Оценка объема выборки в задачах классификации и прогнозирования

А. П. Мотренко, В. В. Стрижов

Московский физико-технический институт,
ВЦ ФИЦ ИУ РАН

Математические методы распознавания образов 17
г. Светлогорск, 23 сентября 2015.

Цель: разработать метод оценки объема выборки, необходимого для получения статистически достоверных результатов классификации.

Задача: выбрать оптимальную модель к решению задачи классификации — порождающую, разделяющую, либо комбинированную.

Предлагается для выбора оптимальной модели при решении задач классификации оценить объем выборки в рамках каждой из моделей.

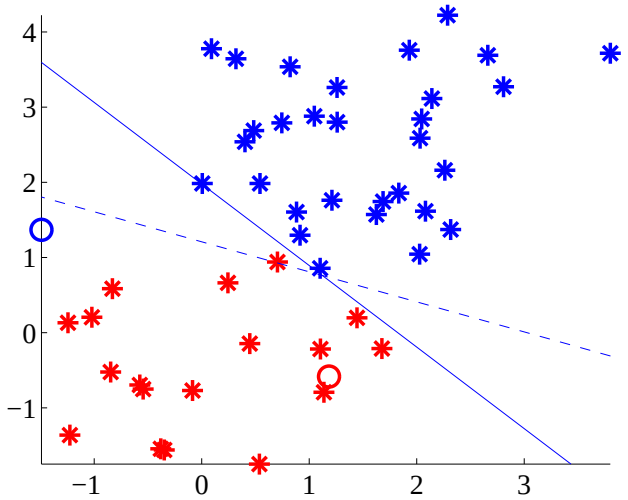
Существующие методы оценки объема выборки

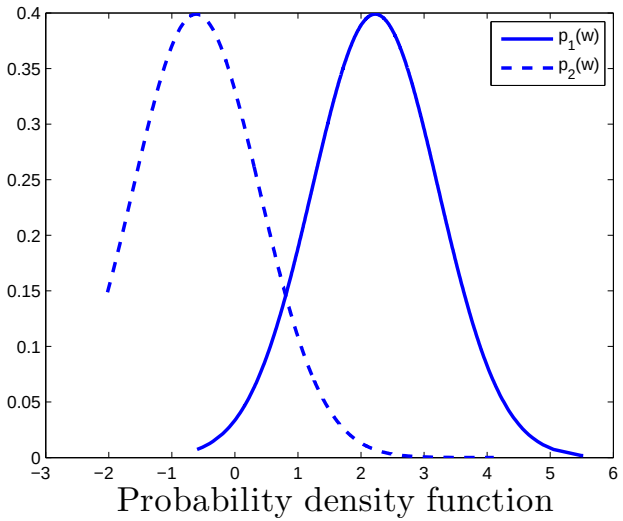
Подход	Формула
Метод доверительных интервалов $Z = \frac{\bar{X}-\mu}{\sigma}\sqrt{m} \to \mathcal{N}(0,1)$ при $H_0 : EX = \mu$	$m = \left(\frac{z_{\alpha/2}\sigma}{\bar{X}-\mu}\right)^2$
Тест на равенство: $Z = \frac{\hat{p}-p_0}{\sqrt{\hat{p}(1-\hat{p})}}\sqrt{m} \to \mathcal{N}(0,1)$ при $H_0 : p = p_0$ против $H_1 : p \neq p_0$	$m = \frac{(z_{\mathbf{Pow}}+z_{\alpha/2})^2 p(1-p)}{(\hat{p}-p_0)^2}$
Тест отношения правдоподобия: $\gamma_m : \chi^2_{p,1-\mathrm{Pow}}(\gamma_m) = \chi^2_{p,\alpha}$	$m = \frac{\gamma_m}{\Delta^*}$, где $\Delta^* = \mathrm{E}_X\left[\frac{-X(\beta-\beta^*)}{1+e^{-X\beta}} - \log\left(\frac{1+e^{-X\beta}}{1+e^{-X\beta^*}}\right)\right]$
Статистика Вальда: $Z = \frac{\hat{\beta}-\beta^0}{\sqrt{\hat{v}}}\sqrt{m} \to \mathcal{N}(0,1)$ при $H_0 : \beta = \beta^0$	$\hat{m} = \frac{\left(\sqrt{V_1}z_{\mathbf{Pow}} - \sqrt{V_0}z_{\alpha/2}\right)^2}{(\beta^1-\beta^0)^2}$
Заданная точность регрессии: $\hat{\beta}_j = t_{1-\alpha/2}(m-n-1)\sqrt{\frac{1-\mathbf{R}^2}{(1-\mathbf{R}^2_j)(m-n-1)}}$	$m^* = \frac{z^2_{\alpha/2}}{\delta^2}\left(\frac{1-\mathbf{R}^2}{1-\mathbf{R}^2_j}\right)\left(\frac{\chi^2_{1-\gamma}(m-1)}{m-n-1}\right) + n + 1$, где $\mathbf{R} = \boldsymbol{\rho}'_{yx}\mathbf{R}^{-1}_{xx}\boldsymbol{\rho}_{yx}$
С помощью метода Bootstrap	$m = \left(\frac{z_{\alpha/2}\sigma}{\bar{X}-\mu}\right)^2$ и $m = \frac{z^2_{\alpha/2}}{(\bar{X}-\mu)^2}\left(\frac{1-R^2}{1-R^2_j}\right) + n$

Пример выборки недостаточного объема

Дана выборка $Z = \{\mathbf{x}_i, y_i\}, i = 1, \ldots, m$, где $y_i \in \{0, 1\}$ — метка класса объекта, $\mathbf{x}_i \in \mathbb{R}^n$ — признаковое описание объекта.

