Оценка объема выборки в задачах классификации

Анастасия Мотренко

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Научный руководитель к.ф.-м.н., н.с. ВЦ РАН В. В. Стрижов

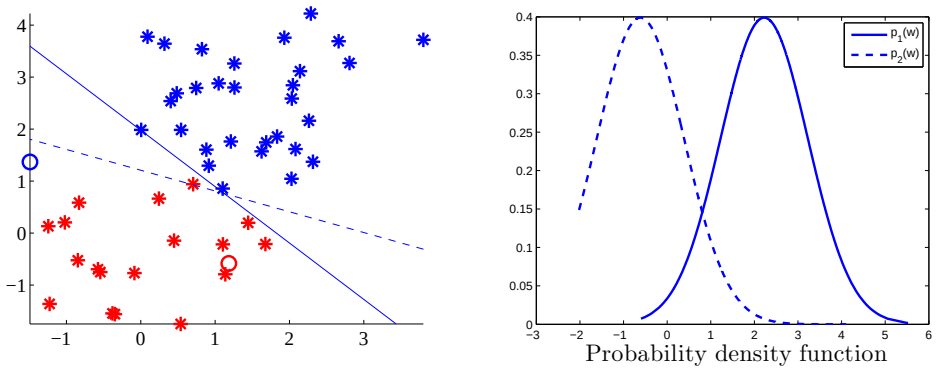
Москва, 2014 г.

Цель

Цель: Предложить метод оценки объема выборки на основе близости между эмпирическими распределениями побвыборок для получения оптимального качества классификации при выборе между порождающим и разделяющим подходами.

1. Определение достаточного объема выборки. Оценка объема выборки на основе расстояния Кульбака-Лейблера
2. Свойства расстояния Кульбака-Лейблера
3. Задача классификации: разделяющий и порождающий подходы
4. Оценка объема выборки при выборе между подходами
5. Основные результаты

Изменение эмпирических распределений при недостаточном объеме выборки



a. Изменение положения разделяющей гиперплоскости $\mathbf{x}^\mathsf{T}\mathbf{w} + c = 0$ при добавлении двух объектов в выборку.

b. Эмпирические плотности распределения $\hat{p}_1$ и $\hat{p}_2$, оцененные на различных подвыборках.

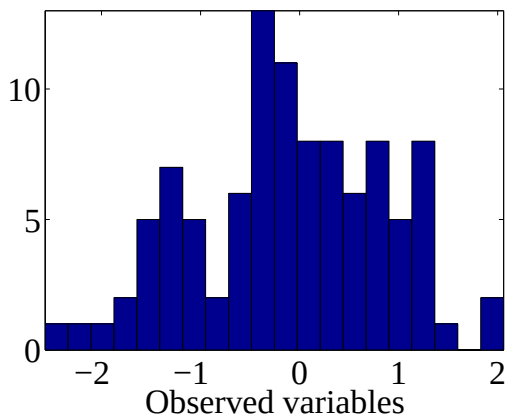
Будем называть объем выборки m^* из распределени P достатоточным, если все выборки X_1, X_2 объема $m > m^*$ из P близки согласно некоторой функции расстояния $D(\hat{P}_1, \hat{P}_2)$ между эмпирическими распределениями, оцененными на этих выборках.

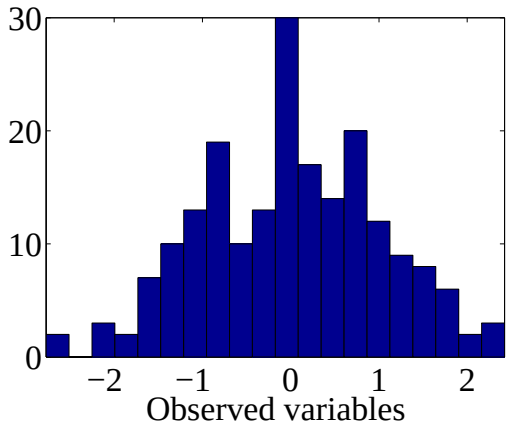
Сравнение гистограмм

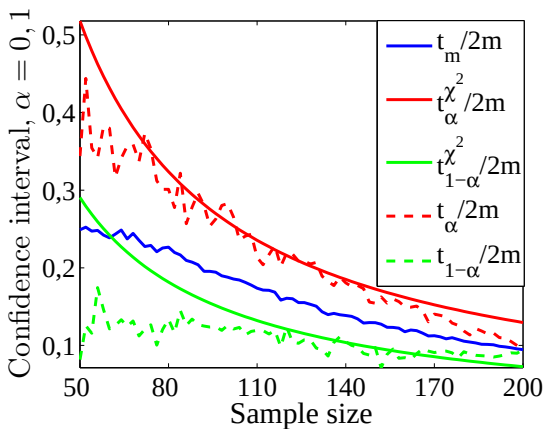
Рассмотрим выборку X объема m и разбиение области значений X на N промежутков $[a_i, a_{i+1}]$. Набор оценок

$$\hat{P}_m(a_i < x \le a_{i+1}) = \frac{n_i}{m} = \hat{p}_i, \quad i = 1, \ldots, N$$

вероятности $p_i = P(a_i < x \le a_{i+1})$ будем называть гистограммой $\hat{P}_m$.







