y = f_1(x), x \in X\},$$

где f_1 – отображение, задаваемое первым слоем.

Аналогичным образом строим второй слой, третий слой и т.д.

Утверждение. *При правильном выборе весов каждый очередной слой правильно классифицирует все вектора, которые правильно классифицировал мастер и еще хотя бы один вектор.*

Таким образом, получаем архитектуру, имеющую конечное число слоев, правильно классифицирующие все прецеденты.

4.1.1. Алгоритм ближайших соседей. Нейроны первого слоя – это биссекторы, разделяющие пары. Второй слой – нейроны *and* , определяющие полиэдры. Третий слой – нейроны *or* , определяющие классы.

Основным недостатком данного метода является слишком большое количество нейронов. Уменьшить количество нейронов можно путем удаления внутренних ячеек:

$$R_i = \{x : d(x, x_i) < d(x, x_j), i \neq j\}.$$

4.2. Алгоритм, основанный на подборе весов для сети с заданной архитектурой. Идея данного метода состоит в том, чтобы ввести критерий в виде функции стоимости, которую необходимо минимизировать.

Пусть

L – число слоев в сети;

k_r – число нейронов в слое r, где $r = 1,2,...,L$;

k_L – число выходных нейронов;

$k_0 = l$ – размер входа;

$x(i) = (x_1(i), x_2(i),..., x_{k_0}(i))$ – входной вектор признаков;

$y(i) = (y_1(i), y_2(i),..., y_{k_L}(i))$ а – выходной вектор, который должен быть правильно классифицирован.

Текущем состоянии сеть при обучении дает результат $\hat{y}(i)$ не совпадающий с $y(i)$. Обозначим:

$$J = \sum_{i=1}^{N} \varepsilon(i),$$

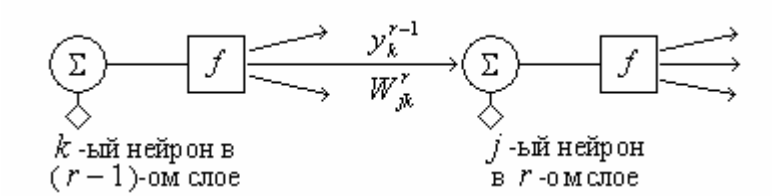
где N – число прецедентов; $\varepsilon(i)$ – ошибка на i-ом прецеденте;

$$\varepsilon(i) = \frac{1}{2}\sum_{m=1}^{k_L} e_m^2(i) = \frac{1}{2}\sum_{m=1}^{k_L} (y_m(i) - \hat{y}_m(i))^2,$$

где $i = 1,2,...,N$. J – функция всех синоптических весов и порогов. Таким образом, целью обучения является решение оптимизационной задачи:

$$J(W) \to \min,$$

где W – множество синоптических весов.

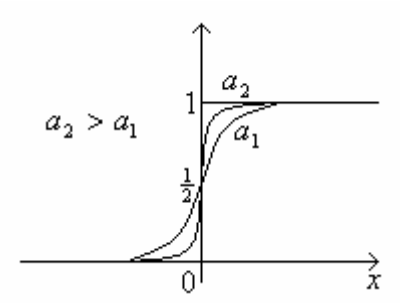


Пусть y_k^{r-1} – выход k-ого нейрона $(r-1)$-ого слоя; W_j^r – весовой вектор (включая порог) j-ого нейрона в r-ом слое, т.е. $W_j^r = \left(W_{j0}^r, W_{j1}^r, ..., W_{jk_{r-1}}^r\right)$, где k_{r-1} – число нейронов в $(r-1)$-ом слое. Таким образом, J – разрывная функция M переменных, где

$$M = \sum_{r=1}^{L} k_{r-1} k_r$$

J разрывна, т.к. разрывна функция активации f:

$$f(x) = \begin{cases} 1, \text{при } x > 0 \\ 0, \text{при } x < 0 \end{cases}$$



4.2.1 Алгоритм обратной волны. Суть – аппроксимация J непрерывной дифференцируемой функцией за счет замены функции активации “сигмовидной” функцией:

$$f(x) = \frac{1}{1+e^{-ax}}$$

Вычислим производную функции:

$$f'(x) = \frac{1}{\left(1+e^{-ax}\right)^2} \cdot ae^{-ax} = a \cdot \frac{1}{1+e^{-ax}} \cdot \frac{e^{-ax}}{1+e^{-ax}} = a \cdot \frac{1}{1+e^{-ax}} \cdot \left(1 - \frac{1}{1+e^{-ax}}\right) = a \cdot f(x) \cdot \left(1 - f(x)\right)$$

