Лекция 10

Коллективные метод, бэггинг, бустинг, голосование по системам закономерностей

Лектор – *Сенько Олег Валентинович*

Курс «Математические основы теории прогнозирования»
4-й курс, III поток

Содержание лекции

Коллективные методы ("бэггинг")

Одним из способов получения ансамбля является использование алгоритмов, обученных по разным обучающим выборкам, которые возникают в результате случайного процесса, лежащего в основе исследуемой задачи. Обычно при решении прикладной задачи в распоряжении исследователя имеется обучающая выборка $\widetilde{S}_t = \{s_1, \ldots, s_m\}$ ограниченного объёма. Однако процесс генерации семейства выборок из генеральной совокупности может быть имитирован с помощью процедуры бутстрэп (bootstrap), которая основана на выборках с возвращениями из $\widetilde{S}_t$. В результате получаются выборки $\widetilde{S}_*^{bg}$, включающие объекты из обучающей выборки $\widetilde{S}_t$. Однако некоторые объекты $\widetilde{S}_t$ могут встречаться в $\widetilde{S}_*^{bg}$ более одного раза, а другие объекты отсутствовать. Предположим, что с помощью процедуры бутстрэп получено T выборок. С помощью заранее выбранного метода, используемого для обучения отдельных алгоритмов распознавания, получим множество, включающее T распознающих алгоритмов $\widetilde{A}_{bg} = \{A_1^{bg}, \ldots, A_T^{bg}\}$.

Коллективные методы "бэггинг"

Для получения коллективного решения может быть использован простейший комитетный метод, относящий объект в тот класс, куда его отнесло большинство алгоритмов . Данная процедура носит название бэггинг (bagging), что является сокращением названия Bootstrap Aggregating. Процедура бэггинг показывает высокий прирост обобщающей способности по сравнению с алгоритмом, обученным с помощью базового метода по исходной обучающей выборке $\widetilde{S}_t$, в тех случаях, когда вариационная составляющая ошибки базового метода высока. К таким моделям относятся в частности решающие деревья и нейросетевые методы. При использовании в качестве базового метода решающих деревьев процедура бэггинг приводит к построению ансамблей решающих деревьев (решающих лесов).
Основной идеей алгоритма бустинг является пошаговое наращивание ансамбля алгоритмов. Алгоритм, который присоединяется к ансамблю на шаге k обучается по выборке, которая формируется из объектов исходной обучающей выборки $\widetilde{S}_t$.

Коллективные методы "бустинг"

В отличие от метода бэггинг объекты выбираются не равноправно, а исходя из некоторого вероятностного распределения, заданного на выборке $\widetilde{S}_t$. Данное распределение вычисляется по результатам классификации с помощью ансамбля, полученного на предыдущем шаге. Приведём схему одного из наиболее популярных вариантов метода бустинг AdaBoost (Adaptive boosting) более подробно. На первом шаге присваиваем начальные значения весов $(w_1^1, \ldots, w_m^1)$ объектам обучающей выборки. Поскольку веса имеют вероятностную интерпретаци, то для них соблюдаются ограничения $\sum_{j=1}^m w_j^1 = 1$, $w_j^1 \in [0, 1]$. Обычно начальное распределение выбирается равномерным $w_j^1 = \frac{1}{m}$, $j = 1, \ldots, m$. Выбираем число итераций T. На итерации k генерируем выборку $\widetilde{S}_k^{bs}$ из исходной выборки $\widetilde{S}_t$ согласно распределению задаваемому весами $(w_1^k, \ldots, w_m^k)$. Обучаем распознающий алгоритм A_k^{bs} по выборке $\widetilde{S}_k^{bs}$.