Литература по оценке расстояния между распределениями

1. S. M. Ali, S. D. Silvey. 1966. A general class of coefficients of divergence of a distribution from another. Journal of Royal Statistical Soiety. Series B (Methodoligical). 1(28):131-142.
2. I. Csiszar and P. Shields. 2004. Information theory and statistics: A tutorial. Foundations and Trend in Communications and Information Theory, 4:417–528.
3. A. L. Gibbs, F. E. Su. 2002. On Choosing and bounding probability metrics. International Statistical Review. 3(70):419–435.

Расстояние Кулльбака-Лейблера между распределениями Q иP

$$D_{\mathrm{KL}}(Q||P) = \sum P \cdot f\left(\frac{Q}{P}\right), \quad \text{где } f(t) = t \ln t.$$

Свойства:

1. Расстояние $D_{\mathrm{KL}}(Q, P)$ определено на всех парах распределений с одним носителем.
2. Значение $D_{\mathrm{KL}}(Q, P)$ минимально при $P = Q$.
3. Пусть P_m — гистограмма, построенная по выборке из распределения P. При $m \to \infty$ имеет место

$$2m \cdot D_{\mathrm{KL}}(\hat{P}_m||P) \to \chi^2_N.$$

Theorem (Мотренко, 2014)

Случайная величина $2m \cdot D_{KL}(P||\hat{P}_m) \to \chi^2_N$ *по распределению при* $m \to \infty$.

$$D_{\mathrm{KL}}(P||\hat{P}_m) \sim \frac{1}{2m} \sum_{i=1}^{N} \frac{(n_i - mp_i)^2}{n_i}.$$

Пусть $G_m(x)$ — ф-ия распределения величины $\sum_{i=1}^{N} \frac{(n_i - mp_i)^2}{mp_i}$, $F_m(x)$ — случайной величины $\sum_{i=1}^{N} \frac{(n_i - mp_i)^2}{n_i}$. Нужно показать, что $|F_m(x) - F_{\chi^2_{N-1}}| < \varepsilon$. Так как

$$|F_m(x) - F_{\chi^2_{N-1}}| < |F_m(x) - G_m(x)| + |G_m(x) - F_{\chi^2_{N-1}}|,$$

достаточно показать, что $|F_m(x) - G_m(x)| < \varepsilon/2$.

Схема доказательства

Покажем, что $\forall \varepsilon > 0 \, \exists \, m_0$: при всех $m > m_0$ выполняется

$$\mathrm{P}\left(\left|\frac{(n_i - mp_i)^2}{n_i} - \frac{(n_i - mp_i)^2}{mp_i}\right| \leq \frac{\varepsilon}{N}\right) > 1 - \varepsilon.$$

Согласно ЦПТ, $\frac{n_i - mp_i}{p_i(1-p_i)\sqrt{m}} \to \mathcal{N}(0,1)$, причем

$$\mathrm{P}\left(\left|\frac{n_i - mp_i}{p_i(1-p_i)\sqrt{m}}\right| < C\right) \geq 2\Phi(C) - 1 - \frac{2A}{\sqrt{m}}.$$

Пусть $\forall i$ выполнено $1 - p < p_i < p$. Тогда с вероятнотью $\mathrm{P}_m(\varepsilon) \geq 2\Phi(C_{m,\varepsilon}) - 1 - \frac{2A}{\sqrt{m}}$ при $m > [4C^2(1-p)^2]$

$$\left|\frac{(n_i - mp_i)^2}{n_i} - \frac{(n_i - mp_i)^2}{mp_i}\right| = \frac{|n_i - mp_i|^3}{mn_i p_i} \leq \frac{C^3(1-p)^3 p}{\sqrt{m}} < \frac{\varepsilon}{N}.$$

Выберем

$$C_{m,\varepsilon} = \frac{\epsilon^{1/3} m^{1/6}}{(1-p_i)(2p_i N)^{1/3}}, \; \mathrm{P}_m(\varepsilon) = 2\Phi(C_\varepsilon) - 1 - \frac{2A}{\sqrt{m}}.$$

Следствие 1: $2m \cdot D_{\mathrm{KL}}(\hat{P}_m || \hat{P}'_m) \leq \chi^2_{2N}$ в пределе при $m \to \infty$.
Следует из неравенства треугольника

$$D_{\mathrm{KL}}(\hat{P}_m || \hat{P}'_m) \leq D_{\mathrm{KL}}(\hat{P}_m || P) + D_{\mathrm{KL}}(P || \hat{P}'_m).$$

Следствие 2: Пусть объем выборок X, X' растет таким образом, что $m/l \to \rho$, $0 < \rho < \infty$. Тогда

$$2\frac{ml}{m+l} \cdot D_{\mathrm{KL}}(\hat{P}_m || \hat{P}_l) \leq \chi^2_{2N}$$

в пределе при $m, l \to \infty$.

