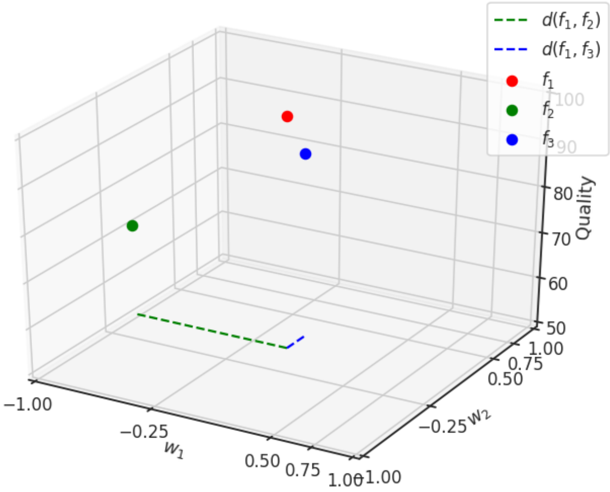
Функции расстояния на вероятностных пространствах

Московский Физико-Технический Институт

2021

Мотивация

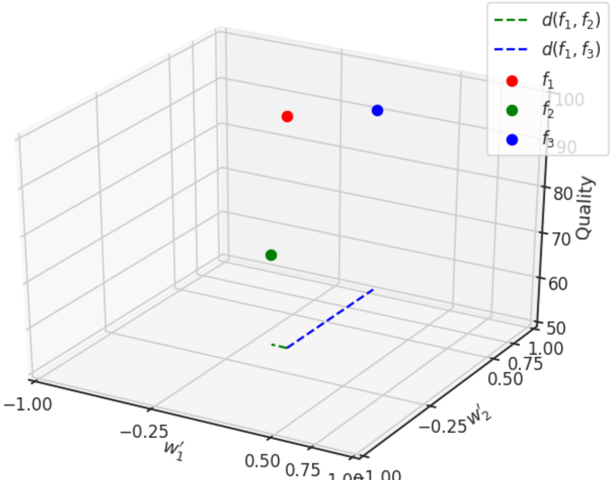
Какая модель ближе к f_1?



Мотивация

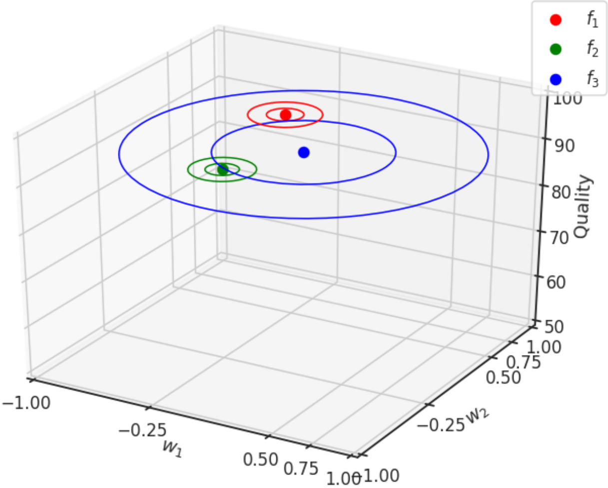
Какая модель ближе к f_1?

Смена метрики$\approx$смена координат. Разные метрики могут отражать различные свойства пространства моделей.



Мотивация

Какая модель ближе к f_1?



Определение и свойства

Пусть заданы пространство параметров $\boldsymbol{w}$.
Функция расстояния d — функция, определенная на паре распределений $p_1, p_2 \to \mathbb{R}_+$.
Возможные свойства:

- Аксиомы метрики
  - $d(p_1, p_1) = 0$
  - $d(p_1, p_2) = d(p_2, p_1)$
  - $d(p_1, p_2) \leq d(p_1, p_3) + d(p_3, p_2)$
- (Адуенко, 2017)
  - $d \in [0, 1]$
  - d определена в случае несовпадения носителей p_1, p_2
  - d близка к нулю, если p_2 — малоинформативное распределение
- Эксплуатационные критерии
  - Есть аналитическая формула
  - Требования к сложности вычисления

Total variation

Для двух вероятностных мер P_1, P_2 на множестве $\mathfrak{A}$

$$TV = \sup_{\mathfrak{a} \in \mathfrak{A}} |P_1(\mathfrak{a}) - P_2(\mathfrak{a})|$$