Вычисляем взвешенную ошибку по формуле $\varepsilon_k = \sum_{j=1}^{m} w_j^k e_j^k$, где $e_j^k = 1$, если алгоритм A_k^{bs} неправильно классифицировал объект s_j и $e_j^k = 0$ в противном случае. В том случае, если $\varepsilon_k \geq 0.5$ или $\varepsilon_k = 0$ игнорируем шаг и заново генерируем выборку $\widetilde{S}_k^{bs}$ исходя из весовых коэффициентов $w_j^1 = \frac{1}{m}$, $j = 1, \ldots, m$. В случае если $\varepsilon_k \in (0, 0.5)$ вычисляем коэффициенты $\tau_k = \frac{\varepsilon_k}{1-\varepsilon_k}$
и пересчитываем веса объектов по формуле

$$w_j^{k+1} = \frac{w_j^k (\tau_k)^{1-e_j^k}}{\sum_{j=1}^{m} w_j^k (\tau_k)^{1-e_j^k}} \tag{1}$$

при $j = 1, \ldots, m$.
Процесс, задаваемый формулой (1), продолжается до тех пор, пока не выполнено T итераций . В результате мы получаем совокупность из T распознающих алгоритмов $A_1^{bs}, \ldots, A_T^{bs}$.

Предположим, что нам требуется распознать объект s^*. Пусть $\beta_l^k(s^*) = 1$, если s^* отнесён алгоритмом A_k^{bs} в класс K_l, и $\beta_l^k(s^*) = 0$ в противном случае. Оценка объекта s^* за класс K_l вычисляется по формуле

$$\Gamma_l(s^*) = \sum_{k=1}^{T} \ln \frac{1}{\tau_k} \beta_l^k(s^*).$$

Объект s^* будет отнесён к классу, оценка за которой максимальна. Описанный вариант метода носит название AdaBoost. M1. Эффективность процедур бустинга подтверждается многочисленными экспериментами на реальных данных. В настоящее время существует большое количество вариантов метода, имеющих разное обоснование.

Коллективные методы, основанные на голосовании по системам закономерностей

Одним из эффективных подходов к решению задач прогнозирования и распознавания является использование коллективных решений по системам закономерностей. Под закономерностью понимается распознающий или прогностический алгоритм, определённый на некоторой подобласти признакового пространства или связанный с некоторым подмножеством признаков. В качестве примера закономерностей могут быть приведены представительные наборы, являющиеся по сути подмножествами признаковых описаний, характерных для одного из распознаваемых классов. Аналогом представительный наборов в задач с веществennозначной информацией являются логические закономерности классов. Под логической закономерностью класса K_l понимается область признакового пространства, имеющая форму гиперпараллелепипеда и содержащая только объекты из K_l.