Доказательство. При выполнении условия $m/l \to \rho$, $0 < \rho < \infty$, имеем

$$\frac{l}{m+l} \to \frac{1}{1+\rho}, \quad \frac{m}{m+l} \to \frac{\rho}{1+\rho}$$

и

$$\frac{2ml}{m+l} D_{\mathrm{KL}}(\hat{P}_m || \hat{P}_l) \leq \frac{l}{m+l} 2m D_{\mathrm{KL}}(\hat{P}_m || Q) + \frac{m}{m+l} 2l D_{\mathrm{KL}}(Q || \hat{P}_l) \to \chi^2_{2N}.$$

Критерий близости распределений

Используем статистику $t_{m,l} = \frac{2ml}{m+l} D_{\mathrm{KL}}(\hat{P}_m || \hat{P}_l)$ для проверки гипотезы

$$H_0 : P(x) = P'(x) \text{ при альтернативе } H_1 : P(x) \neq P'(x).$$

Так как критическая область

$$U(\alpha) = \{t : \bar{t}_{1-\alpha} > t \text{ или } t > \bar{t}_\alpha\}, \text{ где } \mathrm{P}(t > \bar{t}_\alpha | H_0) = \alpha,$$

невычислима, предлагается использовать ее приближение

$$U^{\chi^2}(\alpha) = \{t : \bar{t}^{\chi^2}_{1-\alpha} > t \text{ или } t > \bar{t}^{\chi^2}_\alpha\}, \text{ где } \mathrm{P}(t > \bar{t}^{\chi^2}_\alpha | t \sim \chi^2_{2N}) = \alpha.$$

Из следствия 2 ($t_{m,l} \leq \chi^2_{2N}$ при $m, l \to \infty$) получаем

$$\bar{t}_{1-\alpha} < \bar{t}^{\chi^2}_{1-\alpha}, \quad \bar{t}_\alpha < \bar{t}^{\chi^2}_\alpha.$$

Исследование взаимного положения U и U^{χ^2}

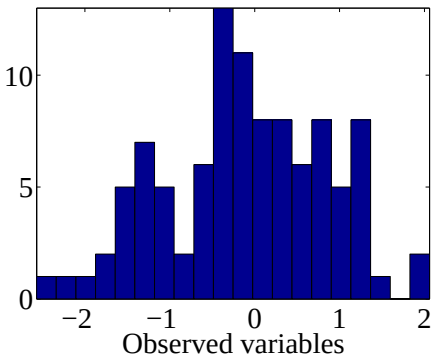
1. $\bar{t}^{\chi^2}_{1-\alpha} < t_{m,l} < \bar{t}_{\alpha}$, тогда $t_{m,l} \notin U$, и $t_{m,l} \notin U^{\chi^2}$.
2. $\bar{t}_{1-\alpha} < t_{m,l} < \bar{t}^{\chi^2}_{1-\alpha}$, тогда $t_{m,l} \notin U$, но $t_{m,l} \in U^{\chi^2}$. Зазор между $\bar{t}_{1-\alpha}$ и $\bar{t}^{\chi^2}_{1-\alpha}$ повышает вероятность ошибки второго рода.

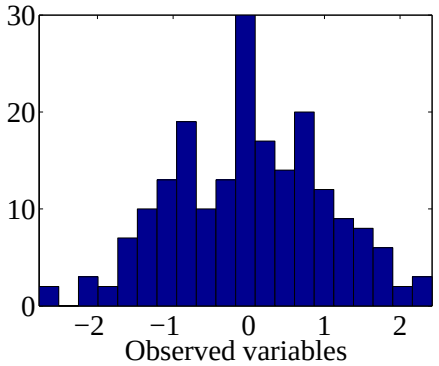
При переходе к одностороннему критерию

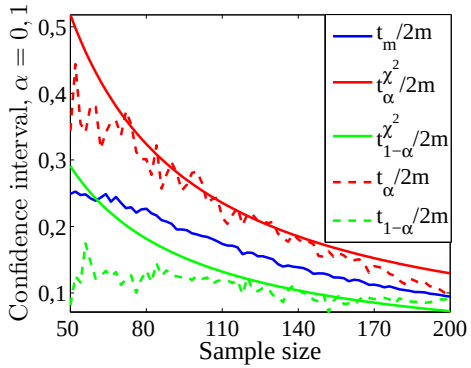
$$U_1(\alpha) = \{t : t > \bar{t}_{\alpha}\}, \quad U_1^{\chi^2}(\alpha) = \{t : t > \bar{t}^{\chi^2}_{\alpha}\},$$

имеем $t_{m,l} \in U_1^{\chi^2} \Rightarrow t_{m,l} \in U_1$.

3. $\bar{t}_{\alpha} < t_{m,l} < \bar{t}^{\chi^2}_{\alpha}$, тогда $t_{m,l} \notin U^{\chi^2}$, но $t_{m,l} \in U$. Зазор между $\bar{t}_{\alpha}$ и $\bar{t}^{\chi^2}_{\alpha}$ повышает вероятность ошибки первого рода.