При данном чисто формальном приеме вектора признаков уже могут отображаться не только в вершины, но и внутрь гиперкуба. Необходимо решить задачу минимизации:

$$J(W) \to \min$$

4.2.1.1. Метод градиентного спуска решения задачи минимизации. Пусть $W = \left\{W_j^r; j = 1,2,...,k_r; r = 1,2,...,L\right\}$. Тогда метод градиентного спуска выглядит так:

$$\Delta W = -\mu \frac{dJ}{dW},$$

где μ а – шаг градиентного спуска. Очевидно, для его реализации необходимо уметь градиент $\frac{dJ}{dW_j^r}$.

4.2.1.2. Вычисление градиента. Аргумент функции активации j-ого нейрона r-ого слоя

$$V_j^r = \sum_{k=1}^{k_{r-1}} W_{jk}^r y_k^{r-1}(i) + W_{j0}^r = \sum_{k=0}^{k_{r-1}} W_{jk}^r y_k^{r-1}(i)$$

принимает различные значения в зависимости от индекса прецедента. В данном случае $y_0^{r-1}(i) = 1$.

Во входном слое, при $r = 1$ $y_k^{r-1}(i) = x_k(i)$, $k = 1,2,...,k_0$. В выходном слое, при $r = L$ $y_k^r(i) = \hat{y}_k(i)$, $k = 1,2,...,k_L$.

Рассмотрим выходной слой $r = L$.

$$\varepsilon(i) = \frac{1}{2}\sum_{m=1}^{k_L}\left(e_m(i)\right)^2 = \frac{1}{2}\sum_{m=1}^{k_L}\left(f\left(V_m^L(i)\right) - y_m(i)\right)^2 = \varepsilon\left(V_m^L(i)\right) = \varepsilon\left(V_m^L\left(W_m^L\right), i\right)$$

$$\frac{\partial \varepsilon(i)}{\partial W_j^L} = \frac{\partial \varepsilon(i)}{\partial V_j^L} \cdot \frac{\partial V_j^L}{\partial W_j^L}$$

$\frac{\partial V_j^L}{\partial W_j^L} = y^{r-1}(i)$ – не зависит от j-ого номера нейрона в слое, т.е. имеем одинаковый вектор производных для всех нейронов $(r-1)$-ого слоя.

$$\frac{\partial \varepsilon(i)}{\partial V_j^L} = \left(f\left(V_j^L(i)\right) - y_j(i)\right) \cdot f'\left(V_j^L(i)\right) = e_j(i) \cdot f'\left(V_j^L(i)\right)$$

Следовательно, для последнего слоя $\frac{\partial \varepsilon(i)}{\partial W_j^L} = y^{r-1}(i) \cdot e_j(i) \cdot f'\left(V_j^L(i)\right)$

Рассмотрим скрытый слой $r < L$. Имеется зависимость:

$$V_k^r = V_k^r\left(V_j^{r-1}\right)$$

$$\frac{\partial \varepsilon(i)}{\partial V_j^{r-1}(i)} = \sum_{k=1}^{k_r} \frac{\partial \varepsilon(i)}{\partial V_k^r(i)} \cdot \frac{\partial V_k^r(i)}{\partial V_j^{r-1}(i)}$$

$$\frac{\partial V_k^r(i)}{\partial V_j^{r-1}(i)} = \frac{\partial}{\partial V_j^{r-1}(i)} \left[\sum_{m=0}^{k_{r-1}} W_{km}^r y_m^{r-1}(i)\right],$$

но $y_m^{r-1}(i) = f\left(V_m^{r-1}(i)\right)$, следовательно:

$$\frac{\partial V_k^r(i)}{\partial V_j^{r-1}(i)} = W_{kj}^r \frac{\partial y_j^{r-1}(i)}{\partial V_j^{r-1}(i)} = W_{kj}^r f'\left(V_j^{r-1}(i)\right)$$

$$\frac{\partial \varepsilon(i)}{\partial V_j^{r-1}(i)} = \left[\sum_{k=1}^{k_r} \frac{\partial \varepsilon(i)}{\partial V_k^r(i)} W_{kj}^r\right] \cdot f'\left(V_j^{r-1}(i)\right)$$

Сумма, заключенная в квадратных скобках, известна из предыдущего шага.

