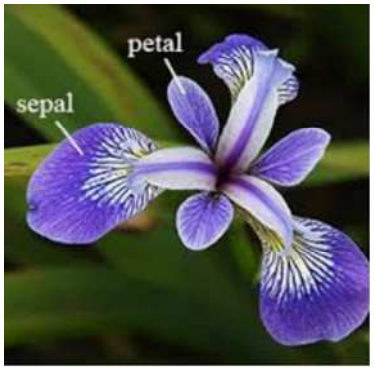
Задача распознавания образов

$A = \{a_1, \ldots, a_n\}$ – множество объектов;

$X = (X_1, \ldots, X_m)$ – набор признаков, $x_{i,j} = X_j(a_i)$; таблица данных

$$\mathbf{X} = (x_{i,j}).$$



	A	B	C	D	E
1	sepal_length	sepal_width	petal_length	petal_width	species
2	5.1	3.5	1.4	0.2	setosa
3	4.9	3	1.4	0.2	setosa
4	4.7	3.2	1.3	0.2	setosa
5	4.6	3.1	1.5	0.2	setosa
6	5	3.6	1.4	0.2	setosa
7	5.4	3.9	1.7	0.4	setosa
8	4.6	3.4	1.4	0.3	setosa
9	5	3.4	1.5	0.2	setosa
10	4.4	2.9	1.4	0.2	setosa
11	4.9	3.1	1.5	0.1	setosa
12	5.4	3.7	1.5	0.2	setosa
13	4.8	3.4	1.6	0.2	setosa
14	4.8	3	1.4	0.1	setosa
15	4.3	3	1.1	0.1	setosa
16	5.8	4	1.2	0.2	setosa
17	5.7	4.4	1.5	0.4	setosa
18	5.4	3.9	1.3	0.4	setosa
19	5.1	3.5	1.4	0.3	setosa
20	5.7	3.8	1.7	0.3	setosa
21	5.1	3.8	1.5	0.3	setosa

Y – зависимая **качественная** переменная, $D_Y = \{\omega_1, \ldots . \omega_K\}$,
ω_l - l-й «класс», «образ» (или $D_Y = \{1, \ldots, \omega, \ldots, K\}$)

Требуется построить по наблюдениям оптимальное по некоторому критерию качества решающее правило классификации:

для любого нового $x = (x_1, \ldots, x_m) \in R^m$: $x \xrightarrow{f} \omega$.

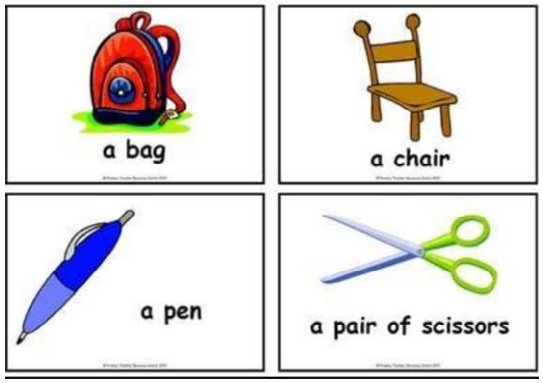
$L(y', y)$ - *функция потерь*, возникающих в случае принятия решения $f(x) = y'$, когда истинное значение есть y.

$$L(y', y) \geq 0.$$

функция потерь может быть несимметричной:

$$L(y', y) \neq L(y, y').$$

$L(y', y) = \begin{cases} 0, & y' = y \\ 1, & y' \neq y \end{cases}$ - индикаторная функция потерь.