Theorem (Мотренко, 2014)

Критерий состоятелен:

$$\lim_{m\to\infty} \mathrm{P}(t_m \in U|H_1) = 1.$$

Пусть $P \neq P'$, тогда найдется $x^* \in \mathbb{R} : P(x^*) \neq P'(x^*)$. Тогда найдется $\{a_1, \ldots, a_{N-1}\}$, что для некоторого i

$$P(a_i < x \leq a_{i+1}) = p_i \neq p'_i = P'(a_i < x \leq a_{i+1}).$$

При больших m с $\mathrm{P} > (2\Phi(C_1) - 1)(2\Phi(C_2) - 1)$ выполнено

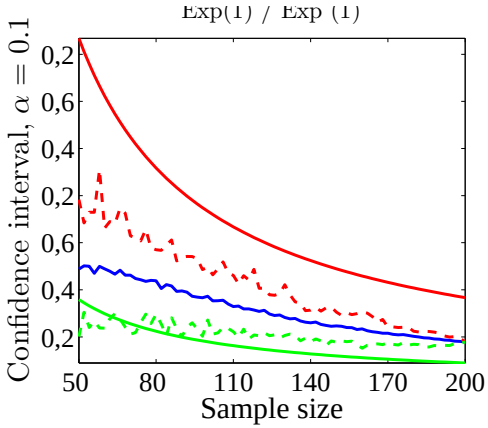
$$|n_i - mp_i| < C_1\sqrt{m}, \quad |n'_i - mp'_i| < C_2\sqrt{m}.$$

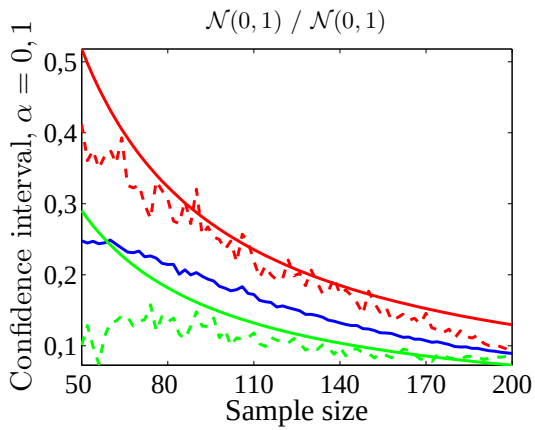
Для любого $\varepsilon > 0$ выберем
$C_1 = C_2 = C_\varepsilon : \mathrm{P} > (2\Phi(C_\varepsilon) - 1)^2 > 1 - \varepsilon$. Тогда

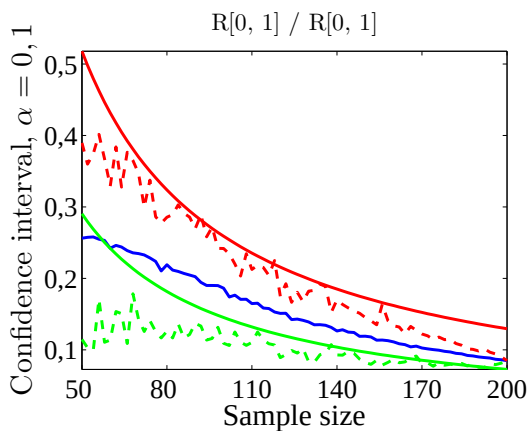
$$\frac{(n_i - n'_i)^2}{n_i} > m\frac{(p_i - p'_i)^2}{p_i} + O(\sqrt{m}) > Cm.$$

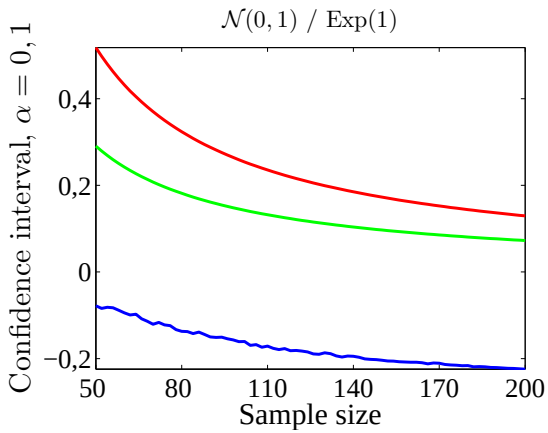
Следовательно, для $\forall\alpha \in (0, 1)\ \mathrm{P}(t_m > \bar{t}_\alpha) \to 1$ при $m \to \infty$.

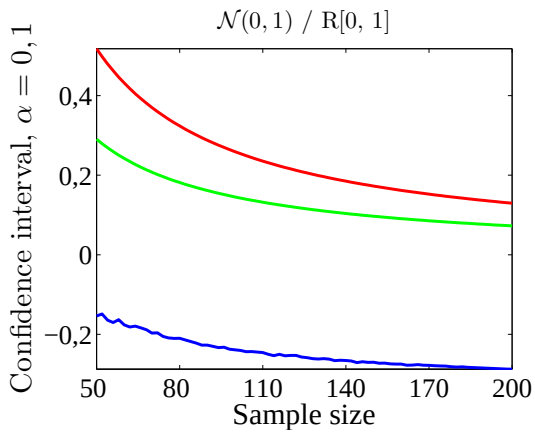
Пример применения критерия к различным парам распределений