4.2.1.3. Описание алгоритма.

0. Начальное приближение. Случайно выбираются веса небольших значений: W_{jk}^r, $r = 1,2,...,L$, $j = 1,2,...,k_r$, $k = 01,2,...,k_{r-1}$.

1. Прямой проход. Для каждого вектора прецедента $x(i)$, $i = 1,2,...,N$ вычисляются все $V_j^r(i)$, $y_j^r(i) = f\left(V_j^r(i)\right)$, $j = 1,2,...,k_r$, $r = 1,2,...,L$. Вычисляется текущее значение ценовой функции $J(W)$:

Цикл по $i = 1,2,...,N$ (по прецедентам):

Вычислить:

$y_k^0(i) = x_k(i)$, $k = 1,2,...,k_0$.

$y_0^0(i) = 1$.

Цикл по $r = 1,2,...,L$ (по слоям):

Цикл по $j = 1,2,...,k_r$ (по нейронам в слое):

$$V_j^r(i) = \sum_{k=0}^{k_{r-1}} W_{jk}^r y_k^{r-1}(i)$$

$$y_j^r(i) = f\left(V_j^r(i)\right)$$

Конец цикла по j.

Конец цикла по r.

Конец цикла по i.

$$J(W) = \sum_{i=1}^{N} \frac{1}{2}\left(y_j^L(i) - y_j(i)\right)^2$$

2. Обратный проход. Для каждого значения $i = 1,2,...,N$ и $j = 1,2,...,k_L$ вычисляется $\frac{\partial \varepsilon(i)}{\partial V_j^L(i)}$. Затем последовательно необходимо вычислить $\frac{\partial \varepsilon(i)}{\partial V_j^r(i)}$ для всех $r = L-1,...,1$ и $j = 1,2,...,k_r$:

Цикл по $i = 1,2,...,k_r$ (по нейронам в слое):

Вычислить:

$e_j(i) = y_j^L(i) - y_j(i)$

$\delta_j^L(i) = e_j(i) \cdot f'\left(V_j^{r-1}(i)\right)$

Цикл по $r = L, L-1,...,2$ (по слоям):

Цикл по $j = 1,2,...,k_r$ (по нейронам в слое):

$$e_j^{r-1}(i) = \sum_{k=1}^{k_r} \delta_k^r(i) \cdot W_{kj}^r$$

$$\delta_j^{r-1}(i) = e_j^{r-1}(i) \cdot f'\left(V_j^{r-1}(i)\right)$$

Конец цикла по j.

Конец цикла по r.

Конец цикла по i.

3. Пересчет весов. Для всех $r = 1,2,...,L$ и $j = 1,2,...,k_r$ $W_j^r(new) = W_j^r(old) + \Delta W_j^r$, где

$$\Delta W_j^r = -\mu \sum_{i=1}^{N} \frac{\partial \varepsilon(i)}{\partial V_j^r(i)} y^{r-1}(i).$$

- Останов алгоритма может происходить по двум критериям: либо $J(W)$ стала меньше порога, либо градиент стал очень мал.
- От выбора μ зависит скорость сходимости. Если μ мало, то скорость сходимости также мала. Если μ велико, то и скорость сходимости высока, но при такой скорости можно пропустить $\min$.
- В силу много экстремальности существует возможность спустить в локальный минимум. Если данный минимум по каким-то причинам не подходит, надо начинать алгоритм с другой случайной точки.
- Данный алгоритм быстрее, чем алгоритм с обучением.

Г Л А В А 3

КОМИТЕТНЫЕ МЕТОДЫ РЕШЕНИЯ ЗАДАЧ РАСПОЗНАВАНИЯ

§1. Теоретико-множественная постановка задачи выбора алгоритма.

Пусть J – индексное множество; D_j, $j \in J$ – подмножество некоторого множества (например, множества алгоритмов); $D = \{D_j \mid j \in J\}$ – система подмножеств. Пусть Y – множество, в котором необходимо найти решение. Задача заключается в нахождении такого элемента $y \in Y$ такое, что $y \in D_j \ \ \forall j \in J$.

