## 4.1. Наивный байес и центроидный классификатор

$$P(x^{(k)}|y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x^{(k)}-\mu_{yk})^2}{2\sigma^2}}, \quad x^{(k)}, \quad k = \overline{1 \ldots n}$$

$\Delta$

$$\ln \prod_{k=1}^{n} P(x^{(k)}|y) = \sum_{k=1}^{n} \ln P(x^{(k)}|y) =$$

$$= -n \ln \sqrt{2\pi\sigma^2} + \sum_{k=1}^{n} -\frac{(x^{(k)}-\mu_{yk})^2}{2\sigma^2} =$$

В евклидовом пространстве $\rho^2(x^{(k)}, \mu_{yk}) = \sum_{k=1}^{n}(x^{(k)} - \mu_{yk})^2$

тогда метрика минимальна: центр $\mu_y$ ближе всего к $x$, относим объект $x$ к классу $y$

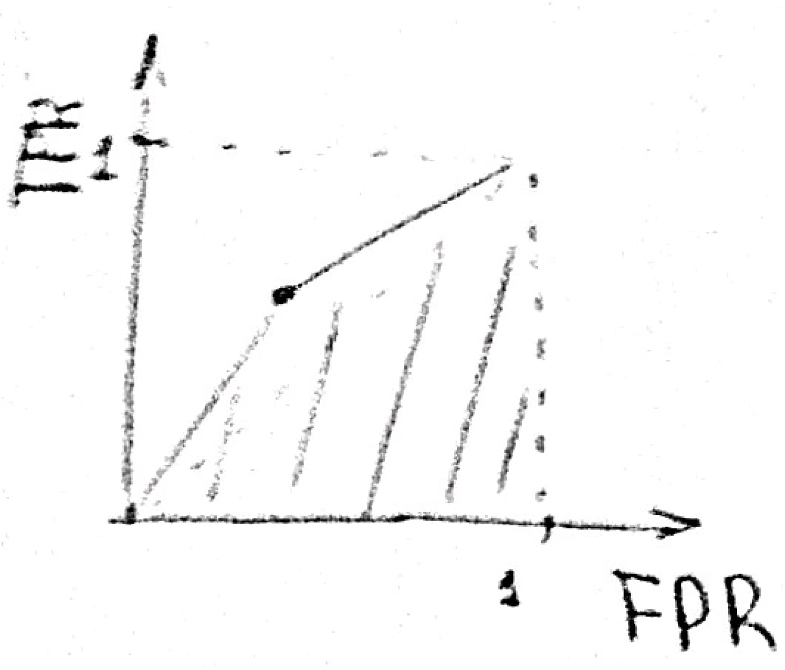
можно представить ф-ю в виде $= C_1 - \frac{\rho^2(x^{(k)}, \mu_{yk})}{C_2}$

где $C_1$, $C_2$ - константы.

Видно, что при уменьшении $\rho$, вероятность попадания в класс увеличивается. $\square$.

## 4.2. ROC-AUC случайных ответов

$a(x)=1$ с вероятностью $p$, $a(x)=0$ с вероятностью $1-p$




$$TPR = \frac{TP}{TP+FN}, \quad FPR = \frac{FP}{FP+TN}$$

$\Delta$

Пусть $n$ - размер выборки

$k$ - доля $a(x) = 1$

Тогда
$TP = p \cdot k \cdot n$
$FP = p(1-k)n$
$TN = (1-p)(1-k)n$
$FN = (1-p)k \cdot n$

$$TPR = \frac{pkn}{pkn + (1-p)kn} = \frac{pk}{pk + k - pk} = p$$

$$FPR = \frac{p(1-k)n}{p(1-k)n + (1-k)n - p(1-k)n} = p$$

т.к. TPR = FPR, то ROC curve имеет вид прямой, т.е. ROC-AUC = 0.5, независимо от p и доли класса 1 (h) в выборке. □.

## 4.3. Ошибка 1NN и оптимального байесовского классификатора

$$E_B = \min\{P(1|x), P(0|x)\}, \quad E_N = P(y \neq y_n)$$

Δ. $E_N = P(y \neq y_n) = P(y=1, y_n=0) + P(y=0, y_n=1) =$

$= P(0|x)\cdot P(1|x_n) + P(1|x)P(0|x_n) \approx$

$\approx 2P(0|x)P(1|x) \circleddash$ (распределение непрерывны, $\underset{n\to\infty}{x_n \approx x}$)

$\circleddash\ 2(1-E_B(x))E_B(x) \leq 2E_B$

Т.е. метод 1NN асимптотически имеет мат. ожидание ошибки не более чем вдвое больше по сравнению с оптимальным байесовским классификатором. □.