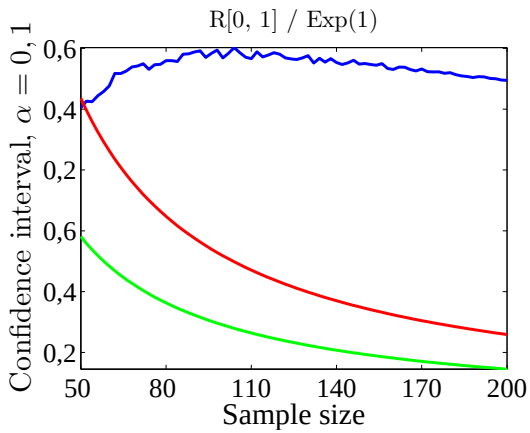






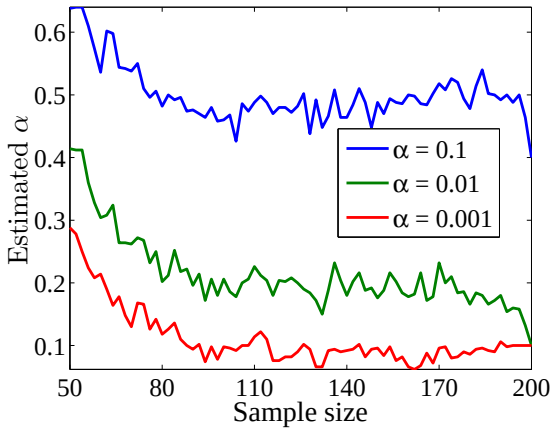




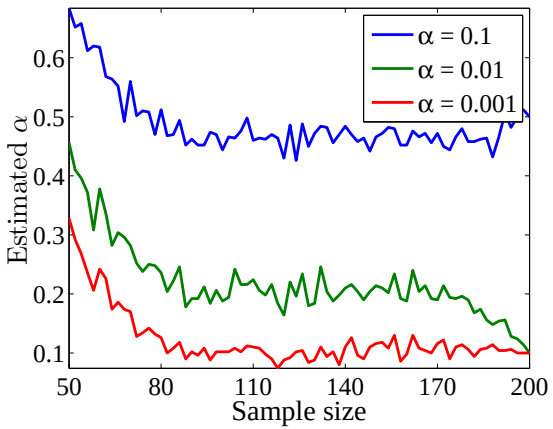


Уровень значимости критерия

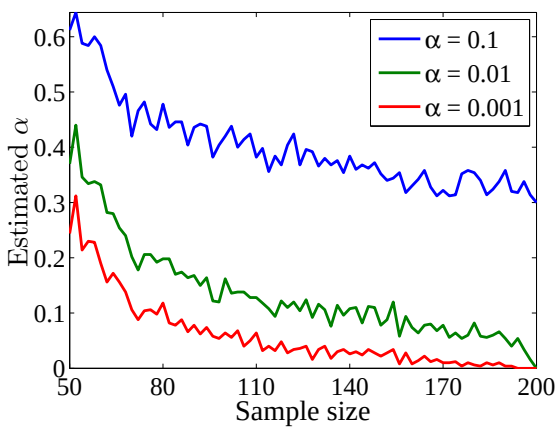
Зависимость фактического уровня значимости от объема выборки различных уровнях значимости критерия χ^2_{2N}.



(o) $Exp(1)||Exp(1)$



(p) $\mathcal{N}(0,1)||\mathcal{N}(0,1)$



(q) $R(0,1)||R(0,1)$

Задача классификации

Дана выборка $D = \{(\mathbf{x}_i, y_i)\}, i = 1, \ldots, m$
$y_i \in \{0, 1\}$, $\mathbf{x}_i \in \mathbb{R}^n$.
Классификатор:

$$a(\mathbf{x}) = [\mathbf{w}^\mathsf{T}\mathbf{x} + c > 0] \in \{0, 1\}.$$

Оптимизация параметров разделяющего

$$\{\mathbf{w}_\mathsf{D}, c_\mathsf{D}\} = \underset{\{\mathbf{w}_\mathsf{D}, c_\mathsf{D}\} \in \mathbb{R}^{n+1}}{\operatorname{argmax}} L_\mathsf{D}(\mathbf{w}_\mathsf{D}, c_\mathsf{D}), \text{ где } L_\mathsf{D} = \ln \prod_{i=1}^{m} p(y_i|\mathbf{x}_i).$$

и порождающего классификаторов:

$$\{\mathbf{w}_\mathsf{G}, c_\mathsf{G}\} = \underset{\{\mathbf{w}_\mathsf{G}, c_\mathsf{G}\} \in \mathbb{R}^{n+1}}{\operatorname{argmax}} L_\mathsf{G}(\mathbf{w}_\mathsf{G}, c_\mathsf{G}), \text{ где } L_\mathsf{G} = \ln L_\mathsf{D}(\mathbf{w}_\mathsf{G}, c_\mathsf{G}) + \ln \prod_{i=1}^{m} p(\mathbf{x}_i).$$

Вероятность $\varepsilon(a_\mathsf{D})$ ошибки классификатора вида $a(\mathbf{x})$:

$$\varepsilon(a) = P \cdot \mathrm{P}(\mathbf{w}^\mathsf{T}\mathbf{x} + c \geq 0 | y = 0) + (1 - P)\mathrm{P}(\mathbf{w}^\mathsf{T}\mathbf{x} + c < 0 | y = 1).$$