Пример. Пусть $X_1 = \{x_1, x_2, ..., x_{m_1}\}$, $X_2 = \{x_{m_1+1}, x_{m_1+2}, ..., x_m\}$, $x_j \in \Omega$, $J = \{1,2,...,m\}$.

$$F : \Omega \to \{0,1\} \text{ так, что } F(x) = \begin{cases} 0, \text{ при } x \in X_1 \\ 1, \text{ при } x \in X_2 \end{cases}$$

Тогда D_j – множество алгоритмов, дающих правильную классификацию x_j:

$$D_j = \left\{ F \mid F : \Omega \to \{0,1\}, \quad F(x_j) = \begin{cases} 0, \text{при } 1 \le j \le m_1 \\ 1, \text{иначе} \end{cases} \right\}, \ j = 1,2,...,m$$

Определение. *Пусть* $J' \in J$, $D' = \{D_j \mid j \in J'\}$. *Тогда система подмножеств* D' *называется совместной, если* $\bigcap_{J'} D_j \neq \varnothing$.

В примере условием совместности является не пересекаемость множеств X_1 и X_2. Тогда, очевидно, что в пересечении $\bigcap_J D_j$ лежит $\Phi : \Omega \to \{0,1\}$, где

$$\Phi(x_j) = \begin{cases} 0, \text{при } 1 \le j \le m_1 \\ 1, \text{иначе} \end{cases}$$

Тогда возникает вопрос: что делать, если $D^* = \bigcap_{j \in J} D_j = \varnothing$? Существует два способа решения данной проблемы:

1) Смягчить условия, описывающие D_j, т.е. построить $\tilde{D} = \{\tilde{D}_j \mid j \in J, D_j \subseteq \tilde{D}_j\}$.
2) Решить задачу поиска максимальных совместных подсистем системы $D' = \{D_j \mid j \in J\}$, $J' \subset J$

Определение. *Теоретико-множественная задача называется разрешимой в классе* Y, *если* $Y \bigcap D^* \neq \varnothing$, *где* $D^* = \bigcap_{j \in J} D_j$.

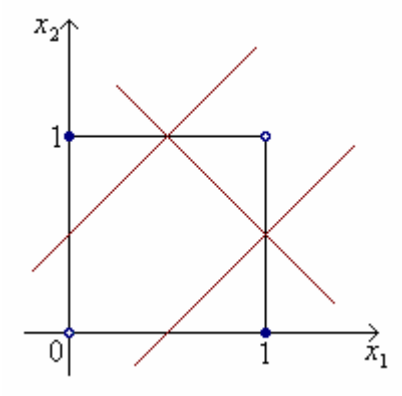
§2. Комитеты.

Нас интересует случай, когда теоретико-множественная задача не разрешима. Идея комитетного метода распознавания состоит в использовании нескольких классификаторов, каждый из которых дает свой результат. Далее по какому-либо общему правилу на основе полученных результатов от каждого классификатора выдается итоговый результат.

Определение. *Для исходной системы* D *и числа* p: $0 \le p < 1$ *конечное подмножество* $K \subseteq Y$ *называется* p*-комитетом в классе* Y, *если для всех* $j \in J$ *выполнено неравенство* $|K \bigcap D_j| > p|K|$ *(относительная доля* K, *лежащая в* D_j, *превосходит* p*). При* $p = ½$ p*-комитет называется просто комитетом.*

Пример комитета для несовместной системы. Рассмотрим задачу исключающего или. $x_0 = (0,0)$, $x_1 = (1,1)$, $x_2 = (0,1)$, $x_3 = (1,0)$. Пусть D а – множество линейных класси-

фикаторов. Опишем множество D^*: $D_0=\{F:F(x_0)=0\}$, $D_1=\{F:F(x_1)=0\}$, $D_2=\{F:F(x_2)=1\}$, $D_3=\{F:F(x_3)=1\}$, $D^*=D_0\bigcap D_1\bigcap D_2\bigcap D_3\neq\varnothing$. Пусть $Y=D$. Построим комитет $K=\{f_1,f_2,f_3\}\subset D$:



$$f_1=\left(-x_1+x_2-\frac{1}{2}>0\right)\ f_1\in D_0\bigcap D_1\bigcap D_2$$

$$f_2=\left(x_1-x_2-\frac{1}{2}>0\right)\ f_2\in D_0\bigcap D_1\bigcap D_3$$

$$f_3=\left(-x_1-x_2+\frac{3}{2}>0\right)\ f_3\in D_1\bigcap D_2\bigcap D_3$$

	z_1	z_2	Класс	f_1	f_2	f_3
x_0	0	0	B(0)	0	0	1
x_1	1	1	B(0)	0	0	0
x_2	0	1	A(1)	1	0	1
x_3	1	0	A(1)	0	1	1

$K\bigcap D_0=\{f_1,f_2\}$, $K\bigcap D_1=K$,
$K\bigcap D_2=\{f_1,f_3\}$, $K\bigcap D_3=\{f_2,f_3\}$.
$\left|K\bigcap D_j\right|\geq 2>\frac{1}{2}|K|=\frac{3}{2}$.

