1 Понятия

$\varphi : X \to 0, 1$ - предикат на множестве объектов. Предикат φ покрывает объект x, если $\varphi(x) = 1$

Предикат называют **закономерностью**, если он покрывает достаточно много объектов класса c и при этом не слишком много объектов других классов.

Закономерности которые описываются простой логической формулой называют правилами (rule). Процесс поиска правил по выборке называют **поиском знаний из данных** (knowledge discovery). Алгоритмы объединяющие несколько правил называются **логическими алгоритмами классификации**.

2 Информативность

2.1 Эвристическое определение информативности

$D = (X, Y), |D| = \ell$

P - число объектов класса c в выборке D

$p(\varphi)$ - объекты из P, которые покрывает φ

N - число объектов класса не c

$n(\varphi)$ - объекты из N, которые покрывает φ

$P + N = \ell$

Требуется построить информативный предикат φ, который одновременно $p(\varphi) \to \max$ и $n(\varphi) \to \min$.

Доля негативных среди всех выделяемых объектов

$$E(\varphi, D) = \frac{n}{p + n}$$

доля выделяемых позитивных

$$D(\varphi, D) = \frac{p}{\ell}$$

Определение. Предикат φ называется логической ε, δ-закономерностью для класса c, если $E \leq \varepsilon$ и $D \geq \delta$ при заданных достаточно малом ε и достаточно большом δ из отрезка $[0, 1]$.

2.2 Статистическая инфомративность

Пусть X - вероятностное пространство. Справедлива гипотеза о независимости событий $x : y^*(x) = c$ и $x : \varphi(x) = 1$ Тогда вероятность реализации пары (p, n) распределена по гипергеометрическому закону

$$h(p,n) = \frac{C_P^p C_N^n}{C_{P+N}^{p+n}}$$

Чем меньше вероятность пары (p, n), тем более значимой является связь между y^* и φ. Другими словами, если реализовалось маловероятное событие, то, скорее всео, оно не случайно, а закономерно.

Определение. Информативность предиката $\varphi(x)$ относительно класса c по выборке D есть

$$I(\varphi, D) = -\ln h(p,n)$$

Предикат $\varphi(x)$ будем назвать статистической закономерностью для класса c, если $I(\varphi, D) \geq I_0$ при заданном достаточно большом I_0.

2.3 Энтропийная инфомративность

Если имеется два исхода ω_1 и ω_2 с вероятностями q_0 и $q_1 = 1 - q_0$, то количество информации, связанное с исходом ω_i по определению равно $-\log_2 q_i$.

Энтропия, определяемая как матожидание количества информации:

$$H(q_0, q_1) = -q_0 \log_2 q_0 - q_1 \log_2 q_1$$

Будем считать появление объекта c исходом ω_0, а появление объекта любого другого класса исходом ω_1. Тогда можно вычислить энтропию выборки

$$\hat{H}(P,N) = H\left(\frac{P}{P+N}, \frac{N}{P+N}\right)$$

Допустим предикат φ выдели p объектов из P и n объектов из N. Тогда энтропия подвыборки $x \in X|\varphi(x) = 1$ есть $\hat{H}(p,n)$.

вероятностность появления объекта из этой выборки оценивается как $(p+n)/(P+N)$.

Аналогично для подвыборки $x \in X|\varphi(x) = 0$, энтропия равна $\hat{H}(P-p, N-n)$, а вероятность появления объекта из неё оценивается как $(P-p+N-n)/(P+N)$. Таким образом, энтропия всей выборки после получения информации φ становится равна

$$\hat{H}_\varphi(p,n) = \frac{p+n}{P+N}\hat{H}(p,n) + \frac{P+N-p-n}{P+N}\hat{H}(P-p, N-n)$$

Уменьшение энтропии составляет

$$IGain(\varphi, D) = \hat{H} - \hat{H}_\varphi$$

Так же это называют информационным выигрышем — количество информации об исходном делении выборки на два класса «c» и «не c», которое содердится в предикате φ.

Определение. Предикат φ является закономерностью по энтропийному критенрию информативности, есои $IGain(\varphi, D) > G_o$ при достаточно большом G_0.

Теорема. Энтропийный критерий $IGain$ асимптотически эквивалентен статистическому I

$$IGain(\varphi, D) \to_{\ell \to +\infty} \frac{1}{\ell \log_2} I(\varphi, D)$$