Свойства:

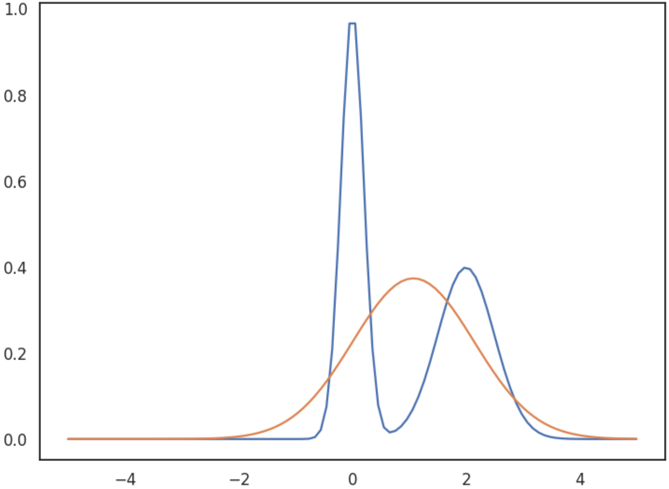
- $0 \leq TV \leq 1$
- TV — метрика
- $TV = 0 \iff P_1 = P_2$
- Лемма Шеффе: для дифференцируемых распределений с плотностями f_i на $\mathbb{R}^d$:

$$TV = \frac{1}{2} \int |f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})| d\boldsymbol{x} = \frac{1}{2} ||f_1 - f_2||_1.$$

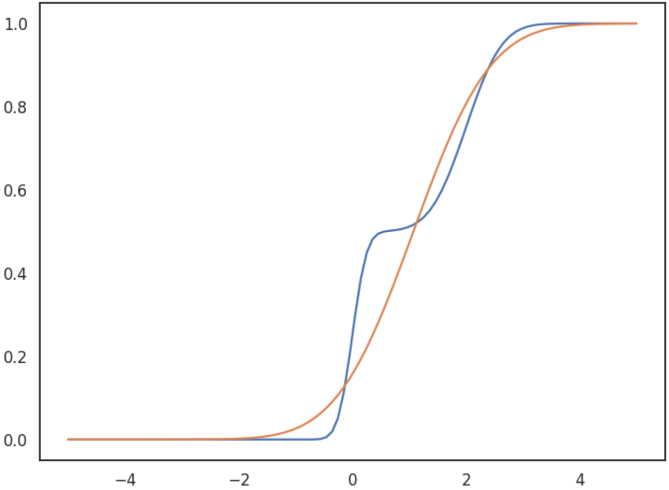
- $TV(\prod_i P_1^i, \prod_i P_2^i) \leq \sum_i TV(P_1^i, P_2^i)$
- Соответствует статистике в KS-тесте

Total variation: пример

Приближение гауссовой смеси нормальным распределением.