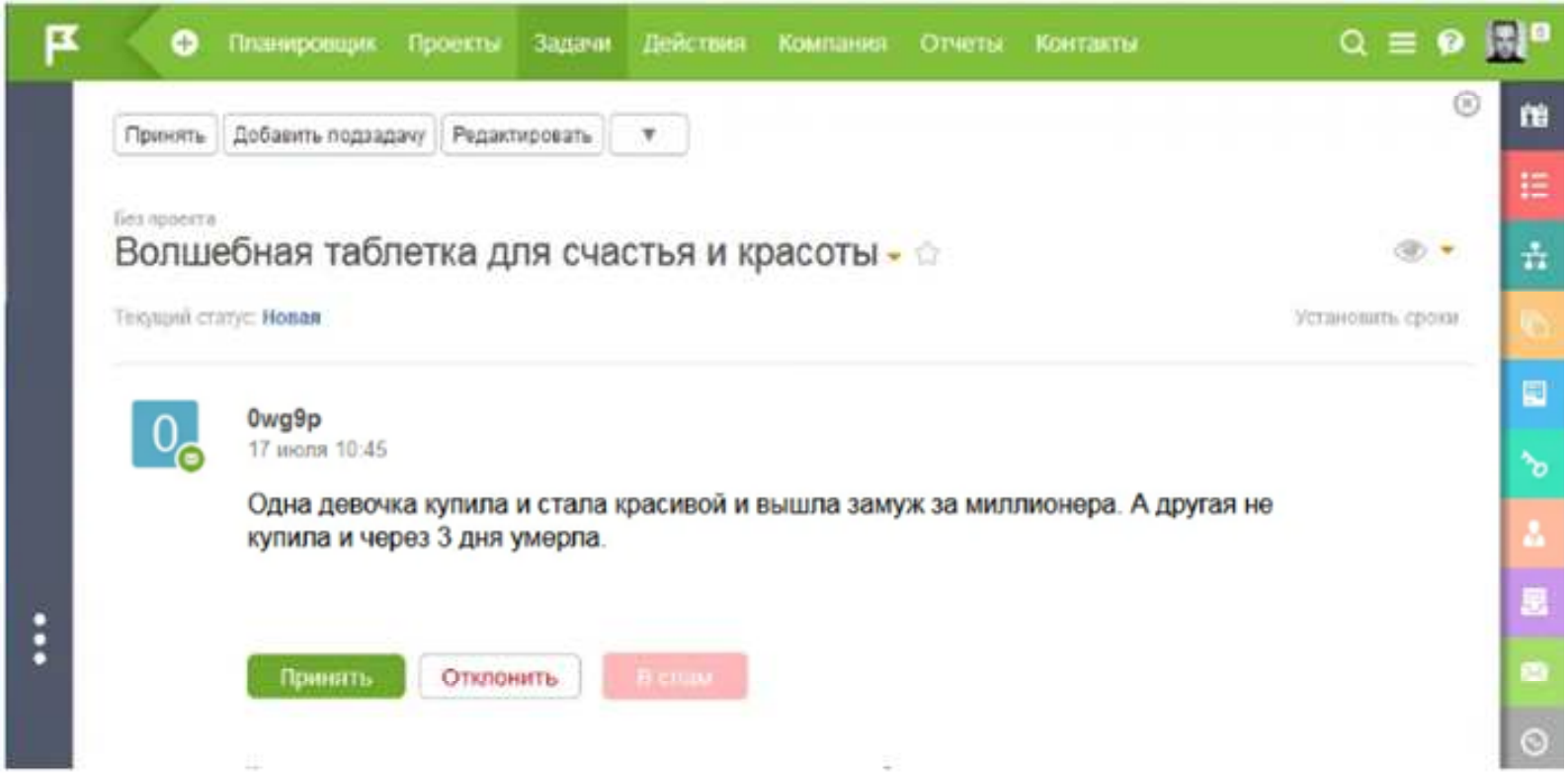


a pen

Критерий качества решающей функции – например, риск неправильного распознавания $R_f = E_{X,Y} L(f(X), Y)$.

Если функция потерь – индикаторная, то $R_f = 1 \cdot P[f(X) \neq Y] + 0 \cdot P[f(X) = Y] = P[error]$ - вероятность ошибки классификации.

Пример: задача распознавания спама



Признаки: слова из заданного набора;
значения: $X_j = 1$ - слово встречается в письме; $X_j = 0$ иначе.

Образ $\omega = 1$ - письмо – спам, $\omega = 2$ – не спам.