При изменении состава выборки существенно изменяются параметры модели $\mathbf{x}^\mathsf{T}\mathbf{w} + c = 0$ и оценка априорного распределения $p(\mathbf{w})$.





Назовем объем m^* выборки из распределения P **достаточным**, если для всех выборок X_1, X_2 объема $m > m^*$ из P распределения $\hat{p}_1(\mathbf{w})$ и $\hat{p}_2(\mathbf{w})$ близки согласно некоторой функции расстояния $D(\hat{p}_1||\hat{p}_2)$.

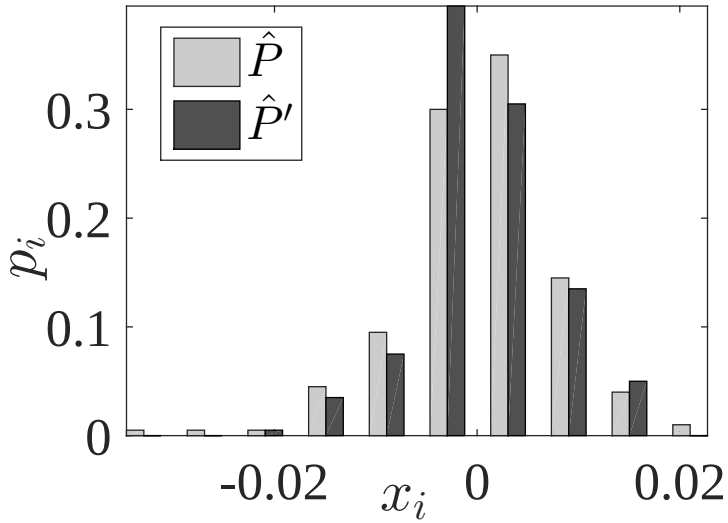
Непараметрический случай: сравнение гистограмм

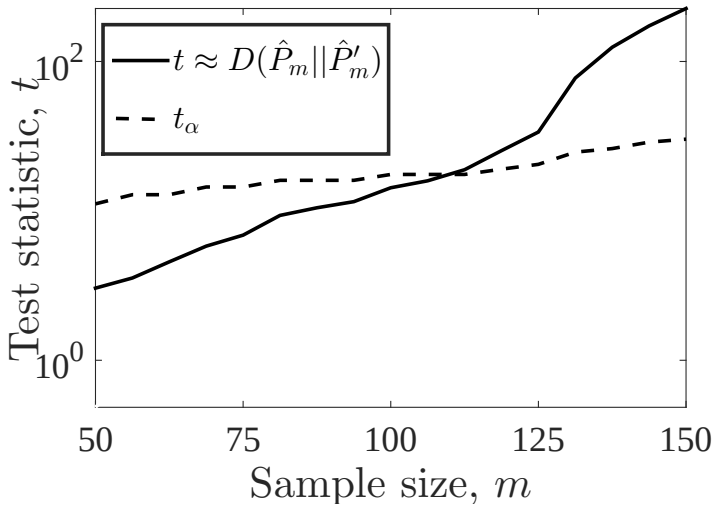
Рассмотрим выборку Z объема m и разбиение множества значений Z на N интервалов (a_i, a_{i+1}). Набор оценок

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

вероятности $p_i = P(a_i < z \leq a_{i+1})$ назовем гистограммой $\hat{P}_m$.

Требуется ввести меру сходства гистограмм $D(\hat{P}_m || \hat{P}'_m)$ и найти $m^* : D(\hat{P}_m || \hat{P}'_m) > D_0$:





f-Дивергенции между гистограммами

f-дивергенция[a] между распределениями Q и P:

$$D_f(Q||P) = \sum P \cdot f\left(\frac{Q}{P}\right).$$

Дивергенция Кульбака-Лейблера: $f(t) = t \ln t$, $D_{\mathrm{KL}}(Q||P) = \sum Q \ln\left(\frac{Q}{P}\right)$.

Свойства D_f (для $f(t)$, строго выпуклых и дважды дифференцируемых в $t = 1$):

1. Дивергенция $D_f(Q, P)$ определена на всех парах распределений с одинаковым носителем.
2. $D_f(Q, P)$ достигает минимума при $P = Q$.
3. При $m \to \infty$ выполнено

$$\frac{2m}{f''(1)} \cdot D_f(\hat{P}_m||P) \to \chi^2_N,$$

где $\hat{P}_m$ — гистограмма, построенная по выборке X_m из распределения P.

[a] 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 (Methodological). 1(28):131-142.

Утверждение

Пусть

(a) функция f строго выпукла и дважды дифференцируема в единице,

(b) распределение $P : \mathbb{R} \to [0, 1]$ таково, что для всех $x \in \mathbb{R}$ $P(x) \neq 0$,

(c) выборки Z и Z' выбираются независимо из распределения P, причем $\frac{1}{\rho} \leq m'/m \leq \rho$ при всех m, m' для некоторого $0 < \rho \leq 1$.