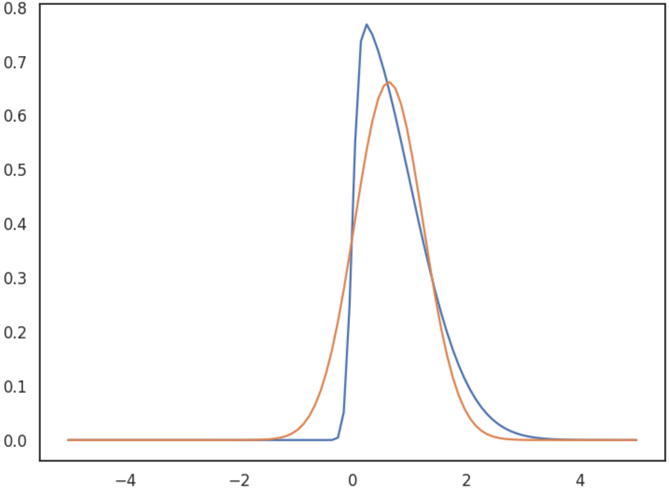


Total variation: пример

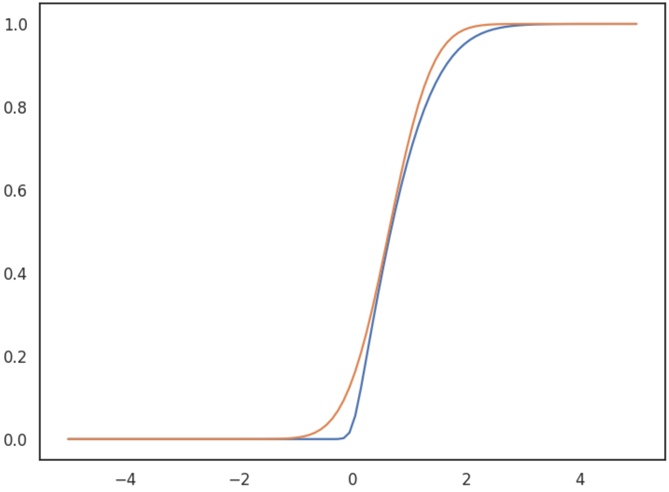


Total variation: пример

Приближение скошенного распределения нормальным распределением.



Total variation: пример



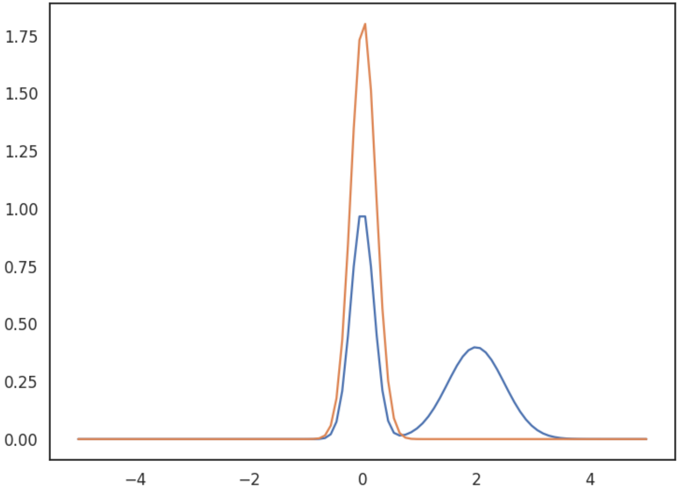
Расстояние Хеллингера

$$H = \sqrt{\int (f_1(\boldsymbol{x}) - f_2(\boldsymbol{x}))^2 \, d\boldsymbol{x}} = ||\sqrt{f_1} - \sqrt{f_2}||_2$$

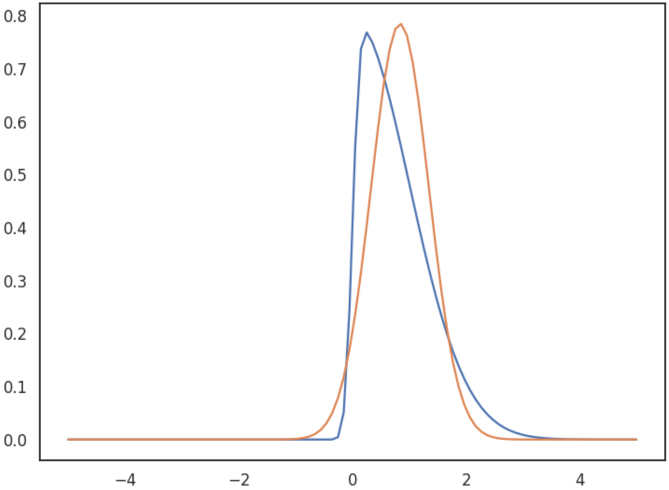
Свойства:

- $0 \leq H \leq 2$
- H — метрика
- $H = 0 \iff P_1 = P_2$
- $H^2(\prod_i P_1^i, \prod_i P_2^i) \leq \sum_i H^2(P_1^i, P_2^i)$
- $1 - H^2 = 1 - \int \sqrt{f_1(\boldsymbol{x}) f_2(\boldsymbol{x})} d\boldsymbol{x}$