Логическая закономерность

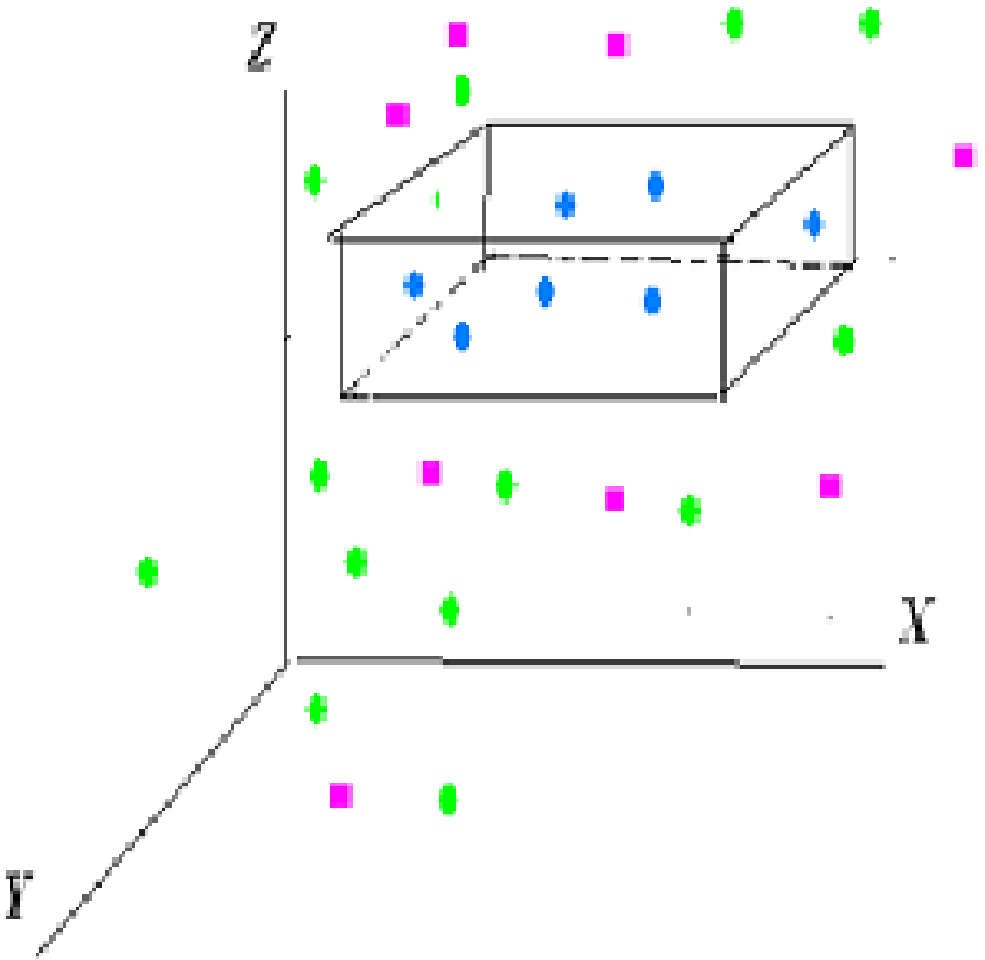

Рис 1. Пример логической закономерности.

Логическая закономерность

Логическая закономерность класса K_j, которую обозначим $\boldsymbol{r}(j)$, описывается с помощью предикатов вида

$$Pt_i^{\boldsymbol{r(j)}} = \text{”}a_i^{\boldsymbol{r}(j)} \leq X_i \leq b_i^{\boldsymbol{r}(j)} \tag{2}$$

где $i = 1, \ldots, n$. Отметим, что для несущественных для закономерности $\boldsymbol{r}(j)$ признаков отрезок $[a_i^{\boldsymbol{r}(j)}, b_i^{\boldsymbol{r}(j)}]$ соответствует области допустимых значений X_i. Для существенных признаков отрезок $[a_i^{\boldsymbol{r}(j)}, b_i^{\boldsymbol{r}(j)}]$ является некоторым подмножеством области допустимых значений X_i. Полностью $\boldsymbol{r}(j)$ задаётся конъюнкцией предикатов (2):

$$\mathrm{Pt}^{\boldsymbol{r}(j)} = Pt_i^{\boldsymbol{r(j)}} \& \ldots \& Pt_n^{\boldsymbol{r(j)}}.$$

Логическая закономерность

Для конъюнкции $\mathrm{Pt}^{\boldsymbol{r}(j)}$ должны выполняться следующие условия:

- в обучающей выборке $\widetilde{S}_t$ должен существовать объект s^* из класса K_j, для которого $\mathrm{Pt}^{\boldsymbol{r}(j)} = 1$;
- в обучающей выборке $\widetilde{S}_t$ не должно содержаться объектов, не принадлежащих класса K_j, для которых $\mathrm{Pt}^{\boldsymbol{r}(j)} = 1$;
- $\mathrm{Pt}^{\boldsymbol{r}(j)}$ доставляет экстремум некоторому функционалу качества $\Phi(\mathrm{Pt})$, заданному на множестве всевозможных предикатов, удовлетворяющих условиям 1), 2)

На практике используются следующие функционалы качества:

- число объектов из класса K_j в обучающей выборке, для которых $\mathrm{Pt}^{\boldsymbol{r}(j)} = 1$;
- доля объектов из класса K_j в обучающей выборке, для которых $\mathrm{Pt}^{\boldsymbol{r}(j)} = 1$;