Тогда случайная величина $\frac{2m}{f''(1)} \cdot D_f(\hat{P}_m || \hat{P}_{m'})$ в пределе ограничена сверху и снизу случайными величинами из распределения χ^2_N

$$C_1\chi^2_N \leq \frac{2m}{f''(1)} \cdot D_f(\hat{P}_m || \hat{P}_{m'}) \leq C_2\chi^2_N, \text{ при } m,\ m' \to \infty.$$

Аппроксимация D_{KL} с помощью χ^2

Схема доказательства: при $P \to Q$: $P(x)f\left(\frac{Q(x)}{P(x)}\right) =$

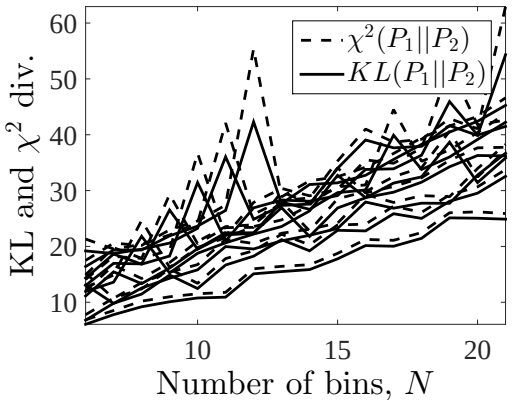
$$P(x)f\left(\frac{Q(x)}{P(x)}\right) \quad + \quad \to 0, \text{ т.к. } f(1) = 0$$

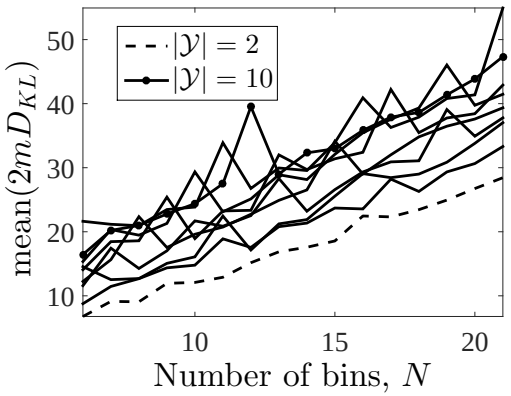
$$f'(1)(Q(x) - P(x)) \quad + \quad \text{суммируется по } x \text{ к } 0,$$

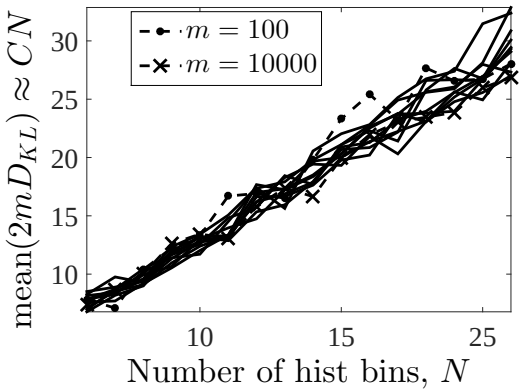
$$\frac{f''(1)}{2}\frac{(Q(x) - P(x))^2}{P(x)} \quad + \quad P(x)o\left(\left(\frac{Q(x)}{P(x)} - 1\right)^3\right),$$

тогда $$D_f(\hat{P}_m||P) = \sum_{i=1}^{N} p_i f\left(\frac{\hat{p}_i}{p_i}\right) \approx \frac{f''(1)}{2m}\sum_i \frac{(n_i - mp_i)^2}{mp_i}.$$

Зависимость параметров аппрокимации C_1, C_2 от размера выборки m и классов $|\mathcal{Y}|$.