Расстояние Хеллингера: пример



Расстояние Хеллингера: пример



KL-дивергенция

$$KL(P_1, P_2) = \int \log \frac{f_1(x)}{f_2(x)} f_1(x) dx$$

- $KL \geq 0$
- KL — не метрика, т.к. не симметрична
- KL — не метрика, т.к. нарушается свойство треугольника
- $KL = 0 \iff P_1 = P_2$
- $KL(\prod_i P_1^i, \prod_i P_2^i) = \sum_i KL(P_1^i, P_2^i)$
- Если есть зависимость между двумя случайными величинами $\boldsymbol{w}, \gamma$, то

$$KL\left(p_1(\boldsymbol{w}, \gamma), p_2(\boldsymbol{w}, \gamma)\right) = KL(p_1(\boldsymbol{w}), p_2(\boldsymbol{w})) + \int_{\boldsymbol{w}} p_1(\boldsymbol{w}) \int_{\gamma} \log \frac{p_1(\gamma|\boldsymbol{w})}{p_2(\gamma|\boldsymbol{w})} p_1(\gamma|\boldsymbol{w}) d\gamma d\boldsymbol{w}$$

Энтропия

Дифференциальная энтропия — обобщение энтропии для непрерывных распределений:

$$h(\boldsymbol{w}) = -\int_{\boldsymbol{w}} \log f(\boldsymbol{w}) f(\boldsymbol{w}) d\boldsymbol{w}$$

- Не инварианта относительно замены переменных
  - $h(\boldsymbol{F}(\boldsymbol{w})) \leq h(\boldsymbol{w}) + \int f(\boldsymbol{w}) \log \left| \frac{\partial \boldsymbol{F}}{\partial \boldsymbol{w}} \right| d\boldsymbol{w}$
  - Если $\boldsymbol{F}$ — неравенство превращается в равенство
- Может принимать отрицательные значения

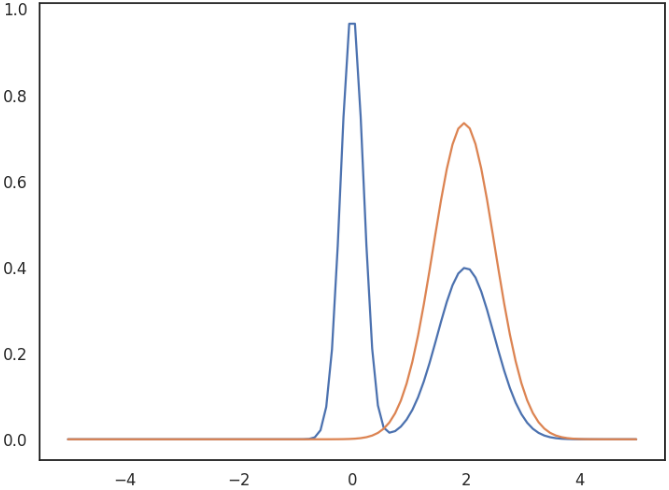
KL можно рассматривать как модификацию энтропии, которая

- Инварианта относительно замены переменных
- Всегда положительна

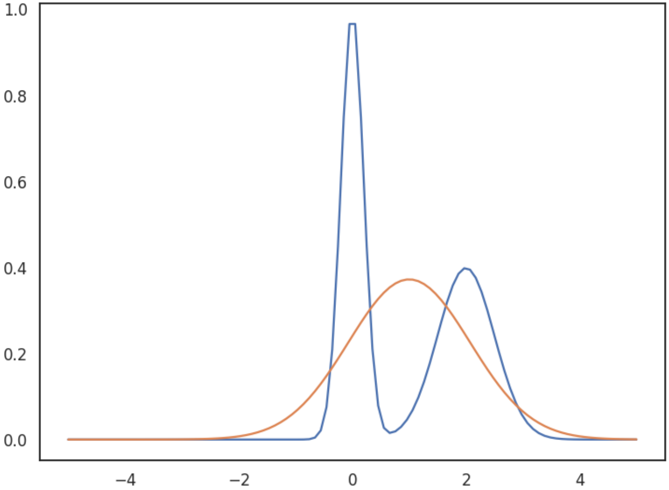
Интерпретации $KL(P_1, P_2)$:

- Количество информации, которую можно получить, если использовать P_1 вместо P_2
- Количество информации, которую придется потратить на кодирование данных, распределенных по P_1, если декодер рассчитан на коды из P_2.