Логическая закономерность

Наряду с полными логическими закономерностями, для которых выполняются все условия 1) – 3), используются также частичные логические закономерности, для которых допускаются некоторые нарушения условия 2). То есть допускается существование небольшой доли нарушений условия 2) для тех объектов, для которых выполняется условие $\mathrm{Pt}^{\boldsymbol{r}(j)} = 1$. На этапе обучения для каждого из классов K_j ищется множество логических закономерностей $\widetilde{R}_j$. Предположим, что нам требуется распознать новый объект s^* . Для каждого из классов K_j ищется число закономерностей из $\widetilde{R}_j$ для которых $\mathrm{Pt}^{\boldsymbol{r}(j)}(s^*) = 1$. При этом доля таких закономерностей считается оценкой за класс K_j . Для классификации используется стандартное решающее правило, т.е. объект относится в класс, оценка за который максимальна. Поиск оптимальной системы логических закономерностей производится по набору $\widetilde{S}_e$ случайно выбранных из обучающей выборки $\widetilde{S}_t$ эталонных объектов (опорных эталонов).

Логическая закономерность

Закономерности для класса K_j строится по каждому из опорных эталонов $s_i \in \widetilde{S}_e$. При этом поиск оптимальных границ

$$[a_1^{\boldsymbol{r}(j)}, b_1^{\boldsymbol{r}(j)}, \ldots, a_n^{\boldsymbol{r}(j)}, b_n^{\boldsymbol{r}(j)}]$$

для закономерности $\boldsymbol{r}(j)$ осуществляется сначала на некоторой неравномерной сетке пространства, которая задается с помощью разбиения интервала значений каждого из признаков. После нахождения оптимальных границ на заданной сетке, поиск продолжается на заданной в окрестности этого оптимального решения, но уже на более мелкой сетке. Процесс заканчивается, если при переходе к более мелкой сетке не удается найти логическую закономерность с более высоким критерия качества Φ. . Задача поиска оптимальной логической закономерности на каждой сетке сводится к поиску максимальной совместной подсистемы некоторой системы неравенств. Логические закономерности, построенные для случайно выбранных «опорных» эталонов класса K_j объединяются в одно множество $\widetilde{R}_j$.

Статистически взвешенные синдромы

Коллективные решения в методе СВС принимается по информации о принадлежности векторного описания распознаваемого объекта так называемым “синдромам” из некоторого множества $\widetilde{Q}$. Под "синдромом"понимается такая область признакового пространства, в которой содержание объектов одного из классов, отличается от содержания объектов этого класса в обучающей выборке или по крайней мере в одной из соседних областях. Пример синдромов, характеризующих разделение объектов из классов K_1 (+) и K_2 (0) приведён на рисунке 2. Синдромы ищутся для каждого из распознаваемых классов с помощью построения оптимальных разбиений интервалов допустимых значений единичных признаков или совместных двумерных областей допустимых значений пар признаков.