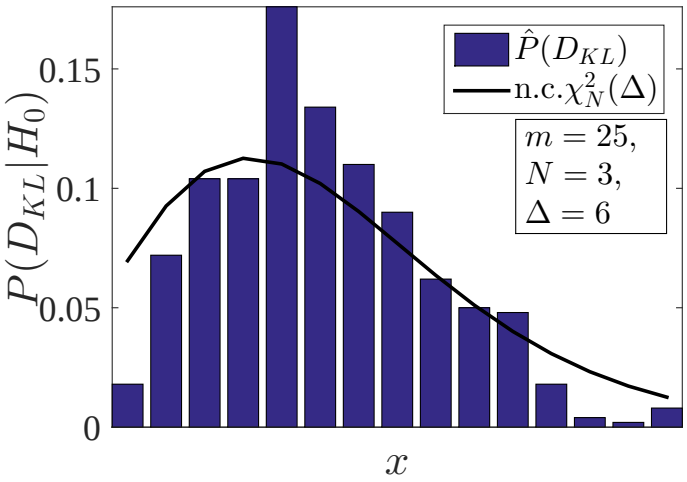


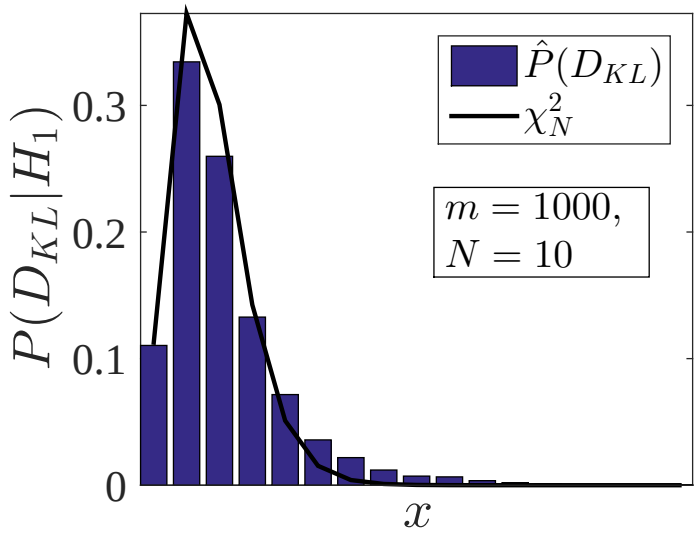


Для оценки объема выборки сформулируем задачу двух выборок:

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

- При H_1 с.в. $C_1'\chi^2_N(\Delta) \leq D_{\mathrm{KL}}(\hat{P}_m, \hat{P}'_m) \leq C_2'\chi^2_N(\Delta)$, где $\chi^2(\Delta)$ — нецентральное хи-квадрат распределение с параметром нецентральности $\Delta = \frac{1}{\mathbb{D}p_i}\sum_{i=1}^{N}\left[\mathbb{E}(p_i - p_i')\right]^2$.
- При H_1 с.в. $C_1\chi^2_N \leq D_{\mathrm{KL}}(\hat{P}_m, \hat{P}'_m) \leq C_2\chi^2_N$.





Оценка объема выборки

Для оценки объема выборки m^*: $2mD_{\mathrm{KL}}(\hat{P}_m||\hat{P}_{m'}) = t_{m^*} \in U(\alpha)$
приблизим критическую область

$$U(\alpha) = \{t : t < t_\alpha\}, \text{ где } \mathrm{P}(t < t_\alpha|H_0) = \alpha,$$

с помощью $U^{\chi^2}(\alpha) = \{t : t < n.c.\chi^2_\alpha\}$, где $\mathrm{P}\big(t > \chi^2_\alpha|t \sim \chi^2_N(\Delta)\big) = \alpha$.

Для соблюдения условия $\mathrm{P}(H_0|H_1) = \mathrm{P}(t_m > t_\beta|H_1) \geq 1 - \beta$
приблизим t_β квантилью $\chi^2_{N,\beta}$.

Из предельных неравенств следует:
$t_\alpha < \chi^2_{N,\alpha}(\Delta) \Rightarrow U(\alpha) \subseteq U^{\chi^2}(\alpha)$, т.е. $\hat{\alpha} < \alpha$,
$t_\beta < \chi^2_{N,\beta} \Rightarrow \hat{\beta} < \beta$.

Тогда

$$m^* : \ \chi^2_{N,\beta} < t_{m^*} < \chi^2_{N,\alpha}(\Delta).$$

Задача классификации:

$$\hat{y} = \underset{y \in \{0,1\}}{\operatorname{argmax}} p(y|\mathbf{x}, \mathbf{w}).$$

• Разделяющая модель: параметры $\mathbf{w}$ максимизируют условное правдоподобие

$$\mathbf{w} = \mathbf{w}_{\mathsf{D}} = \underset{\mathbf{w} \in \mathcal{W}}{\operatorname{argmax}} L_{\mathsf{D}}(\mathbf{w}; \mathbf{y}, X), \quad L_{\mathsf{D}}(\mathbf{w}) = \ln p(\mathbf{y}|X, \mathbf{w}).$$

• Порождающая модель: параметры $\mathbf{w}$ максимизируют совместное правдоподобие

$$\mathbf{w} = \mathbf{w}_{\mathsf{G}} = \underset{\mathbf{w} \in \mathcal{W}}{\operatorname{argmax}} L_{\mathsf{G}}(\mathbf{w}; \mathbf{y}, X), \quad L_{\mathsf{G}}(\mathbf{w}) = \ln p(\mathbf{y}, X|\mathbf{w}).$$

- Разделяющий классификатор имеет вид $\hat{y} = a_{\mathsf{D}}(\mathbf{x}) = [\mathbf{w}_D^{\mathsf{T}}\mathbf{x} > 0]$.

$\mathbf{y} = [y_1, \ldots, y_m]^T$ — случайный вектор с независимыми компонентами y_i:

$$p(y_i = 1|\mathbf{x}_i, \mathbf{w}) = \frac{1}{1 + \exp(-\mathbf{x}_i^{\mathsf{T}}\mathbf{w})},$$

- Порождающий классификатор: $\hat{y} = a_{\mathsf{G}}(\mathbf{x}) = [\mathbf{w}_{\mathsf{G}}^{\mathsf{T}}\mathbf{x} > 0]$, где $\mathbf{w} = [c, \boldsymbol{\beta}]$.

