

Обзор статьи Revisiting Classifier Two-Sample Tests

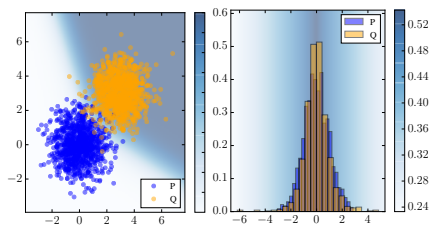
Владислав Скрипнюк

Октябрь 2017

Two sample tests

- ▶ Даны две выборки

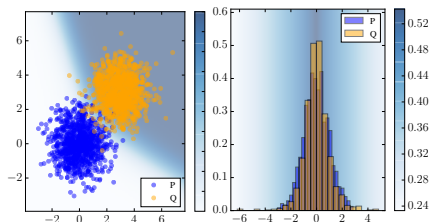
$$S_P \sim P^n \text{ и } S_Q \sim Q^m$$



Two sample tests

- ▶ Даны две выборки

$$S_P \sim P^n \text{ и } S_Q \sim Q^m$$



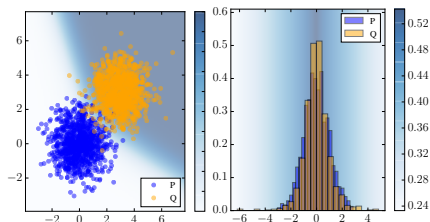
- ▶ Необходимо проверить гипотезу

$$P = Q ?$$

Two sample tests

- ▶ Даны две выборки

$$S_P \sim P^n \text{ и } S_Q \sim Q^m$$

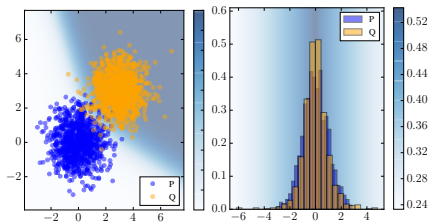


- ▶ Необходимо проверить гипотезу

$$P = Q ?$$

- ▶ Сможет ли классификатор отделить S_P от S_Q ?

Two sample tests



- ▶ Даны две выборки

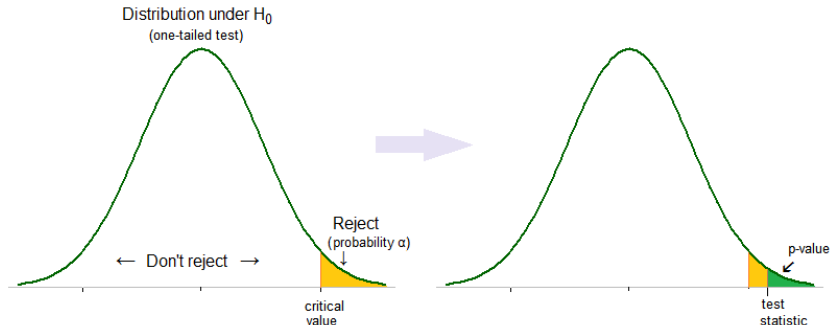
$$S_P \sim P^n \text{ и } S_Q \sim Q^m$$

- ▶ Необходимо проверить гипотезу

$$P = Q ?$$

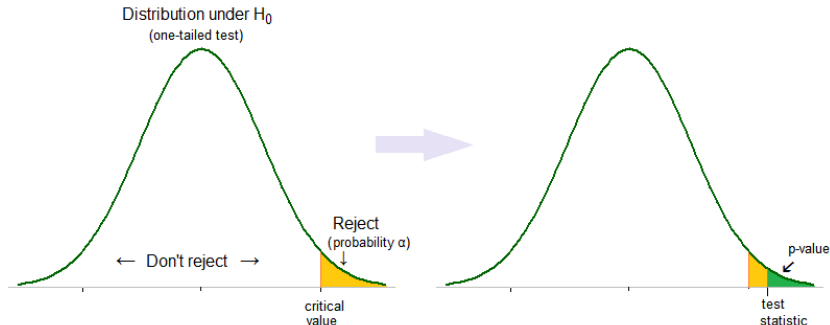
- ▶ Сможет ли классификатор отделить S_P от S_Q ?
- ▶ Численная метрика качества для GAN