Следовательно, K есть комитет в классе линейных классификаторов.

Определение. *Пусть* $A,B\subseteq\Omega$ *(подмножества, возможно, бесконечные) и* $\widetilde{F}=\{F|F:\Omega\to R\}$ *– класс функционалов. Набор функционалов* $\{F_1,F_2,...,F_q\}$ *называется разделяющим комитетом для множеств* A *и* B*, если*

$$\left|\{k|F_k(a)>0\}\right|>\frac{1}{2}q,\ \forall a\in A$$

$$\left|\{k|F_k(b)<0\}\right|>\frac{1}{2}q,\ \forall b\in B$$

Утверждение. *Чтобы набор* $\{F_1,F_2,...,F_q\}$ *был разделяющим комитетом для* A *и* B *необходимо, чтобы для каждой пары* $a\in A$ *и* $b\in B$ *нашелся такой* F_k*, что* $F_k(a)>0$ *и* $F_k(b)<0$.

Доказательство. Если n_a – число функционалов $F_k(a)>0$, n_b – число функционалов $F_k(b)<0$, то

$$n_a+n_b>\frac{1}{2}q+\frac{1}{2}q=q$$

И, т.к. найдется функционал, обладающий обоими свойствами, утверждение доказано.

ч.т.д.

Теорема. *Пусть* $X=R^l$, $l\geq 2$; $A=\{x_1,x_2,...,x_{m_1}\}$, $B=\{x_{m_1+1},x_{m_2+2},...,x_m\}$, $0<m_1<m$. *И пусть* $x_k\neq 0$, $\forall k=1,2,...,m$ *(нет нулевой точки);* $x_i\neq x_j\alpha$, $\alpha\neq 0$, $\forall i,j,\alpha$ *(не коллинеарны). Тогда для таких* A *и* B *существует разделяющий комитет в классе аффинных функционалов:* $\widetilde{F}=\{F|F(x)=(W,x)+W^0,W\in R^l,W^0\in R\}$.

Доказательство. Построим комитет из $2m-1$ элементов (функционалов):

$$K=\{F_1,F_1',F_2,F_2',...,F_{m-1},F_{m-1}',F_m\}$$

Для каждого функционала необходимо найти W_k и W_k^0 – пару, которая определяет функционал $F_k=(W_k,x)+W_k^0$, причем $(x_k,W_k)=0$, т.е. $W_k\perp x_k$ и $\forall r\neq k$, $r=1,2,...,m$ $(W_k,x_r)\neq 0$, т.е. W_k не ортогонален остальным x_r. Другими словами каждая гиперпло-

скость должна иметь направляющий вектор, ортогональный своему прецеденту и не ортогональный всем остальным.

Пусть $\delta_k = \frac{1}{2}\min_{r\neq k}\left|(W_k, x_r)\right| > 0$. Выберем W_k^0 следующим образом:

$$W_k^0 = \begin{cases} \delta_k, \text{ при } k = 1,2,...,m_1 \\ -\delta_k, \text{ при } k = m_1+1,...,m \end{cases}$$

$$F_k'(x) = -(W_k, x) + W_k^0$$

$$F_k(x) = (W_k, x) + W_k^0$$

Покажем, что построенное множество функционалов является комитетом для A и B. Рассмотрим

$$F_k(x_k) = (W_k, x_k) + W_k^0 = W_k^0 = \begin{cases} > 0, \text{ при } k \le m_1 \\ < 0, \text{ при } k > m_1 \end{cases}$$

$$F_k'(x_k) = -(W_k, x_k) + W_k^0 = W_k^0 = \begin{cases} > 0, \text{ при } k \le m_1 \\ < 0, \text{ при } k > m_1 \end{cases}$$

$F_k'(x)$ и $F_k(x)$ правильно классифицируют x_k. Посмотрим, как будет работать каждый такой функционал на остальных x_r:

$$F_k(x_r) = (W_k, x_k) + W_k^0$$

Т.к. $W_k^0 < (W_k, x_k)$, то знак $F_k(x_r)$ определяется знаком (W_k, x_r).