Байесовское распознавание образов

Вероятностная модель

Предположим, известны:

$P(\omega)$ - априорная вероятность образа ω;

$P(x|\omega)$ - закон условного распределения переменных для образа ω,

Тогда по формуле Байеса:

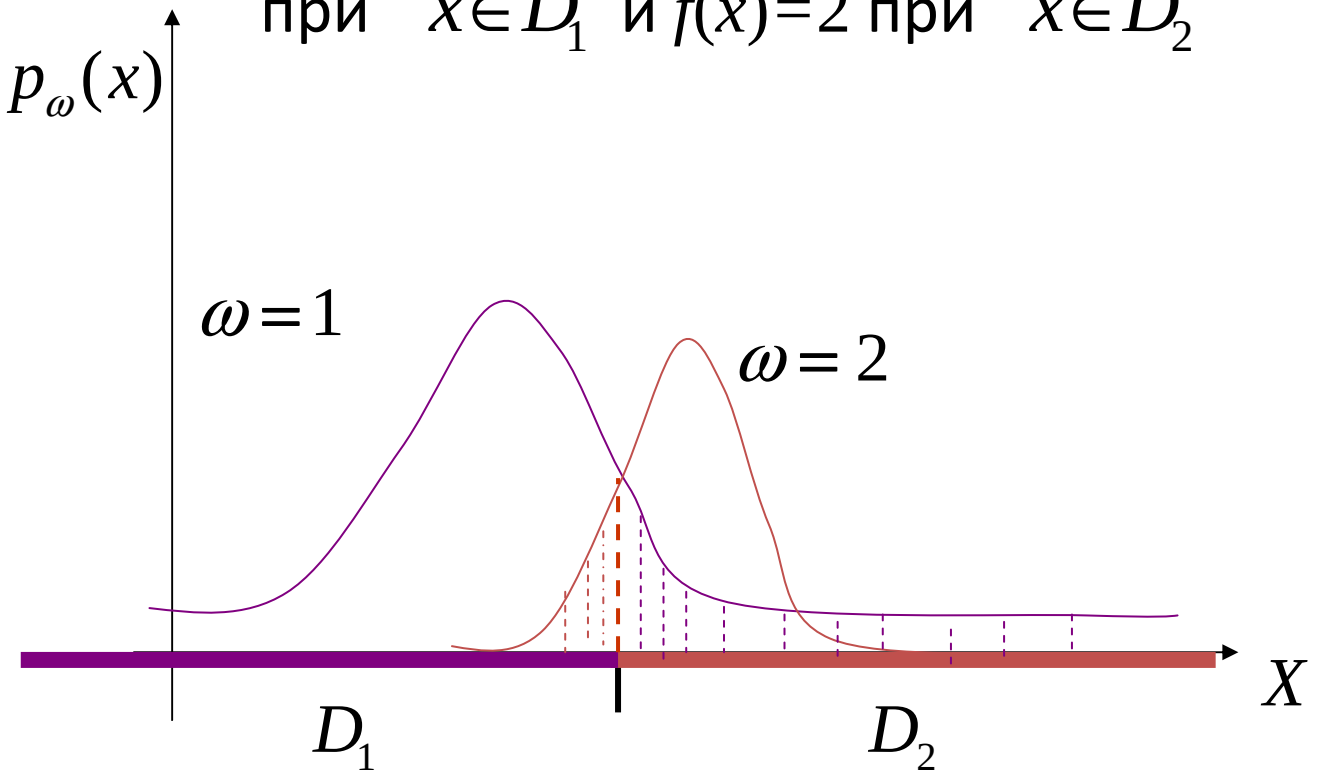
$$P(\omega|x) = \frac{P(\omega)P(x|\omega)}{P(x)},$$

где

$P(\omega|x)$ - апостериорная вероятность класса ω для точки x,

$P(x) = \sum_{\omega=1}^{K} P(\omega)P(x|\omega)$ (по формуле полной вероятности).

Пример. Решающая функция: $f(x)=1$ при $x \in D_1$ и $f(x)=2$ при $x \in D_2$



$$P[error] = P(1)\int_{D_2} p(x|1)dx + P(2)\int_{D_1} p(x|2)dx$$

Байесовская решающая функция

$$f_B : x \to \omega^* : \; P(\omega^* \mid x) = \arg\max_{\omega} P(\omega \mid x)$$

- максимизирует апостериорную вероятность образа в точке x.

Формула Байеса $\Rightarrow \; P(\omega^* \mid x) = \arg\max_{\omega} P(\omega) P(x \mid \omega)$

(так как $P(x)$ не зависит от ω).