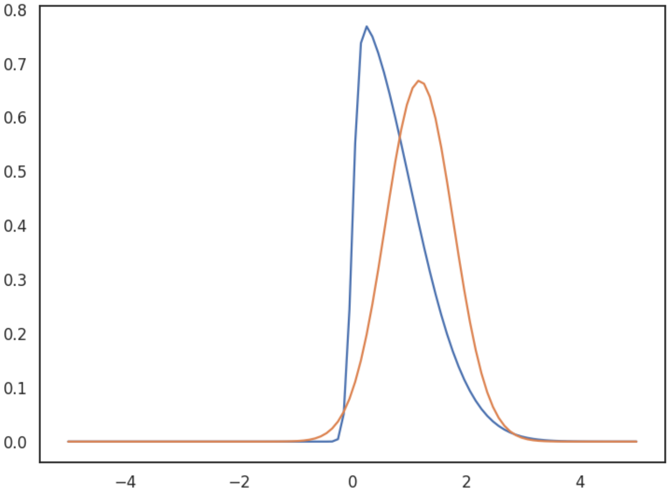
KL: пример



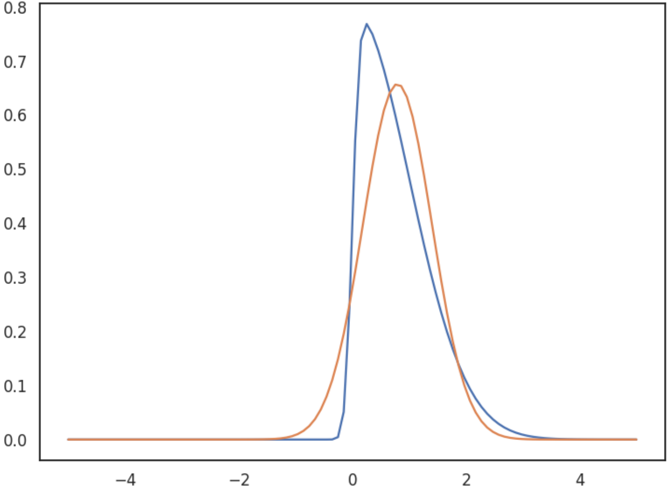
KL: пример



KL: пример



KL: пример

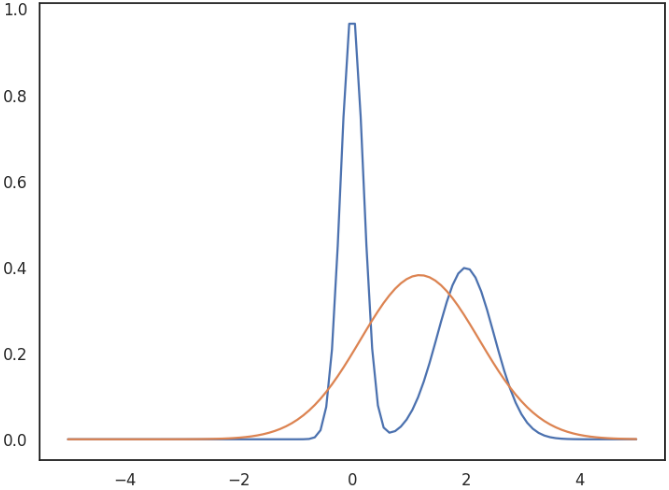


JS

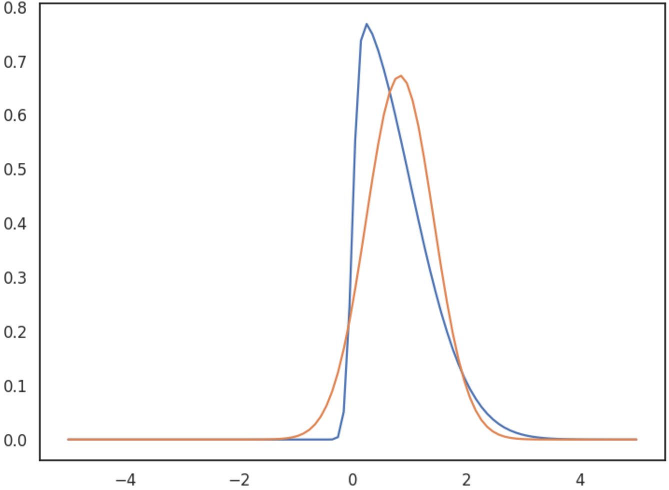
$$JS(P_1, P_2) = \frac{1}{2} KL\left(P_1 \middle| \frac{1}{2}P_1 + \frac{1}{2}P_2\right) + \frac{1}{2} KL\left(P_2 \middle| \frac{1}{2}P_1 + \frac{1}{2}P_2\right)$$