Статистически взвешенные синдромы

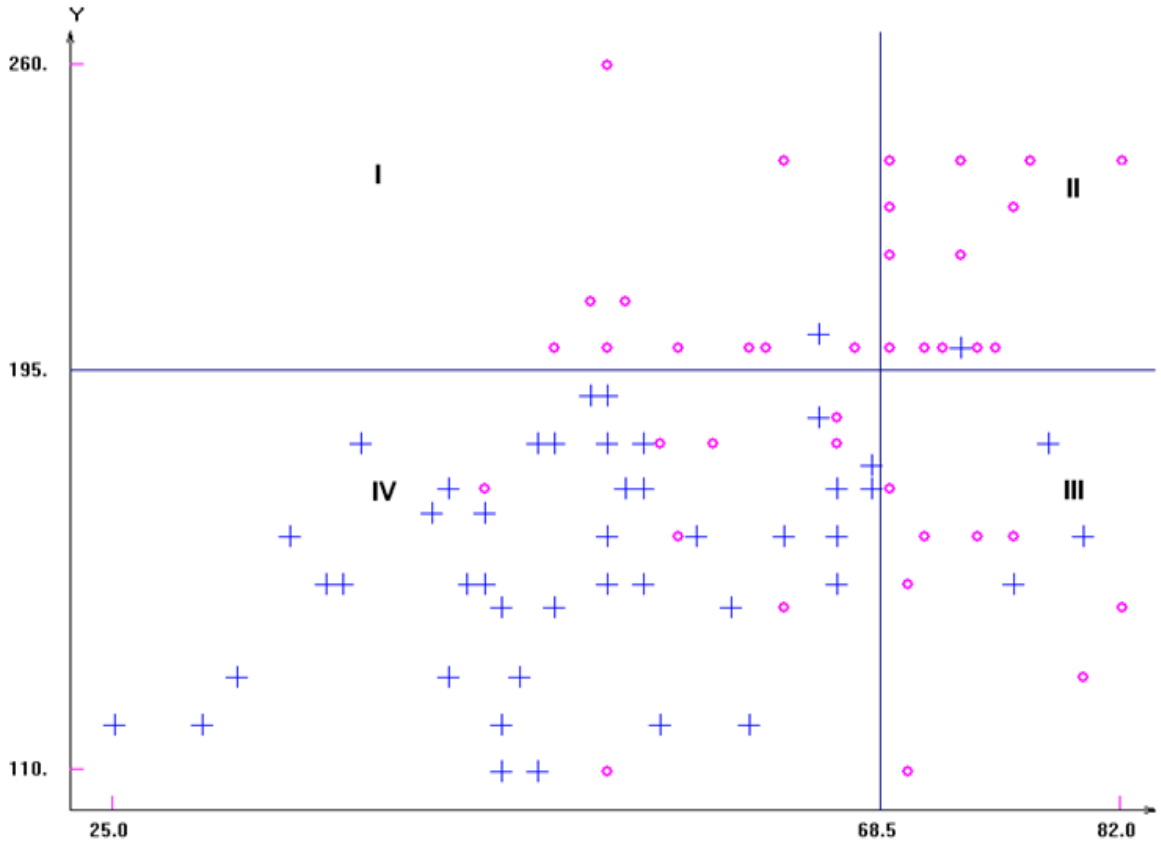

Рис 2. Видно, описания объектов из сосредоточены главным образом в нижнем левом квадранте «синдроме».

Статистически взвешенные синдромы

При этом поиск производится внутри нескольких семейств разбиений, имеющих различный уровень сложности. В ходе поиска выбирается разбиение с максимальным значением функционала качества. Используется два функционала качества, зависящих от обучающей выборки $\widetilde{S}_t$, распознаваемого класса K_l, и разбиения R:

- интегральный $F_i(\widetilde{S}_t, K_l, R)$;
- локальный $F_{loc}(\widetilde{S}_t, K_l, R)$.

Обозначим через $q_1, \ldots, q_r$ элементы некоторого разбиения R . Пусть ν_0^l является долей объектов класса K_l в обучающей выборке $\widetilde{S}_t$, ν_i^l - доля объектов K_l среди объектов, описания которых принадлежат элементу q_i, m_i- число объектов, описания которых принадлежат элементу q_i. Интегральный функционал определяется формулой

$$F_i(\widetilde{S}_t, K_l, R) = \sum_{i=1}^{r} (\nu_0^l - \nu_i^l)^2 m_i.$$

Статистически взвешенные синдромы

Локальный функционал определяется формулой

$$F_i(\widetilde{S}_t, K_l, R) = \max_{i=1,\ldots,r} (\nu_0^l - \nu_i^l)^2 m_i.$$

Поиск разбиений с максимальным значением одного из функционалов производится в рамках одного из четырёх семейств. Примеры разбиений для каждого из семейств приведены на рисунке.
Семейство I включает всевозможные разбиения интервалов допустимых значений отдельных признаков на два интервала с помощью одной граничной точки.
Семейство II включает всевозможные разбиения интервалов допустимых значений отдельных признаков на 3 интервала с помощью двух граничных точек.