Так как $P(\omega)\, p(x \mid \omega) = P(\omega \mid x) \cdot P(x)$, то для б.р.ф. выполняется:

$$P(1)\, p(x \mid 1) < P(2)\, p(x \mid 2) \text{ при } x \in D_2$$
$$P(1)\, p(x \mid 1) > P(2)\, p(x \mid 2) \text{ при } x \in D_1.$$

Вероятность ошибки распознавания:

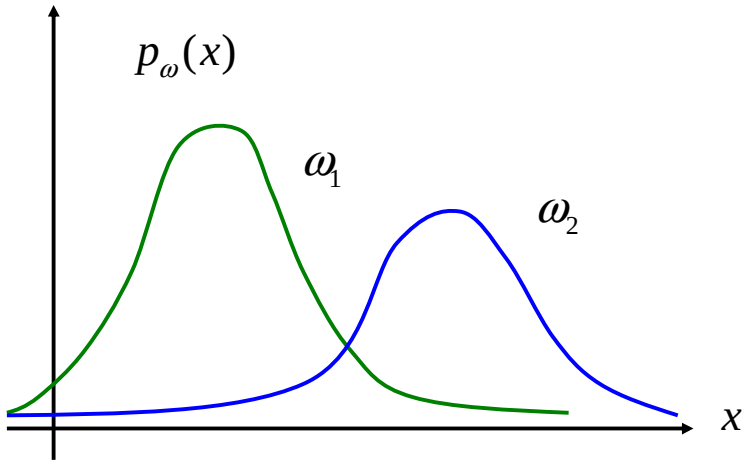
$$P[error] = \int_{D_2} P(1)\, p(x \mid 1) dx + \int_{D_1} P(2)\, p(x \mid 2) dx$$

$$\Downarrow$$

Теорема. Байесовская решающая функция минимизирует вероятность ошибочного распознавания.

Случай двух классов и одной переменной

Модель нормального распределения для каждого класса:
ω_1 : плотность $p_1(x) \sim N(\mu_1, \sigma_1)$, ω_2 :плотность $p_2(x) \sim N(\mu_2, \sigma_2)$;
заданы априорные вероятности $P(1)$, $P(2)$ $\quad (P(1)+P(2)=1)$.



На границе принятия решения для б.р.ф. должно выполняться:

$$P(1)\,p(x\,|\,1) = P(2)\,p(x\,|\,2) \text{ или}$$

$$\ln\frac{p(x\,|\,1)}{p(x\,|\,2)} + \ln\frac{P(1)}{P(2)} = 0.$$

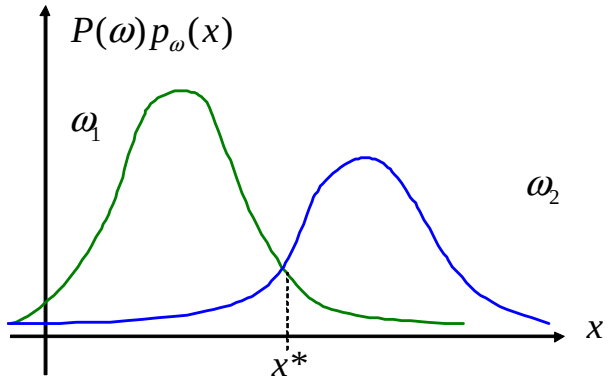
$$\ln\left(\frac{e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}}{\sqrt{2\pi}\sigma_1}\Big/\frac{e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}}{\sqrt{2\pi}\sigma_2}\right)+\ln\frac{P(1)}{P(2)}=0;$$

$$\ln\frac{\sigma_2}{\sigma_1}+\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}+\frac{(x-\mu_2)^2}{2\sigma_2^2}\right)+\ln\frac{P(1)}{P(2)}=0;$$