Рассмотрим $1 \le k \le m-1$. $F_k'(x_k)$ и $F_k(x_k)$ голосуют правильно, т.е. x_k соответствует правильное положение гиперплоскостей. $F_k'(x_r)$ и $F_k(x_r)$ имеют разные знаки. Следовательно, каждая пара F_k' и F_k правильно классифицирует на всех x_k и дает одну правильную классификацию на остальных x_r. Таким образом, количество правильно голосующих за x_k равно $2 + (m-2) = m$.

ч.т.д.

§3. Комитеты линейных функционалов.

Пусть $A = \{x_1, x_2, ..., x_{m_1}\}$, $B = \{x_{m_1+1}, x_{m_2+2}, ..., x_m\}$, $A, B \subseteq R^l$ – конечные множества в пространстве признаков; $x_1, x_2, ..., x_m$ – точки общего положения.

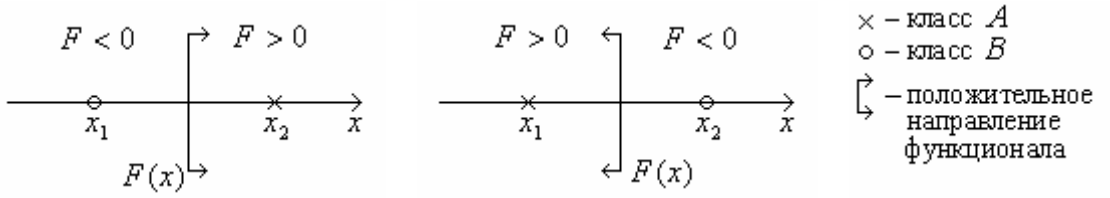
Определение. *Точки $x_1, x_2, ..., x_m$ пространства R^l называются точками общего положения, если никакая $l+1$ точка не лежит в гиперплоскости размерности $l-1$.*

Пример. Пусть $l = 2$, т.е. рассматривается пространство R^2 (плоскость). Тогда точки $x_1, x_2, ..., x_m$ – точки общего положения, если никакие три из них не лежат на одной прямой.

Теорема. *Существует разделяющий комитет аффинных функционалов, состоящий из не более, чем m членов при нечетном m и не более, чем $m-1$ при четном m.*

Доказательство. Рассмотрим случай $l = 1$, т.е. пространство R^1.

Пусть $m = 2$, $m_1 = 1$. Тогда возможны два случая.



Для первого случая (рис. слева) функционал имеет вид:

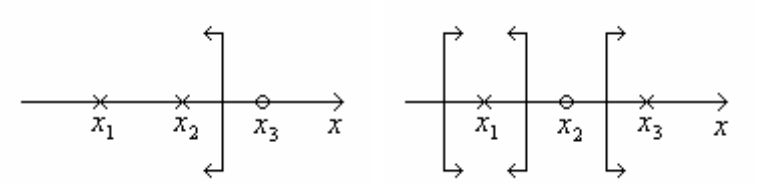
$$F(x) = x - \frac{x_1 + x_2}{2}$$

Для второго случая (рис. справа) функционал имеет вид:

$$F(x) = -\left(x - \frac{x_1 + x_2}{2} \right)$$

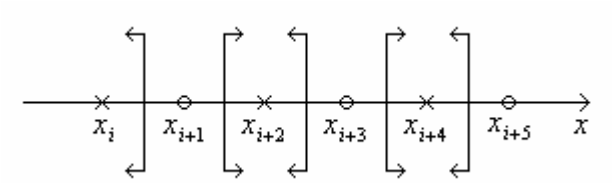
$|k| = 1$ – количество функционалов для худшего случая.

Пусть $m = 3$, $m_1 = 2$. Тогда возможны следующие варианты.



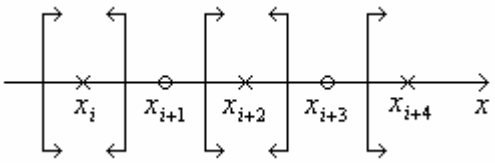
Все случаи вида показанного на рис. слева сводятся к предыдущему ($m = 2$, $m_1 = 1$). Во всех остальных случаях функционалы надо располагать аналогично рис. справа. Для худшего случая $|k| = 3$.

Пусть $m = 2n$ (четное количество точек). Рассмотрим худший из возможных вариантов.