Статистическая проверка гипотез



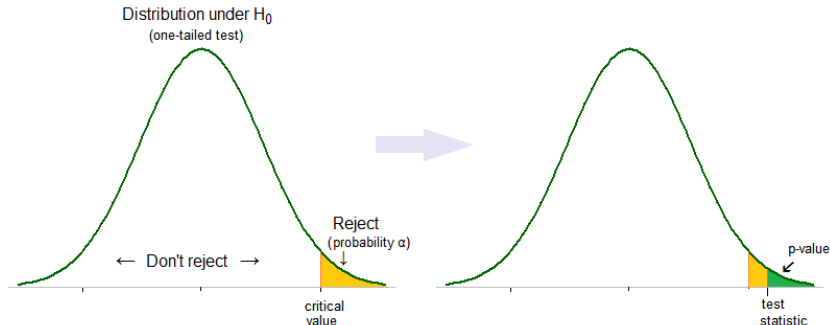
- Определить уровень значимости α

Статистическая проверка гипотез



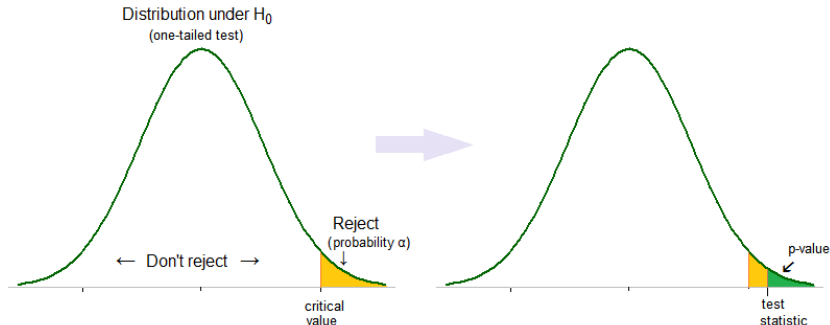
- ▶ Определить уровень значимости α
- ▶ Вычислить статистику $\hat{t} = T(X)$

Статистическая проверка гипотез



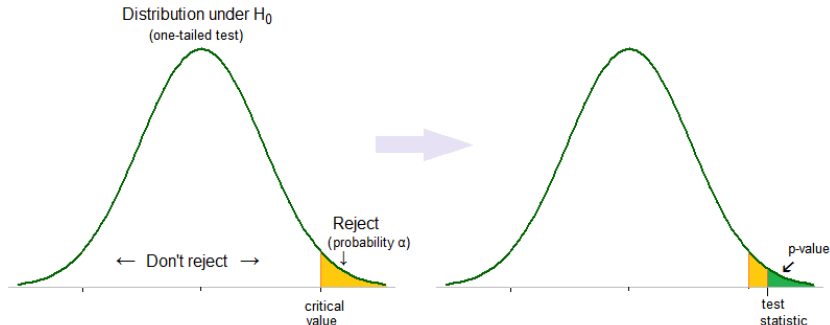
- ▶ Определить уровень значимости α
- ▶ Вычислить статистику $\hat{t} = T(X)$
- ▶ Вычислить p-value $\hat{p} = P(T > \hat{t} | H_0)$

Статистическая проверка гипотез



- ▶ Определить уровень значимости α
- ▶ Вычислить статистику $\hat{t} = T(X)$
- ▶ Вычислить p-value $\hat{p} = P(T > \hat{t} | H_0)$
- ▶ Отвергнуть H_0 , если $\hat{p} < \alpha$

Статистическая проверка гипотез



- ▶ Определить уровень значимости α
- ▶ Вычислить статистику $\hat{t} = T(X)$
- ▶ Вычислить p-value $\hat{p} = P(T > \hat{t} | H_0)$
- ▶ Отвергнуть H_0 , если $\hat{p} < \alpha$
- ▶ Получим $\pi = 1 - \beta$

Classifier Two-Sample Test

- ▶ собрать датасет

$$\mathcal{D} = \{(x_i, 0)\}_{i=1}^n \cup \{(y_i, 1)\}_{i=1}^n = \{(z_i, l_i)\}_{i=1}^{2n}$$

Classifier Two-Sample Test

- ▶ собрать датасет

$$\mathcal{D} = \{(x_i, 0)\}_{i=1}^n \cup \{(y_i, 1)\}_{i=1}^n = \{(z_i, l_i)\}_{i=1}^{2n}$$