Параметры $\boldsymbol{\beta} = \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$, $c = \ln \frac{P}{1-P} - \frac{1}{2}\boldsymbol{\beta}^T(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0)$ максимизируют совместное правдоподобие

$$\ln p(\mathbf{y}, X|\mathbf{w}) = L_G(\boldsymbol{\mu}_1, \boldsymbol{\mu}_0, \Sigma, P) \to \max_{\boldsymbol{\mu}_1, \boldsymbol{\mu}_0, \Sigma, P},$$

Предполагается, что $\mathbf{x}|y \sim \mathcal{N}(\mu_y, \Sigma)$, т.е.

$$p(\mathbf{x}|y) = \frac{1}{\sqrt{(2\pi)^n|\Sigma^{-1}|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_y)^T\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}_y)\right).$$

тогда имеет место

$$\frac{p(1|\mathbf{x})}{p(0|\mathbf{x})} = \frac{Pp(\mathbf{x}|1)}{(1-P)p(\mathbf{x}|0)} = \frac{P}{1-P}\exp\left(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{x} + \tilde{c}\right) = \exp\left(\mathbf{w}^{\mathsf{T}}\mathbf{x}\right),$$

Вероятность ошибки разделяющего и порождающего подходов

Оценим вероятность ошибки ε каждого классификаторов a_{D} и a_{G} частотой ошибок на выборке $Z = \{\mathbf{x}_i, y_i\}_{i=1}^m$ *:

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

• При $m \to \infty$ выполняется $\varepsilon_{\mathsf{G}} > \varepsilon_{\mathsf{D}}$.
Учитывая соотношение $\ln L_{\mathsf{G}}(\mathbf{w}) = \ln L_{\mathsf{D}}(\mathbf{w}) + \ln \prod_{i=1}^m p(\mathbf{x}_i)$, получаем:

$$\prod_{i=1}^{m} p(y_i|\mathbf{x_i}, \mathbf{w}_{\mathsf{D}}) > \prod_{i=1}^{m} p(y_i|\mathbf{x}_i, \mathbf{w}_{\mathsf{G}}),$$

поэтому при $m \to \infty$ с высокой вероятностью выполняется

$$a_{\mathsf{D}}(\mathbf{x}) > a_{\mathsf{G}}(\mathbf{x}) \;\text{ для } y = 1,\; a_{\mathsf{D}}(\mathbf{x}) < a_{\mathsf{G}}(\mathbf{x}) \;\text{ для } y = 0.$$

• При $m \approx 1$ за счет регуляризатора $\ln \prod_{i=1}^m p(\mathbf{x}_i)$ в L_{G} выполняется $\varepsilon_{\mathsf{G}} < \varepsilon_{\mathsf{D}}$.

---

* A. Y. Ng, M. I. Jordan. On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes, 2002.

Достоинства моделей:

- Разделяющие модели более точны $\varepsilon_D < \varepsilon_G$ в случае если априорное распределение $p(X)$ плохо описывает истинное распределение данных.
- При малых m, $\varepsilon_D > \varepsilon_G$; порождающие модели также позволяют учитывать неразмеченные данные.

Линейная комбинация:

$$\mathbf{w}_\lambda = \underset{\mathbf{w}}{\operatorname{argmax}} L_\lambda(\mathbf{w}), \quad L_\lambda(\mathbf{w}) = \lambda L_D(\mathbf{w}) + (1-\lambda) L_G(\mathbf{w})$$

- $\lambda \to 1 \Leftrightarrow \mathbf{w} \to \mathbf{w}_D$;
- $\lambda \to 0 \Leftrightarrow \mathbf{w} \to \mathbf{w}_G$.

Комбинированная вероятностная модель

Построим комбинированную модель с параметрами $\mathbf{w}_\lambda = (\mathbf{w}_\mathsf{D}, \mathbf{w}_\mathsf{G})$, где

$$\mathbf{w}_\lambda = \underset{\mathbf{w}}{\operatorname{argmax}} \ln p(X, \mathbf{y}, \mathbf{w}_\lambda),$$

$$p(X, \mathbf{y}, \mathbf{w}_\lambda) = p(\mathbf{w}_\mathsf{D}, \mathbf{w}_\mathsf{G}) \left[\prod_{i\in 1}^{n} p(y_i|\mathbf{x}_i, \mathbf{w}_\mathsf{D})\right] \left[\prod_{i\in 1}^{n} p(\mathbf{x}_i|\mathbf{w}_\mathsf{G})\right].$$