В данном случае функционалы надо располагать, как показано на рис. $|k| = m - 1$.

Пусть $m = 2n - 1$ (нечетное количество точек). Рассмотрим худший из возможных вариантов.



В данном случае функционалы надо располагать, как показано на рис. $|k| = m$. Все остальные случаи можно свести либо к этим двум, либо к предыдущим.

Таким образом, по методу математической индукции существует разделяющий комитет аффинных функционалов из не более, чем m членов при нечетном m и не более, чем $m - 1$ при четном m в пространстве R^1.

Многомерный случай сводится к одномерному следующим образом. Ищем подпространство $W \in R^l$ такое, что $(W, x_i) \neq (W, x_j)$, при $i \neq j$. Проектируем все x_i на соответствующие подпространства, пока не получим одномерную задачу. В многомерном случае для разделения x_i и x_j служит гиперплоскость:

$$(W, x) = \frac{1}{2}\left[(W, x_i) + (W, x_j) \right]$$

ч.т.д.

§4. Функция Шеннона.

Пусть $L_n(m_1, m - m_1)$ – это число гиперплоскостей, достаточное для разделения любых точечных множеств m_1 и $m - m_1$ точек общего положения в пространстве R^n.

Лемма 1. *Если* $m_1 \le m - m_1$, *то*

$$L_n(m_1, m-m_1) \le 2\left\lceil \frac{m_1}{n} \right\rceil$$

Доказательство. Если $m_1 \le n$, то добавим точки общего положения до n. Через n точек из m_1 проводим гиперплоскость:

$$F(x_1) = F(x_2) = ... = F(x_n) = 0$$

Для x_k такого, что $k > n$ $F(x_k) \neq 0$. Выберем $\varepsilon = \frac{1}{2}\min_{n<i\le m_1}|F(x_i)|$ и возьмем гиперплоскости $G_1 = F + \varepsilon$ и $G_2 = F - \varepsilon$. G_1 и G_2 отделяют точки $x_1, x_2, ..., x_n$ от всех остальных.

Аналогичным образом из оставшихся $(m_1 - n)$ точек выделяем еще n и строим еще пару гиперплоскостей. Далее из оставшихся $(m_1 - 2n)$ точек выделяем еще n и строим еще пару гиперплоскостей и т.д. В конце получим $(m_1 - nm)$ точек. Следовательно:

$$L_n(m_1, m-m_1) \le 2\left\lceil \frac{m_1}{n} \right\rceil$$

ч.т.д.

Утверждение 1. *Если* $W_1, W_2, ..., W_q$ *разделяют множества* A *и* B, *и* $r(t)$ *– непрерывная кривая в* R^l *такая, что* $r(0) \in A$, *а* $r(1) \in B$, *то существует* $k \in \{1,2,...,q\}$ *и* $t_0 \in (0,1)$ *такие, что* $(W_{k_0}, r(t_0)) = 0$.

Утверждение 2. *Любая гиперплоскость пересекает кривую* $r(t)$ *не более, чем в* n *точках.*

Доказательство. Рассмотрим линейный функционал W. Запишем условие пересечения гиперплоскости и кривой $r(t)$:

$$(W, r(t)) = 0.$$

Кривая $r(t)$ задана многочленом степени n. Следовательно, $(W, r(t))$ – то же многочлен степени n. Значит, уравнение $(W, r(t)) = 0$ является уравнением степени n. Следовательно, т.к. корни могут быть кратными, данное уравнение имеет не более n корней.

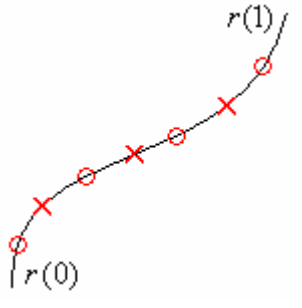
ч.т.д.

Лемма 2. $L_n(m_1, m-m_1) \ge \left\lceil \frac{2m_1 - 1}{n} \right\rceil$.

Доказательство. Построим $L_n(m_1, m-m_1)$. Рассмотрим последовательность точек:

$$0 < t_1 < t_2 < ... < t_m = 1.$$

Пусть $r(t) = (r_1, r_2, ..., r_n)$, где $r_i = r_i(t) = t^i$, $i = 1,2,...,n$. Тогда $x_j = r(t_j) = (t_j, t_j^2, t_j^3, ..., t_j^n)$ – точки в R^n.