- ▶ поделить на train и test

$$\mathcal{D} = \mathcal{D}_{tr} \cup \mathcal{D}_{te}$$

Classifier Two-Sample Test

- ▶ собрать датасет

$$\mathcal{D} = \{(x_i, 0)\}_{i=1}^n \cup \{(y_i, 1)\}_{i=1}^n = \{(z_i, l_i)\}_{i=1}^{2n}$$

- ▶ поделить на train и test

$$\mathcal{D} = \mathcal{D}_{tr} \cup \mathcal{D}_{te}$$

- ▶ обучить классификатор

$$f : \mathcal{X} \rightarrow [0, 1]$$

Classifier Two-Sample Test

- ▶ собрать датасет

$$\mathcal{D} = \{(x_i, 0)\}_{i=1}^n \cup \{(y_i, 1)\}_{i=1}^n = \{(z_i, l_i)\}_{i=1}^{2n}$$

- ▶ поделить на train и test

$$\mathcal{D} = \mathcal{D}_{tr} \cup \mathcal{D}_{te}$$

- ▶ обучить классификатор

$$f : \mathcal{X} \rightarrow [0, 1]$$

- ▶ вычислить C2ST статистику

$$\hat{t} = \sum_{(z_i, l_i) \in \mathcal{D}_{te}} \mathbb{I} \left[\mathbb{I} \left(f(z_i) > \frac{1}{2} \right) = l_i \right]$$

Classifier Two-Sample Test

- ▶ тестируем гипотезу

$$H_0 : P = Q \iff \hat{t} \sim \text{Bin} \left(|\mathcal{D}_{te}|, \frac{1}{2} \right)$$

$$H_1 : P \neq Q \iff \hat{t} \sim \text{Bin} \left(|\mathcal{D}_{te}|, p_1 = \frac{1}{2} + \epsilon \right)$$

Classifier Two-Sample Test

- ▶ тестируем гипотезу

$$H_0 : P = Q \iff \hat{t} \sim \text{Bin} \left(|\mathcal{D}_{te}|, \frac{1}{2} \right)$$

$$H_1 : P \neq Q \iff \hat{t} \sim \text{Bin} \left(|\mathcal{D}_{te}|, p_1 = \frac{1}{2} + \epsilon \right)$$

- ▶ воспользуемся нормальной аппроксимацией (Z-тест)

$$H_0 : \frac{\hat{t}}{|\mathcal{D}_{te}|} \sim \mathcal{N} \left(\frac{1}{2}, \frac{1}{4|\mathcal{D}_{te}|} \right)$$

$$H_1 : \frac{\hat{t}}{|\mathcal{D}_{te}|} \sim \mathcal{N} \left(p_1, \frac{p_1(1-p_1)}{|\mathcal{D}_{te}|} \right)$$

Classifier Two-Sample Test

- ▶ тестируем гипотезу

$$H_0 : P = Q \iff \hat{t} \sim \text{Bin} \left(|\mathcal{D}_{te}|, \frac{1}{2} \right)$$

$$H_1 : P \neq Q \iff \hat{t} \sim \text{Bin} \left(|\mathcal{D}_{te}|, p_1 = \frac{1}{2} + \epsilon \right)$$

- ▶ воспользуемся нормальной аппроксимацией (Z-тест)

$$H_0 : \frac{\hat{t}}{|\mathcal{D}_{te}|} \sim \mathcal{N} \left(\frac{1}{2}, \frac{1}{4|\mathcal{D}_{te}|} \right)$$

$$H_1 : \frac{\hat{t}}{|\mathcal{D}_{te}|} \sim \mathcal{N} \left(p_1, \frac{p_1(1-p_1)}{|\mathcal{D}_{te}|} \right)$$

- ▶ мощность критерия

$$1 - \beta = \Phi \left(\frac{\epsilon \sqrt{|\mathcal{D}_{te}|} - \Phi^{-1}(1 - \alpha)/2}{\sqrt{\frac{1}{4} - \epsilon^2}} \right)$$