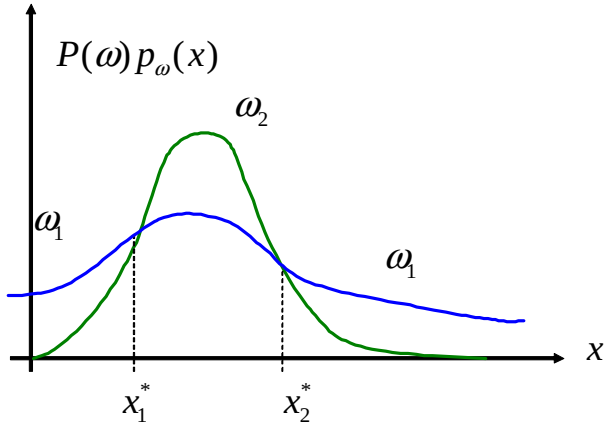
$$\frac{(x-\mu_2)^2}{\sigma_2^2}-\frac{(x-\mu_1)^2}{\sigma_1^2}=2\ln\left(\frac{P(2)}{P(1)}\frac{\sigma_1}{\sigma_2}\right)$$

- квадратное уравнение; получим варианты решения:

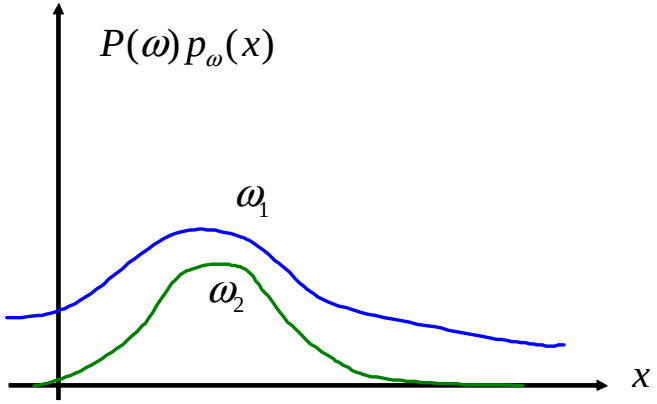
а) единственный корень x^*:



б) два корня x_1^*, x_2^*:



в) нет корней:



Вместо параметров подставим их оценки:

$$\hat{P}(1) = \frac{n_1}{n},\ \hat{P}(2) = \frac{n_2}{n},$$

$$\hat{\mu}_1 = \frac{1}{n_1} \sum_{i:Y(i)=1} x_i, \qquad \hat{\mu}_2 = \frac{1}{n_2} \sum_{i:Y(i)=2} x_i,$$

$$\hat{\sigma}_1^2 = \frac{1}{n_1} \sum_{i:Y(i)=1} (x_i - \hat{\mu}_1)^2, \qquad \hat{\sigma}_2^2 = \frac{1}{n_2} \sum_{i:Y(i)=2} (x_i - \hat{\mu}_2)^2,$$

где n_1, n_2 - число объектов 1 и 2-го образа в выборке.

Получим уравнение оптимальной выборочной разделяющей функции.

«Наивный» байесовский классификатор

Рассмотрим случай произвольных качественных переменных X_1, X_2. Пусть $X_1 \in \{u_1,...,u_L\}$, $X_2 \in \{v_1,...,v_Q\}$.

Предположим, все переменные независимы.

Тогда $\forall x = (u_l, v_q)$:

$$P(x \mid \omega) = P(u_l \mid \omega) \cdot P(v_q \mid \omega),\ \omega = 1,2.$$

Байесовская решающая функция:

$$x \to \omega^*:\ P(\omega^* \mid x) = \arg\max_{\omega} P(\omega) P(u_l \mid \omega) \cdot P(v_q \mid \omega).$$

Выборочные оценки:

$$\hat{P}\,(u_l \mid \omega) = \frac{n_\omega(u_l)}{n_\omega},\quad \hat{P}\,(v_q \mid \omega) = \frac{n_\omega(v_q)}{n_\omega},\ \hat{P}(\omega) = \frac{n_\omega}{n},$$

где $n_\omega(u_l), n_\omega(v_q)$ - число объектов образа ω, у которых переменная X_1 приняла значение u_l, а X_2 приняла значение v_q.

Аналогично для переменных $X_1,...,X_m$, $m>2$, $x=(u_{l,1},...,u_{l,m})$:

$$P(x\,|\,\omega)=P(u_{l,1}\,|\,\omega)\cdot...\cdot P(u_{l,m}\,|\,\omega).$$

Таким образом, многомерная условная вероятность определяется путем нахождения одномерных условных вероятностей.