Задание нормального апостериорного распределения параметров $\mathbf{w}_\lambda$

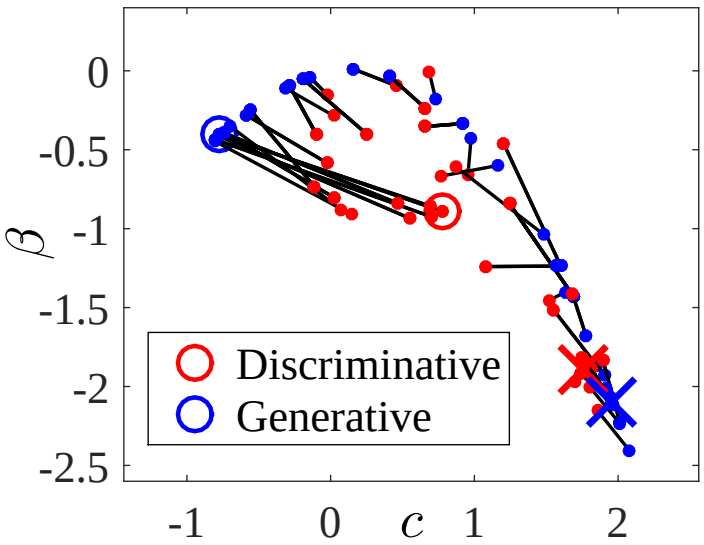
$$p(\mathbf{w}_\mathsf{D}, \mathbf{w}_\mathsf{G}) \propto p(\mathbf{w}_\mathsf{D}) p(\mathbf{w}_\mathsf{G}) \frac{1}{\sigma} e^{\left(-\frac{1}{2\sigma^2}||\mathbf{w}_\mathsf{D} - \mathbf{w}_\mathsf{G}||^2\right)}, \quad \sigma(\lambda) = \left(\frac{\lambda}{1-\lambda}\right)^2.$$

позволяет параметризовать $\lambda \in [0, 1]$ переход от разделяющей к порождающей модели:

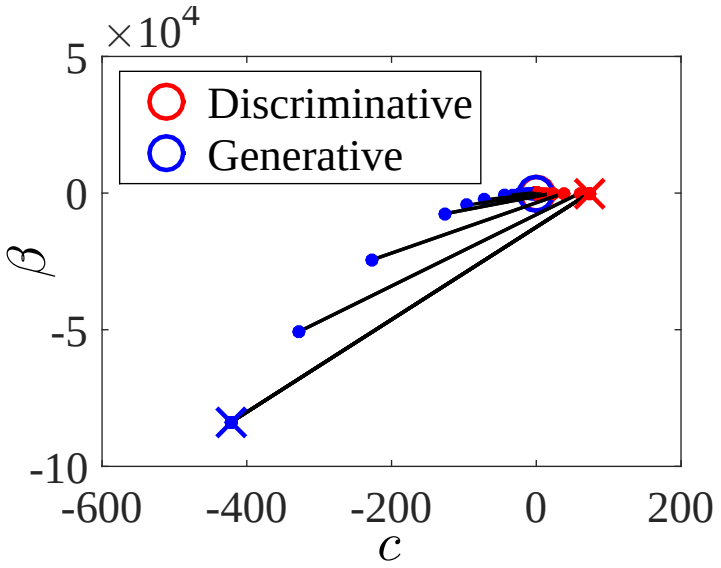
- $\lambda \to 0 \Rightarrow \sigma(\lambda) \to 0 \Rightarrow p(\mathbf{w}_\mathsf{D}, \mathbf{w}_\mathsf{G}) \approx p(\mathbf{w}_\mathsf{D})\delta(\mathbf{w}_\mathsf{D} - \mathbf{w}_\mathsf{G})$.
- $\lambda \to 1 \Rightarrow \sigma(\lambda) \to \infty \Rightarrow p(\mathbf{w}_\mathsf{D}, \mathbf{w}_\mathsf{G}) \approx p(\mathbf{w}_\mathsf{D}) p(\mathbf{w}_\mathsf{G})$.

Оценка параметров $\mathbf{w}_\lambda$

Итерации настройки параметров $\mathbf{w} = [\beta, c]$ при экстремальных значениях параметра λ без дополнительной регуляризации (бесконечная дисперсия априорных распределений $p(\mathbf{w})$).