Статистически взвешенные синдромы

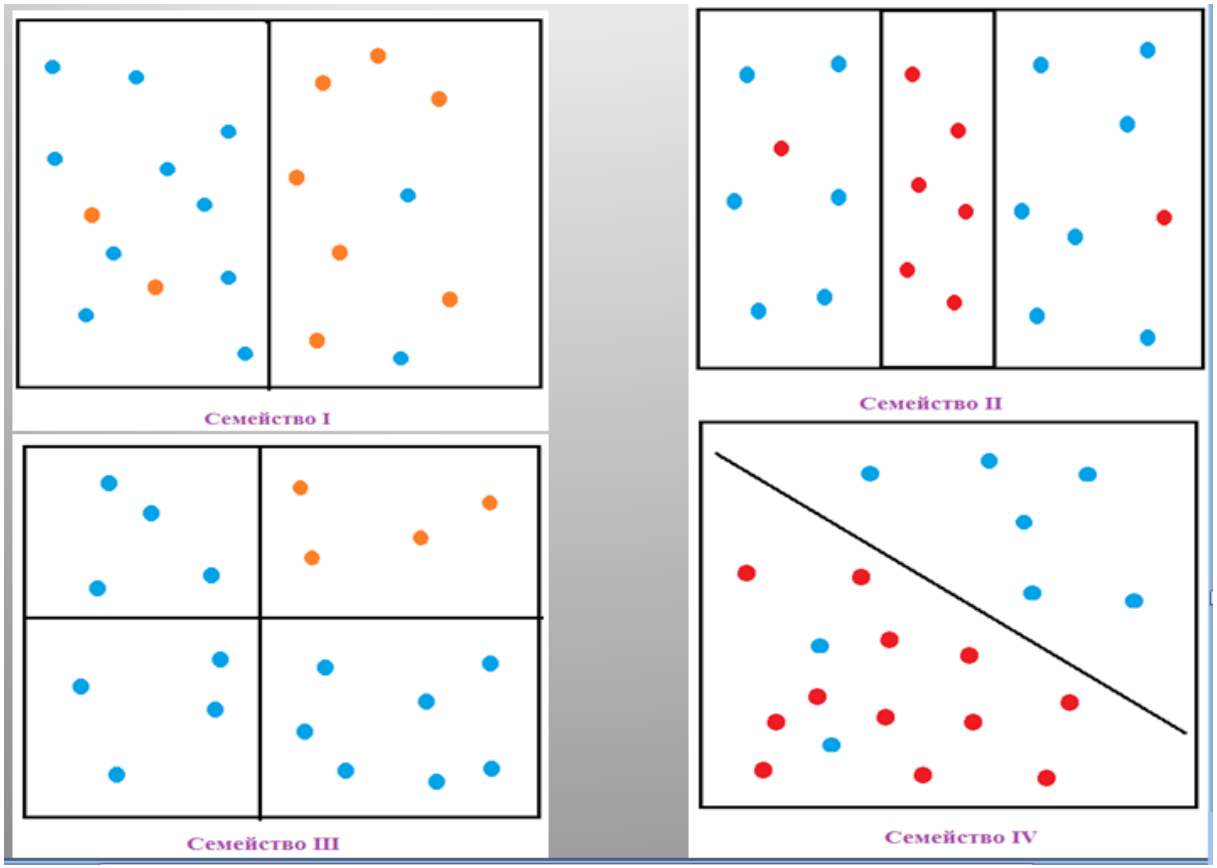

Рис 3. Примеры разбиений для каждого из четырёх семейств, используемых в методе CBC.