Пример. 10 булевых переменных; $n=100$. $|X|=2^{10}$=1024, поэтому для большинства точек x выполняется $\hat{P}(x\,|\,\omega)=0$, хотя наверняка $P(x\,|\,\omega)>0$.

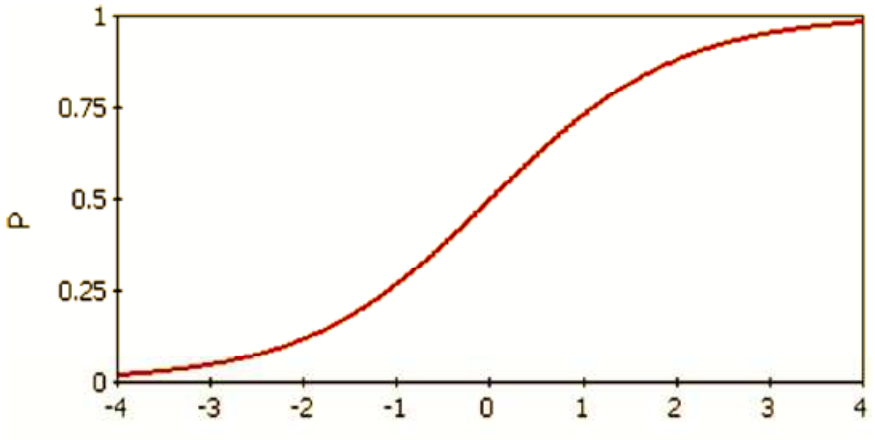
Для одномерных условных вероятностей частоты $\hat{P}(u_{l,1}\,|\,\omega)$ будут положительными $\Rightarrow$ можно оценить вероятность $P(x\,|\,\omega)$ более точно.

Логистическая регрессия

Метод, позволяющий использовать аппарат регрессионного анализа в распознавании образов

Пусть $Y \in \{0,1\}$. Предположим, вероятность $P(Y=1 \mid x) = f(x)$, где $f(x)$ - заданная функция, тогда $P(Y=0 \mid x) = 1 - f(x)$, или $P(y \mid x) = f(x)^y (1 - f(x))^{1-y}$.

Логистическая функция – $\sigma(x) = \sigma(x;\beta) = \dfrac{1}{1+e^{-(\beta_0 + \sum \beta_j x_j)}}$ (сигмоид).



По таблице данных $(x^{(i)}, y^{(i)})$, $i = 1,..., N$ требуется найти параметры модели, оптимальные по некоторому критерию.

Например, пусть максимизируется логарифмическая функция правдоподобия: $\log L(\beta) = \log \prod_{i=1}^{N} P(Y = y^{(i)} \mid X = x^{(i)}) =$

$$= \sum_{i=1}^{N} \log P(Y = y^{(i)} \mid x = x^{(i)})$$

$$= \sum_{i=1}^{N} (y^{(i)} \log \sigma(x^{(i)}; \beta) + (1 - y^{(i)}) \log(1 - \sigma(x^{(i)}; \beta)).$$

Можно показать, что задача выпукла, т.е. решение единственно.
поиск оптимального решения – градиентный метод (после преобразования $\nabla \log L(\beta)$):

$$\beta := \beta + \tau \nabla \log L(\beta) = \beta + \tau \sum_{i=1}^{N} (y^{(i)} - \sigma(x^{(i)}; \beta))\, x^{(i)},$$

где $\tau > 0$ - параметр.

Принятие решения:

если $\sigma(x; \beta) > \frac{1}{2}$, то $y = 1$, иначе $y = 0$.

Характеристики качества решения бинарного классификатора

Рассмотрим бинарную задачу распознавания:

$$D_Y = \{True, False\} = \{+, -\},$$

и некоторую решающую функцию.