Синтетические данные

- ▶ 2 гауссианы

$$x_i \sim \mathcal{N}(0, 1) \quad y_i \sim \mathcal{N}(0, 1)$$

- ▶ гауссиана VS t-распределение

$$x_i \sim \mathcal{N}(0, 1) \quad y_i \sim St(3)$$

- ▶ синусоида

$$x_i \sim \mathcal{N}(0, 1)$$

$$\epsilon_i \sim \mathcal{N}(0, \gamma^2)$$

$$y_i \sim \cos(\delta x_i) + \epsilon_i$$

Синтетические данные

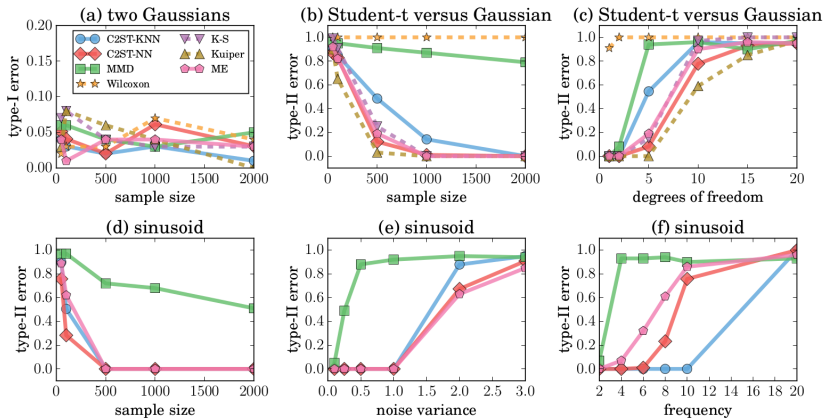


Figure 1: Results (type-I and type-II errors) of our synthetic two-sample test experiments.

NIPS и Emotional Faces



Problem	n^{te}	ME-full	ME-grid	SCF-full	SCF-grid	MMD-quad	MMD-lin	C2ST-NN
Bayes-Bayes	215	.012	.018	.012	.004	.022	.008	.002
Bayes-Deep	216	.954	.034	.688	.180	.906	.262	1.00
Bayes-Learn	138	.990	.774	.836	.534	1.00	.238	1.00
Bayes-Neuro	394	1.00	.300	.828	.500	.952	.972	1.00
Learn-Deep	149	.956	.052	.656	.138	.876	.500	1.00
Learn-Neuro	146	.960	.572	.590	.360	1.00	.538	1.00

Table 1: Type-I errors (first row) and powers (rest of rows) in distinguishing NIPS papers categories.

Problem	n^{te}	ME-full	ME-grid	SCF-full	SCF-grid	MMD-quad	MMD-lin	C2ST-NN
\pm vs. \pm	201	.010	.012	.014	.002	.018	.008	.002
$+$ vs. $-$	201	.998	.656	1.00	.750	1.00	.578	.997

Table 2: Type-I errors (first row) and powers (second row) in distinguishing facial expressions.

GAN






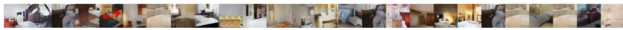






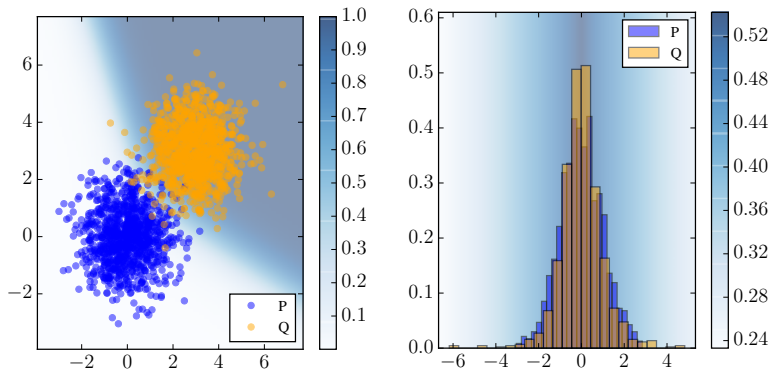
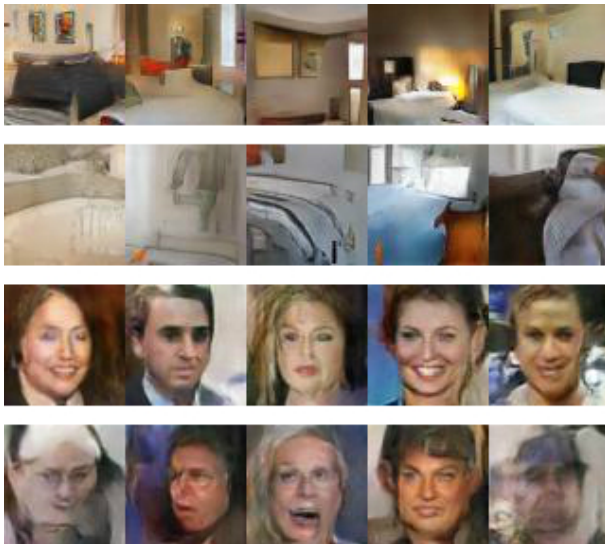
random sample	MMD	KNN	NN
	0.158	0.830	0.999
	0.154	0.994	1.000
	0.048	0.962	1.000
	<u>0.012</u>	0.798	0.964
	0.024	0.748	<u>0.949</u>
	0.019	<u>0.670</u>	0.983
	0.152	0.940	1.000
	0.222	0.978	1.000
	0.715	1.000	1.000
	<u>0.015</u>	0.817	0.987
	0.020	0.784	<u>0.950</u>
	0.024	<u>0.697</u>	0.971