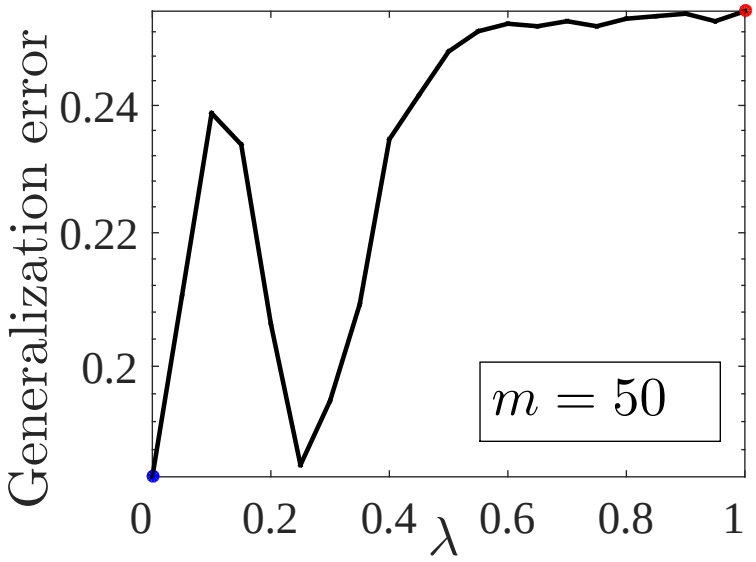


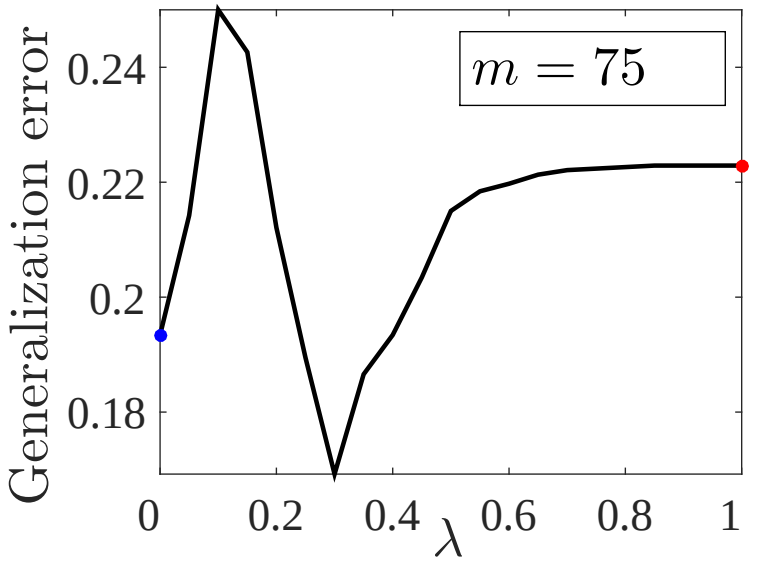
$\lambda = 0$

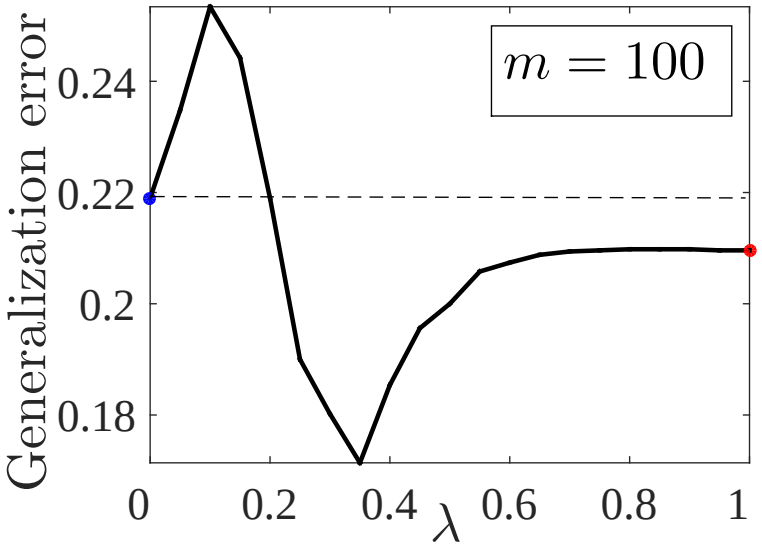


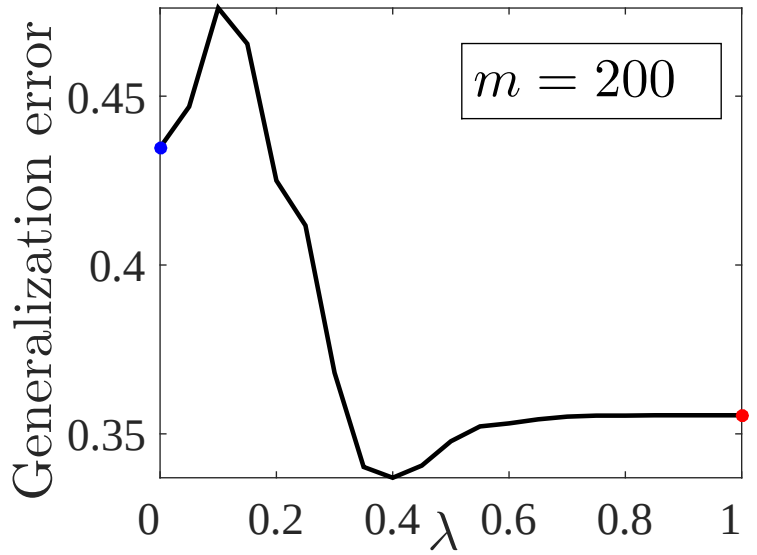
$\lambda = 1$

Пример: ошибка классификации $\hat{\varepsilon}_m(\lambda)$, $x \sim \mathcal{N}(\mu_y, \sigma_y^2)$









Оценка объема выборки

$$D_{\mathrm{KL}}(\lambda) = \int_{\mathbf{w}} p(\mathbf{w}|D_1, \alpha, \lambda) \ln \frac{p(\mathbf{w}|D_1, \alpha, \lambda)}{p(\mathbf{w}|D_2, \alpha, \lambda)} d\mathbf{w}.$$

Выражение для $p(\mathbf{w}|D, \alpha, \lambda)$ найдем с помощью формулы Байеса

$$p(\mathbf{w}_\lambda|D, \alpha, \lambda) = \frac{p(D|\mathbf{w}_\lambda)p(\mathbf{w}_\lambda|\alpha, \lambda)}{p(D|\alpha, \lambda)},$$