Таблица сопряженности

True class	Predicted class	
	Positive	Negative
Positive (Pos)	TP	FN=Pos - TP
Negative (Neg)	FP	TN=Neg - FP

accuracy = (TP+TN)/(Pos+Neg);

recall = TP/Pos (true positive rate, TPR, sensitivity, полнота);

precision = TP/ (TP+FP);

F1= 2•(precision·recall)/(precision+recall)
гармоническое средняя precision и recall (отражает особенности решений в случае несбалансированных классов)

Эмпирическая ошибка

Вместо вероятности ошибки можно использовать ее оценку – частоту ошибки (эмпирическую, подстановочную ошибку) для решающей функции f:

$$\hat{P}_f[error] = \frac{1}{n}\sum_{i=1}^{N} \mathbf{I}[f(x_i) \neq y_i],$$

где $\mathbf{I}[U] = \begin{cases} 1, U = true \\ 0, U = false. \end{cases}$

- метод минимизации эмпирической ошибки:

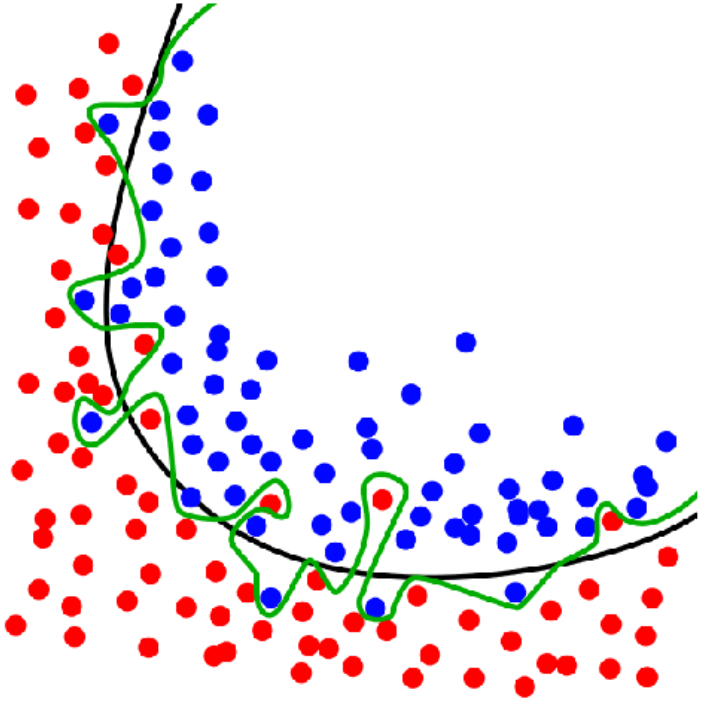
Найти f^* : $\hat{P}_{f*}[error] = \min_f \hat{P}_f[error]$.

Например, нужно найти линейную разделяющую функцию, для которой частота ошибок на обучающей выборке минимальна.

Недостатки метода:

- при ограниченной выборке, частотная оценка может обладать большой погрешностью («занижает» вероятность ошибки); решения – далеки от оптимальных.

- проблема «переобучения»: если класс решающих функций сложный, то можно подобрать такую решающую функцию, которая на обучающей выборке дает низкую частоту ошибки («подстраивается» под шум), но при распознавании новых объектов вероятность ошибки велика.