Об ошибке разделяющего и порождающего классификаторов в предельных случаях

$$L_{\mathrm{G}} = \ln L_{\mathrm{D}}(\mathbf{w}_{\mathrm{G}}, c_{\mathrm{G}}) + \ln \prod_{i=1}^{m} p(\mathbf{x}_i).$$

Учитывая вид оптимизируемых функционалов L_{D}, L_{G}, имеем:

$$\prod_{i=1}^{m} p_{\mathrm{D}}(y_i|\mathbf{x_i}) > \prod_{i=1}^{m} p_{\mathrm{G}}(y_i|\mathbf{x}_i),$$

и при $m \to \infty$ с достаточно большой вероятностью

$$\mathbf{w}_{\mathrm{D}}^{\mathsf{T}}\mathbf{x} + c_{\mathrm{D}} < \mathbf{w}_{\mathrm{G}}^{\mathsf{T}}\mathbf{x} + c_{\mathrm{G}} \quad \text{для } y = 1,$$

$$\mathbf{w}_{\mathrm{D}}^{\mathsf{T}}\mathbf{x} + c_{\mathrm{D}} > \mathbf{w}_{\mathrm{G}}^{\mathsf{T}}\mathbf{x} + c_{\mathrm{G}} \quad \text{для } y = 0.$$

Следовательно, при $m \to \infty$ имеем $\varepsilon_{\mathrm{D}} < \varepsilon_{\mathrm{G}}$.

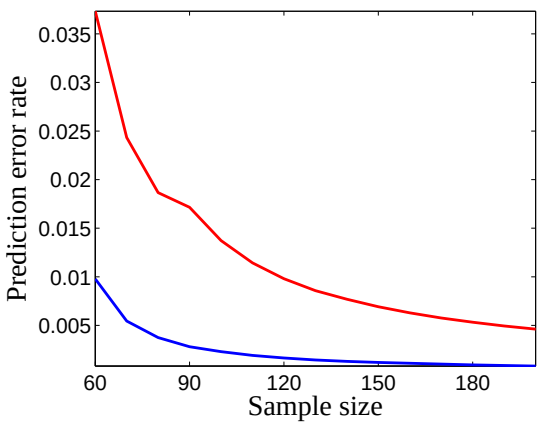
Пусть $m = 1, D = \{(\mathbf{x}, 1)\}$. Тогда

$w_{\mathrm{D}} = \operatorname{argmax} L_D = \mathbf{w}^{\mathsf{T}}\mathbf{x} + c \to \infty$. С другой стороны, слагаемое $\ln \prod_{i=1}^{m} p(\mathbf{x}_i)$ в определении функционала L_{G} выступает в роли регуляризатора, поэтому при $m \approx 1$ имеем $\varepsilon_{\mathrm{G}} < \varepsilon_{\mathrm{D}}$.

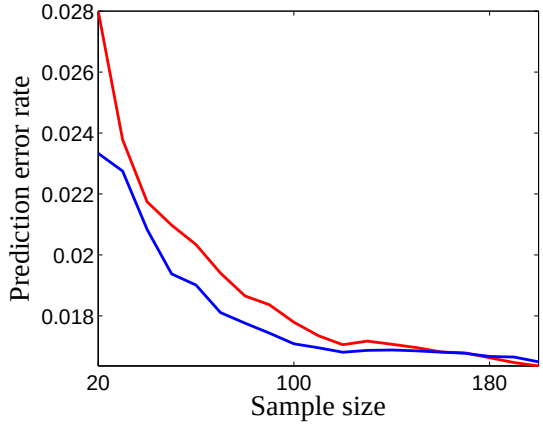
Сравнение $\varepsilon_D(m)$ и $\varepsilon_G(m)$ на синтетических данных

В качестве приближения вероятности ошибки ε рассмотрим частоту ошибки каждого классификатора a на выборке D:

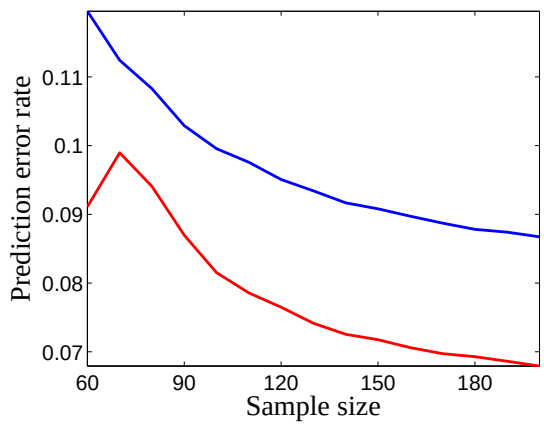
$$\hat{\varepsilon}_m(a) = \frac{1}{m}\sum_{i=1}^{m}[a(\mathbf{x}_i) \neq y_i].$$