- $0 \leq JS \leq 1$
- $\sqrt{JS}$ — метрика
- $JS = 0 \iff P_1 = P_2$

JS: пример

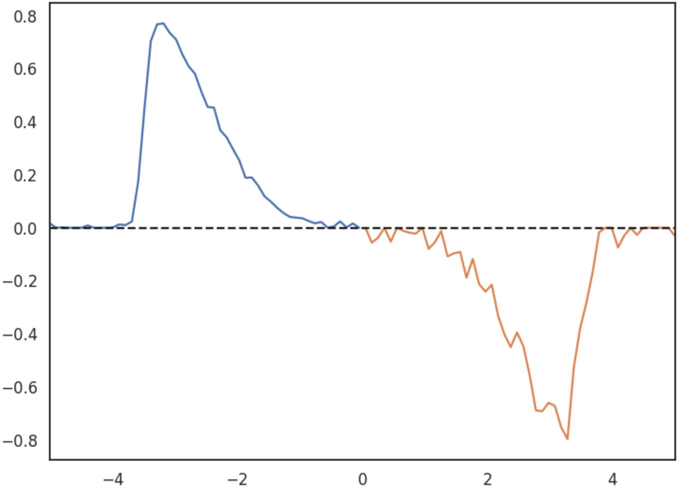


JS: пример



Расстояние Вассерштайна: мотивация

Гаспар Монж: как дешевле всего перенести кучу песка в канаву?



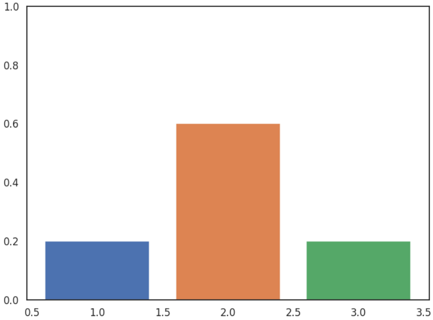
Расстояние Вассерштайна: дискретная задача

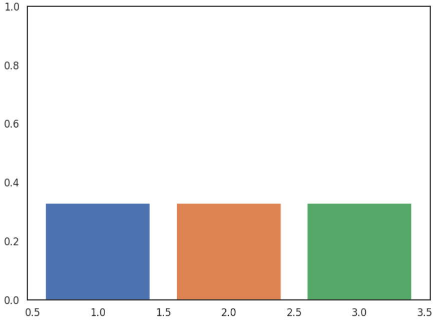
Пусть заданы две дискретных вероятностных меры $p_1(\boldsymbol{w}_i^1), i \in \{1, \dots n_1\}$, $p_2(\boldsymbol{w}_j^2), j \in \{1, \dots n_2\}$.
Пусть задана матрица стоимости $\boldsymbol{C}$: $c_{ij} \in \mathbb{R}_+$.

Требуется найти отображение, задаваемое матрицей элементов t_{ij}, такое что:

- $\sum_i t_{ij} = p_2(\boldsymbol{w}_j^2), \sum_j t_{ij} p_2(\boldsymbol{w}_i^1)$
- $\sum_i \sum_j c_{ij} t_{ij} \to \min$.