Экспериментальное оценивание вероятности ошибки

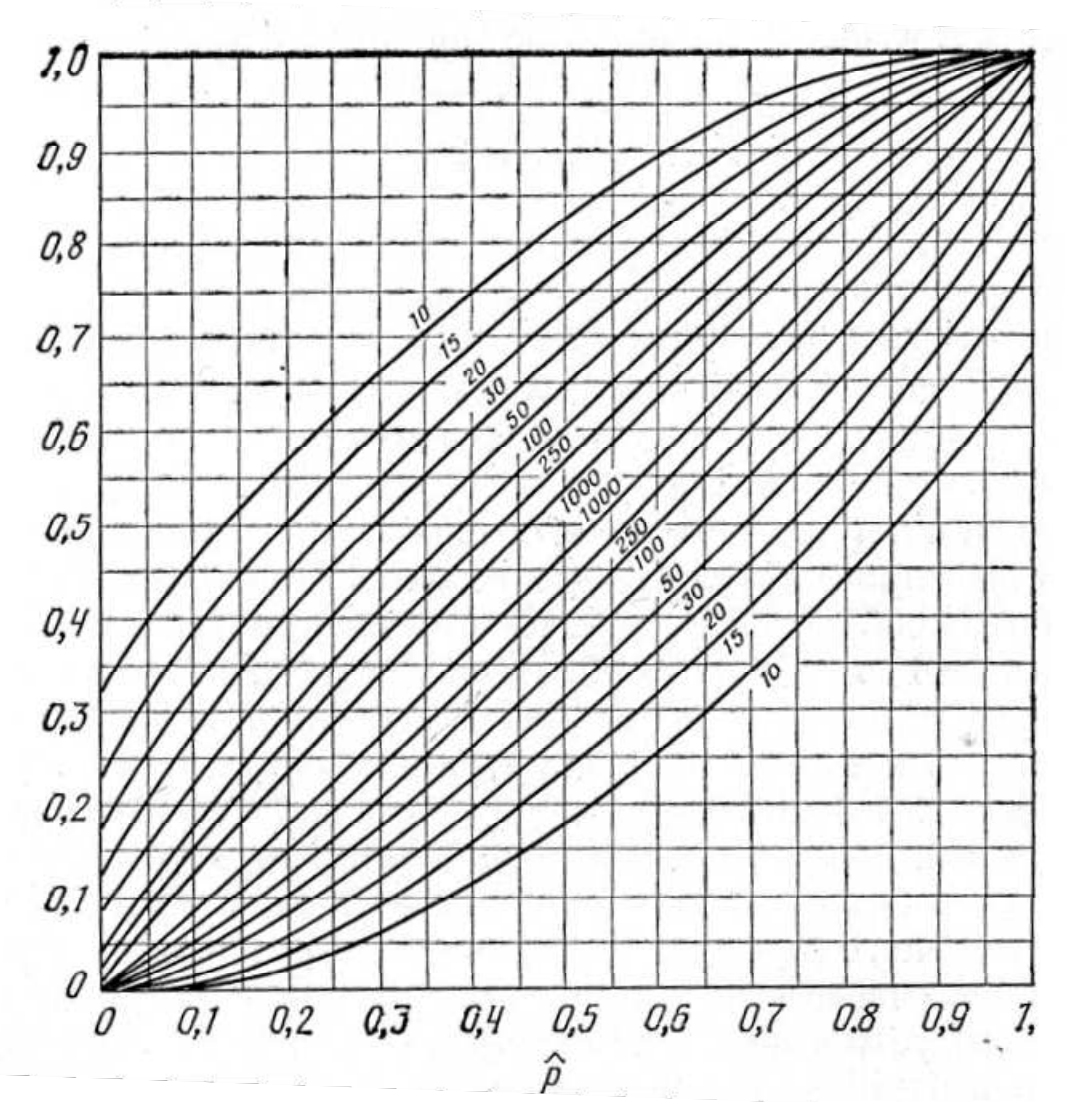
Разделение выборки на обучающую (по которой находится решающая функция) и контрольную (по которой определяется качество решающей функции).

Контрольной (экзаменационной) выборкой называют выборку, которая не используется при формировании решающей функции, а служит для оценки ее качества путем вычисления относительного числа ошибок.

- Более объективно отражает «истинную» неизвестную ошибку.

- При условии независимости наблюдений, частота ошибок подчиняется биномиальному распределению. Зная число ошибок на контрольной выборке, можно найти доверительный интервал, в котором с заданной вероятностью находится неизвестное значение вероятности ошибки.

Номограмма для определения доверительных интервалов в зависимости от частоты и объема выборки.



$\gamma = 0.95$

Метод скользящего экзамена.

Поочередно каждый объект выборки «выбрасывается» из нее, по оставшейся части выборки строится классификатор, с помощью которого затем находится прогноз для данного объекта. Прогнозируемое значение сравнивается с наблюдаемым, после чего объект возвращается в исходную выборку. Процент ошибок показывает качество метода.

- Большая трудоемкость, так как необходимо решить n задач построения решающей функции (n - объем выборки).

Скользящий экзамен оценивает не конкретную решающую функцию, но метод ее построения

Метод L-кратной перекрестной проверки ("L-fold cross-validation").

Исходная выборка случайным образом делится на L частей, приблизительно одинаковых по объему. Затем каждая часть поочередно выступает как контрольная выборка, а оставшиеся части объединяются в обучающую. Показателем качества метода служит усредненная по контрольным выборкам ошибка.