(r) $\sigma = 0.1$



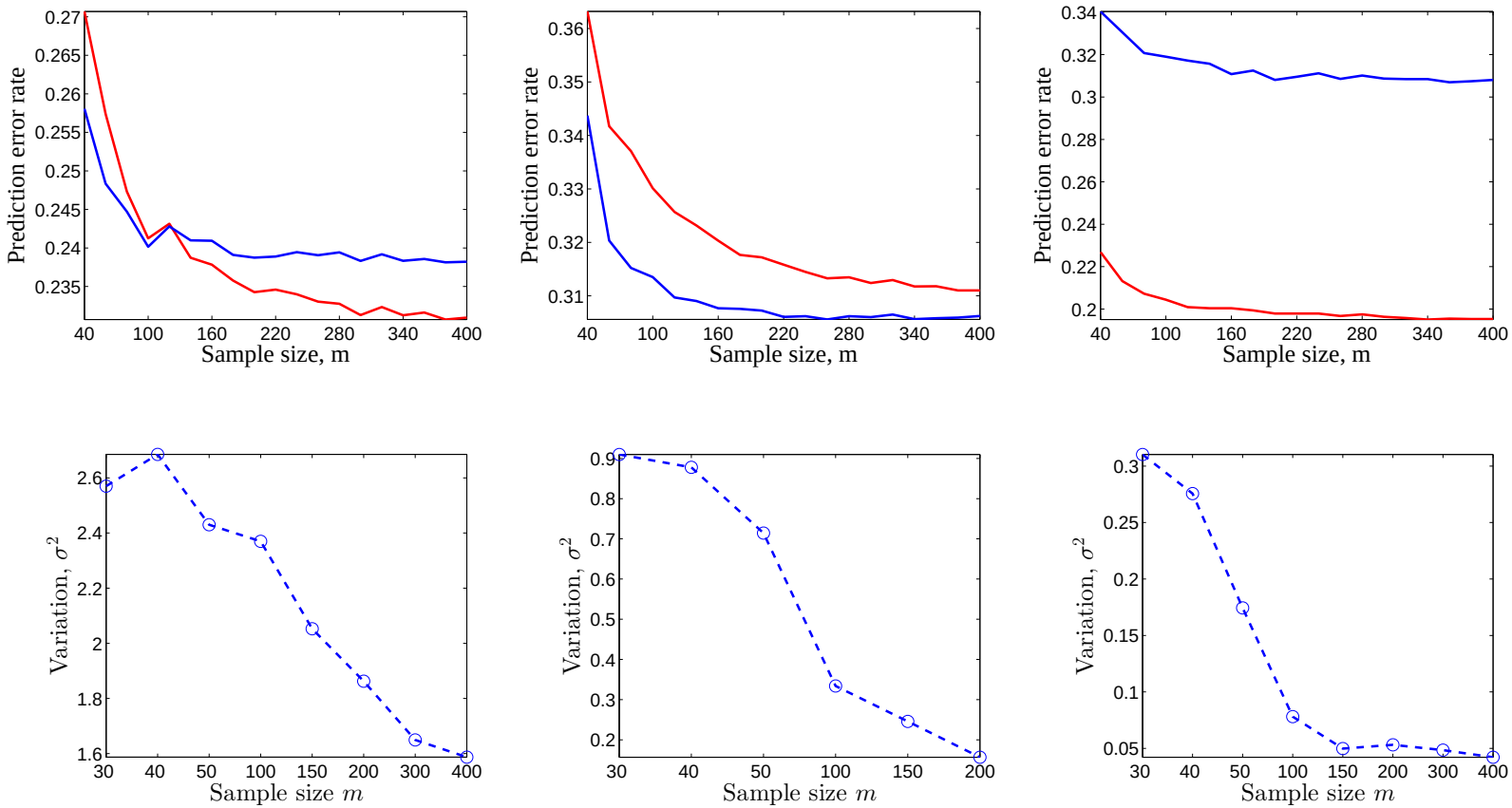
(s) $\sigma = 0.2$



(t) $\sigma = 0.3$

Красный — логистическая регрессия, синий — наивный Байес.

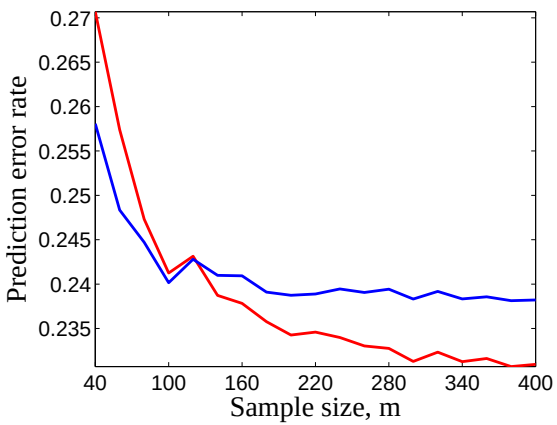
Сравнение $\varepsilon_D(m)$ и $\varepsilon_G(m)$ на синтетических данных