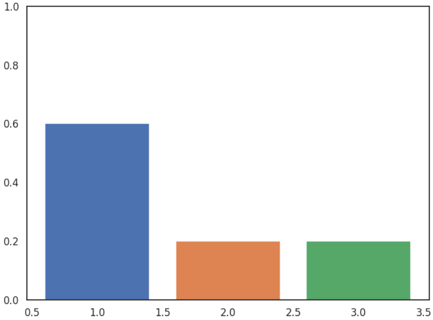
Дискретная задача: пример

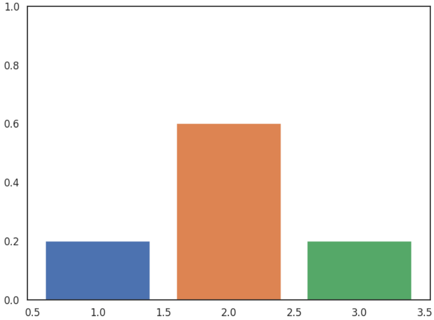




Стоимость: 0.4

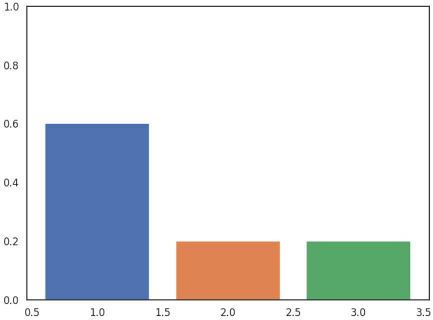
Дискретная задача: пример

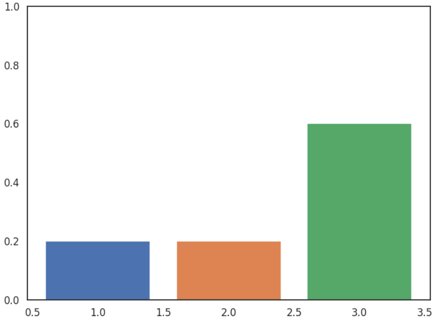




Стоимость: 0.4

Дискретная задача: пример





Стоимость: 0.8

Расстояние Вассерштайна: непрерывная задача

Пусть заданы две непрерывные вероятностные меры $P_1(\boldsymbol{w}^1), \boldsymbol{w}^1 \in \mathbb{W}_1,\ P_2(\boldsymbol{w}^2), \boldsymbol{w}^2 \in \mathbb{W}_2$.
Пусть задана функция стоимости $C : \mathbb{W}_1 \times \mathbb{W}_2 \to \mathbb{R}_+$.

Требуется найти совместное распределение T на $\mathbb{W}_1 \times \mathbb{W}_2$, такое что:

- $\int_{\mathbb{W}_1} dT(\boldsymbol{w}_1, \boldsymbol{w}_2) = P_1, \quad \int_{\mathbb{W}_2} dT(\boldsymbol{w}_1, \boldsymbol{w}_2) = P_2$
- $\int_{\mathbb{W}_1 \times \mathbb{W}_2} C(\boldsymbol{w}_1, \boldsymbol{w}_2) dT(\boldsymbol{w}_1, \boldsymbol{w}_2) \to \min.$

Двойственная задача

$$\max_{\hat{T}_1, \hat{T}_2} \int_{\mathbb{W}_1} \hat{T}_1(\boldsymbol{w}_1) f_1(\boldsymbol{w}_1) d\boldsymbol{w}_1 + \int_{\mathbb{W}_2} \hat{T}_2(\boldsymbol{w}_2) f_2(\boldsymbol{w}_2) d\boldsymbol{w}_2$$

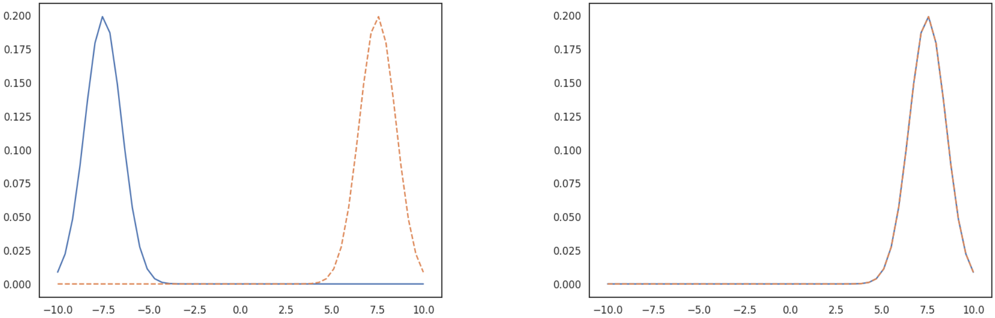
при $\hat{T}_1(\boldsymbol{w}_1) + \hat{T}_2(\boldsymbol{w}_2) \leq C(\boldsymbol{w}_1, \boldsymbol{w}_2)$