Table 3: Results on GAN evaluation. Lower test statistics are best. Full results in Appendix [A](#)

Интерпретируемость



Интерпретируемость



Causal discovery

Причинность $X \rightarrow Y$ понимается как

$$x \sim P(x) \quad \epsilon \sim P(\epsilon)$$

$$y = g(x, \epsilon)$$

99 Tübingen cause-effect pairs

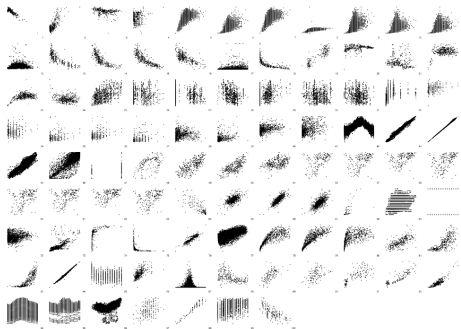
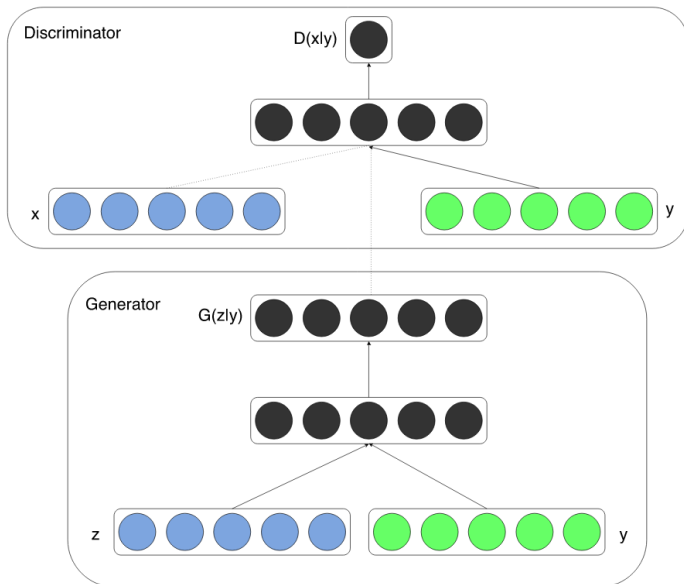


Figure 6: Scatter plots of the cause-effect pairs in the `CauseEffectPairs` benchmark data. We only show the pairs for which both variables are one-dimensional.

Conditional GAN



99 Tübingen cause-effect pairs

Method	ANM-HSIC	IGCI	RCC	CGAN-C2ST	Ensemble	C2ST type
Accuracy	67%	71%	76%	73%	82%	KNN
				70%	73%	NN
				58%	65%	MMD

Table 4: Results on cause-effect discovery on the Tübingen pairs experiment.

Плюсы и минусы

Плюсы

- ▶ Простая и понятная идея
- ▶ Универсальный метод
- ▶ Интерпретируемость

Минусы

- ▶ Тривиальная теория
- ▶ Неубедительные эксперименты

Литература I



David Lopez-Paz, Maxime Oquab
Revisiting Classifier Two-Sample Tests.
[arXiv:1610.06545 \[stat.ML\]](#)



Mehdi Mirza, Simon Osindero
Conditional Generative Adversarial Nets
[arXiv:1411.1784 \[cs.LG\]](#)



Joris M. Mooij et.al.
Distinguishing cause from effect using observational data:
methods and benchmarks
[arXiv:1412.3773 \[cs.LG\]](#)