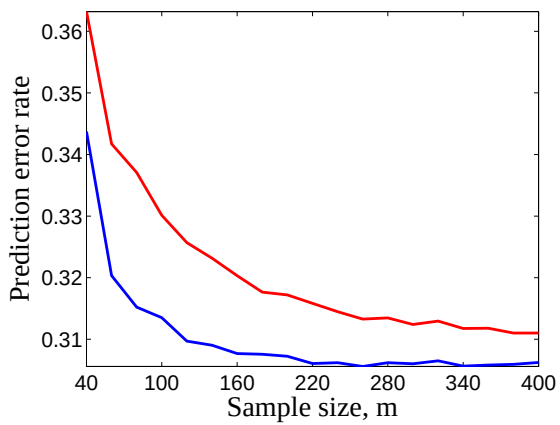


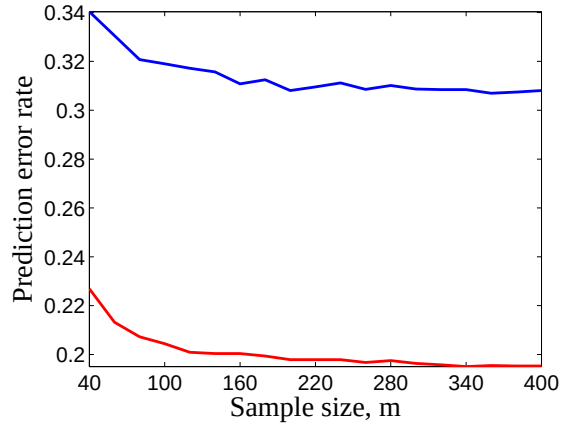
(x) $P = 0.6$ **(y)** $P = 0.7$ **(z)** $P = 0.8$

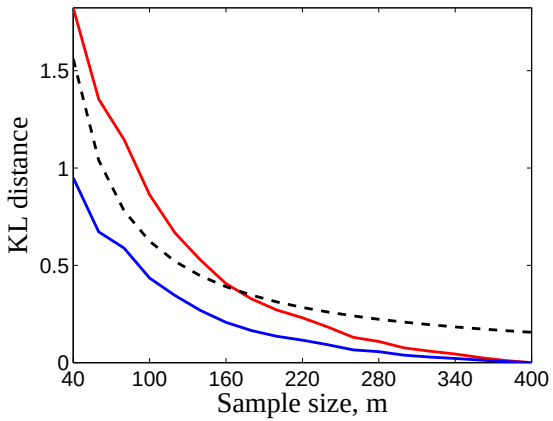
Зависимость объема выборки, при котором происходит пересечение кривых $\varepsilon_D(m)$ и $\varepsilon_G(m)$ от дисперсии σ^2 для синтетических данных с фиксированными $\mu_1 = 1$, $\mu_0 = 0$.

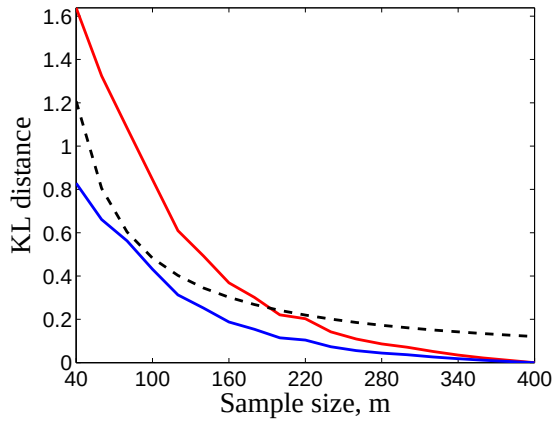
Оценка объема выборки

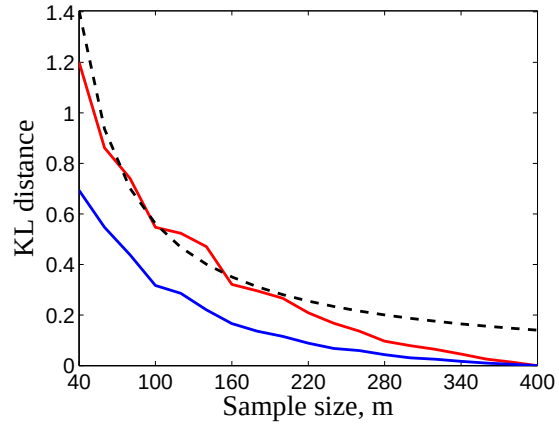












Результаты

- Получены предельные оценки сверху распределения расстояния Кульбака-Лейблера между гистограммами из одного распределения.
- Предложен способ оценки достаточности объема выборки на основе расстояния Кульбака-Лейблера.
- Продемонстрирована возможность использования предложенного метода оценки объема выборки для выбора между порождающим и разделяющим подходами к классификации.

1. *А. П. Мотренко, В. В. Стрижов.* Построение агрегированных прогнозов объемов железнодорожных грузоперевозок с использованием расстояния Кульбака–Лейблера. Информатика и ее применениия, 2014. (принято в печать)
2. *A. Motrenko, V. Strijov and G.-W. Weber.* Sample Size Determination for Logistic Regression. Journal of Computational and Applied Mathematics, 2014, 255(743-752).
3. *А. П. Мотренко.* Оценка плотности совместного распределения. Машинное обучение и анализ данных, 2012. 1-4(428-436).
4. *Вальков А.С., Кожанов Е.М., Мотренко А.П., Хусаинов Ф.И.* Построение кросс-корреляционных зависимостей при прогнозе загруженности железнодорожного узла. Машинное обучение и анализ данных, 2013. 1-5(505-518).