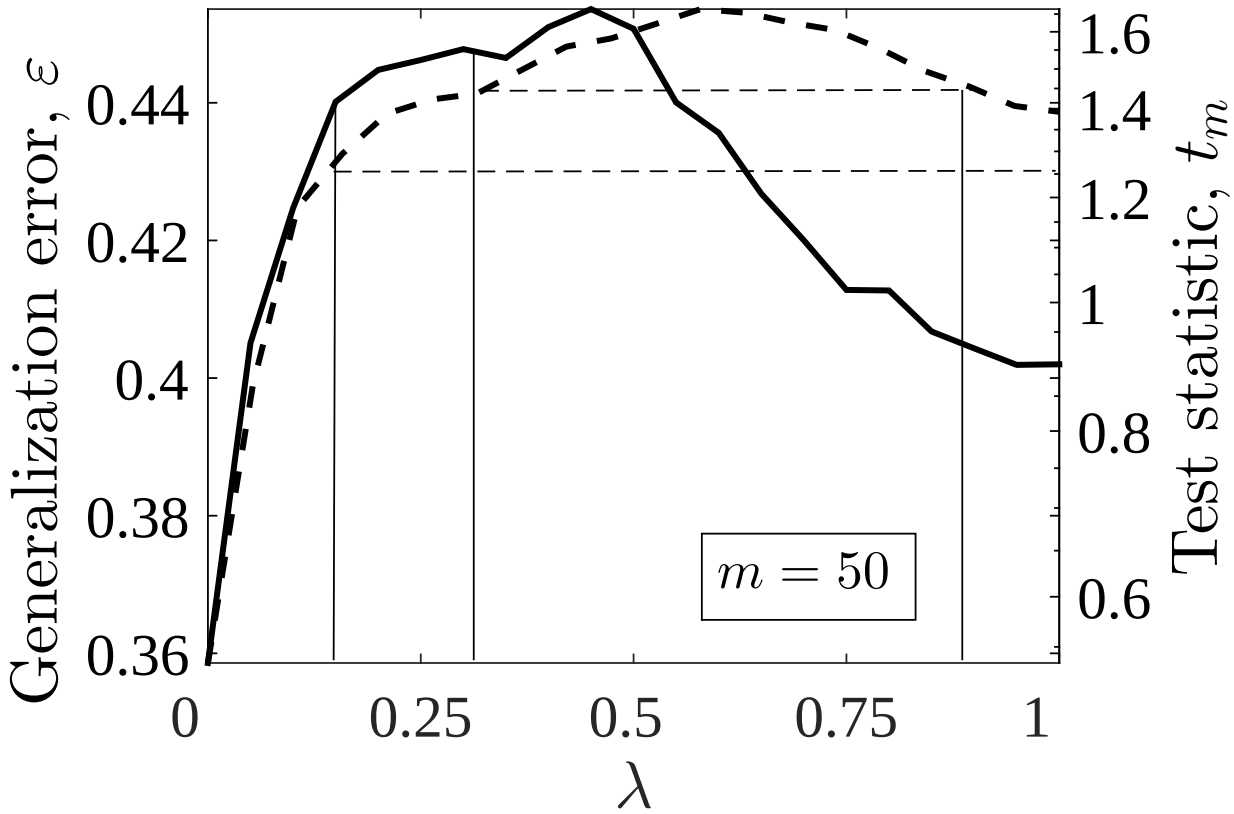
где $p(D|\mathbf{w}_\lambda) = p(\mathbf{y}|X, \mathbf{w}_\mathsf{D})p(X|\mathbf{w}_\mathsf{G})$ — правдоподобие данных, $p(\mathbf{w}_\lambda|\alpha, \lambda) = p(\mathbf{w}_\mathsf{D}, \mathbf{w}_\mathsf{G}|\alpha, \lambda)$ — плотность распределения параметров модели, и

$$p(D|\alpha, \lambda) = \int p(D|\mathbf{w}_\lambda)p(\mathbf{w}_\lambda|\alpha, \lambda)d\mathbf{w}.$$

Для определения m^*

- в разделяющем случае положим
  $p(D|\mathbf{w}_\lambda) = p(\mathbf{y}|X, \mathbf{w}_\mathsf{D})$, $p(\mathbf{w}_\lambda|\alpha, \lambda) = p(\mathbf{w}_\mathsf{D}|\alpha)$;
- в порождающем случае:
  $p(D|\mathbf{w}_\lambda) = p(\mathbf{y}, X|\mathbf{w}_\mathsf{G})$, $p(\mathbf{w}_\lambda|\alpha, \lambda) = p(\mathbf{w}_\mathsf{G}|\alpha)$.

Сравнение $D_{\mathsf{KL}}(\lambda)$ и $\varepsilon_m(\lambda)$



m	50	75	100	200	500
$\cos\big(D_{\mathsf{KL}}(\lambda), \varepsilon_m(\lambda)\big)$	0.916	0.933	0.926	0.969	0.8622

Результаты

- Рассмотрена задача оценки объема выборки в задаче классификаци с учетом используемой модели.
- Предложен способ оценки объема выборки, основанный на анализе распределения параметров модели.
- Предложенный способ проиллюстрирован задачей выбора между порождающим и разделяющим подходами к решению задачи классификации.

А. П. Мотренко, В. В. Стрижов. Построение агрегированных прогнозов объемов железнодорожных грузоперевозок с использованием расстояния Кульбака–Лейблера. Информатика и ее примененииия, 2014, 8(2), 86-97.
A. Motrenko, V. Strijov and G.-W. Weber. Sample Size Determination for Logistic Regression. Journal of Computational and Applied Mathematics, 2014, 255, 743-752.