Статистически взвешенные синдромы

Семейство III включает всевозможные разбиения совместных двумерных областей допустимых значений пар признаков на 4 подобласти с помощью двух граничных точек (по одной точке для каждого из двух признаков).
Семейство IV включает всевозможные разбиения совместных двумерных областей допустимых значений пар признаков на 2 подобласти с помощью прямой граничной линии, произвольно ориентированной относительно координатных осей.
Найденные оптимальные разбиения используются для построения систем синдромов, если соответствующая им максимальная величина функционала качества превосходит некоторое заранее заданное пользователем пороговое значение δ. Причём величина порога зависит от сложности модели разбиений. Порог является минимальным для простейшей одномерной модели I. Для моделей II-IV величина порога домножается на величину , задаваемую пользователем, что позволяет регулировать влияние эффекта переобучения.

Статистически взвешенные синдромы

Одномерные разбиения, найденные внутри семейств I и II могут быть используются при построении не только одномерных, но также и двумерных синдромов. Предположим, что на этапе обучения для класса K_l найдена система синдромов $\widetilde{Q}_l$. Предположим, что описание $\boldsymbol{x}^*$ распознаваемого объекта s^* принадлежит синдромам $q_1, \ldots, q_r$ из системы $\widetilde{Q}_l$. Оценка s^* за класс K_l вычисляется по формуле

$$\Gamma_l(s^*) = \frac{\sum_{i=1}^{r} w_i^l \nu_i^l}{\sum_{i=1}^{r} w_i^l},$$

где ν_i^l - доля класса K_l в синдроме q_i, w_i^l - вес синдрома при классификации класса K_l. Вес синдрома вычисляется по формуле

$$w_i^l = \frac{m_i}{m_i + 1} \frac{1}{\nu_i^l(1 - \nu_i^l)},$$

где m_i - число объектов обучающей выборки с описанием, принадлежащем q_i.

Метод комитетов

Метод комитетов представляет собой реализацию подхода к решению задач распознавания, объединяющего принципы линейного разделения классов и вычисления коллективных решений. Рассмотрим задачу распознавания с двумя классами K_1 и K_2 . Пусть $\widetilde{f} = \{f_1(\boldsymbol{x}), \ldots, f_r(\boldsymbol{x})\}$ является набором линейных функций вида

$$f_i(\boldsymbol{x}) = a_{1i}x_1 + \ldots + a_{ni}x_n,$$

где $\boldsymbol{x} = (x_1, \ldots, x_n)$ является вектор используемых для распознавания признаков, $(a_{1i}, \ldots, a_{ni})$- вектор вещественных параметров, задающих линейную функцию $f_i(\boldsymbol{x})$. Каждая из функций из $\widetilde{f}$ рассматривается в качестве отдельного линейного классификатора, относящего объект с описанием $\boldsymbol{x}$ в класс K_1, если $sign[f_i\boldsymbol{x}] > 0$, и в класс K_2 в противном случае. .

Метод комитетов

Предположим, что для классификации произвольного объекта s с описанием $\boldsymbol{x}$ используется следующее решающее правило метода комитетов:

- объект s относится в класс K_1 , если $\sum_{i=1}^{r} sign[f_i(\boldsymbol{x})] > 0$;
- объект s с описанием $\boldsymbol{x}$ относится в класс K_2 , если $\sum_{i=1}^{r} sign[f_i(\boldsymbol{x})] < 0$;
- в случае, если $\sum_{i=1}^{r} sign[f_i(\boldsymbol{x})] = 0$ происходит отказ от распознавания.

Набор функций $\widetilde{f}$ называется комитетом, если решающее правило метода комитетов правильно классифицирует объекты обучающей выборки.

Метод, основанный на поиске комитетов, потенциально позволяет производить распознавание линейно неразделимых классов, реализуя кусочно-линейную разделяющую поверхность. Обучение сводится к поиску оптимальных (минимальных по числу функций) комитетов. Теоретически показано существование комитета для непротиворечивых данных.