Без ограничения общности положим $x_j \in A$, при j нечетном, и $x_j \in B$, при j четном. Тогда получим непрерывную кривую (см. рис).

Каждая гиперплоскость дает не более, чем n пересечений. Кривая должна иметь $(m-1)$ разделение, т.е. должно быть $(m-1)$ гиперплоскостей. Следовательно, всего гиперплоскостей должно быть не менее, чем $\left\lceil \frac{m-1}{n} \right\rceil$, т.е. $L_n(m_1, m-m_1) \ge \left\lceil \frac{m-1}{n} \right\rceil$

Т.к. $m_1 = \begin{cases} m/2, \text{ при четном } m \\ (m-1)/2, \text{ при нечетном } m \end{cases}$, то $\begin{matrix} m = 2m_1, \text{ при четном } m \\ m = 2m_1 + 1, \text{ при нечетном } m \end{matrix}$.

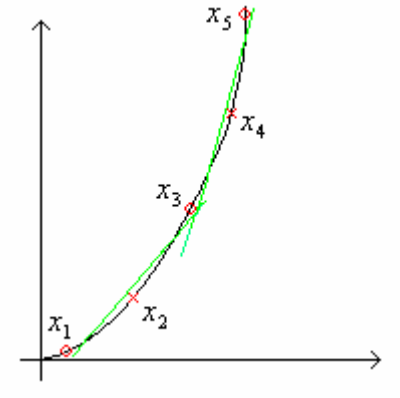
Следовательно,

$$L_n(m_1, m-m_1) \geq \left\lceil \frac{2m_1}{n} \right\rceil, \text{ при нечетном } m,$$

$$L_n(m_1, m-m_1) \geq \left\lceil \frac{2m_1-1}{n} \right\rceil, \text{ при четном } m.$$

Окончательно получаем: $L_n(m_1, m-m_1) \geq \left\lceil \frac{2m_1-1}{n} \right\rceil$, $\forall m$.

ч.т.д.



Пример. Пусть $m=5$, $m_1=2$, $n=2$. Обозначим $A=\{x_1, x_3, x_5\}$ и $B=\{x_2, x_4\}$. Тогда

$$L_n(m_1, m-m_1) = L_2(2,3) \geq \left\lceil \frac{2\cdot 2-1}{2} \right\rceil = 2 \text{ и}$$

$$L_n(m_1, m-m_1) = L_2(2,3) \leq 2\cdot\left\lceil \frac{2}{2} \right\rceil = 2$$

§5. Метод построения комитета.

Пусть X – множество прецедентов; l в – размерность пространства признаков; m_1 и $m-m_1$ – количество прецедентов в каждом классе.

Построим $W(x)$ – линейный функционал такой, что, если $W(x_k)>0$, то объект из класса A ($k=1,2,...,m_1$), и, если $W(x_k)<0$, то объект из класса B ($k=m_1+1, m_2+2, ..., m$). Если данный функционал правильно классифицирует меньше половины объектов, то возьмем его со знаком минус.

Итак, пусть линейный функционал $W(x)$ правильно классифицирует больше половины объектов. Разобьем множество прецедентов X на множество правильно классифицированных объектов X_1 и множество неправильно классифицированных объектов $\overline{X}_1$, т.е. $X = X_1 \bigcup \overline{X}_1$.

Далее строим последовательно пары функционалов W_s и W_s':

$$W_1, W_2, W_2', W_3, W_3', ..., W_s, W_s'$$

Делаем очередной шаг. $X = X_s \bigcup \overline{X}_s$. Пусть на X_s – (s) правильно классифицированных объектов, а на $\overline{X}_s$ – $(s-1)$ правильно классифицированных объектов. Строим пару W_{s+1}, W_{s+1}'. В $\overline{X}_s$ выделяем l точек одного класса. Эти точки можно перевести в X_{s+1}, т.е. $X_{s+1} = X_s + \{l \text{ точек}\}$, а $\overline{X}_{s+1} = \overline{X}_s$.

На каждом шаге множество неправильно классифицированных объектов уменьшается на l, следовательно, процесс сходится.

Общее число функционалов: $1 + 2\cdot\left\lceil \frac{m}{2}\cdot\frac{1}{l} \right\rceil = 1 + \left\lceil \frac{m}{l} \right\rceil$.

Теорема. *Существует комитет линейных функционалов, в котором число членов не превосходит* $\left\lceil \frac{m}{l}+1 \right\rceil$.