Теорема Канторовича-Рубинштейна

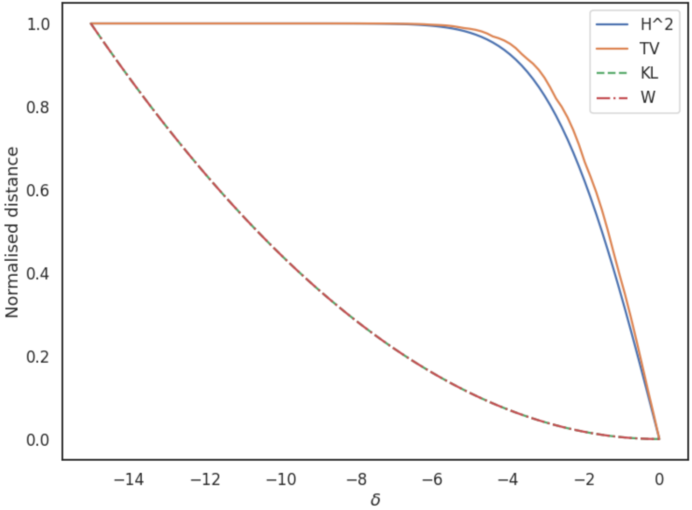
Пусть $\mathbb{W}_1 = \mathbb{W}_2$ и $C = ||\cdot||_1$. Тогда двойственная задача выглядит следующим образом:

$$\max \hat{T} \in \mathrm{Lip}_1 \int_{\mathbb{W}} \hat{T}(\boldsymbol{w}) f_1(\boldsymbol{w}) d\boldsymbol{w} - \int_{\mathbb{W}} \hat{T}(\boldsymbol{w}) f_2(\boldsymbol{w}) d\boldsymbol{w}$$

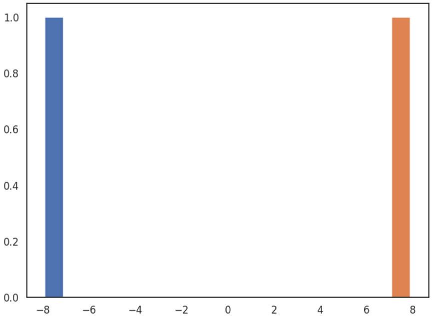
Расстояние между пиками: пример

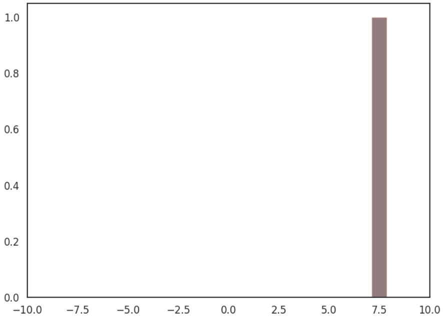


Расстояние между пиками: пример



Расстояние между пиками: пример





Расстояние между пиками: пример

$$TV = 0$$

$$H = 0$$

$$KL = \begin{cases} 0, & \delta = 0 \\ \infty, & \text{иначе.} \end{cases}$$

$$JS = \begin{cases} 0, & \delta = 0 \\ \log 2, & \text{иначе.} \end{cases}$$

$$W = |\delta|.$$

Вывод: при работе с распределениями с различающимися носителями предпочтительно использование расстояния Вассерштайна.

Литература и прочие ресурсы

- Адуенко А. А. Выбор мультимоделей в задачах классификации : дис. – Федер. исслед. центр"Информатика и управление"РАН, 2017.
- Andrew Nobel: Distances and Divergences for Probability Distributions, https://nobel.web.unc.edu/wp-content/uploads/sites/13591/2020/11/Distance-Divergence.pdf
- Про KL с условной вероятонстью: http://akosiorek.github.io/ml/2017/09/10/kl-hierarchical-vae.html
- Kolouri, Cattell, Rohde: Optimal Transport: A Crash Course, http://imagedatascience.com/transport/OTCrashCourse.pdf
- Computational Optimal Transport - https://arxiv.org/pdf/1803.00567.pdf
- Про GAN и WGAN: https://lilianweng.github.io/lil-log/2017/08/20/from-GAN-to-WGAN.html#kullbackleibler-and-jensenshannon-divergence Wasserstein GAN - https://arxiv.org/abs/1701.07875