Домашнее задание №3

Иванов Вячеслав, группа 699

7 ноября 2018 г.

Оглавление

1 Дивергенция Кульбака-Лейблера

Определение. Пусть даны два вероятностных распределения — $p(x)$ и $q(x)$. Тогда *дивергенцией Кульбака-Лейблера* (KL-divergence) называют величину:

$$D_{KL}(p \parallel q) := \int\limits_D p(x) \log \frac{p(x)}{q(x)} dx,\ D := (\mathrm{dom}(p)\ \cap\ \mathrm{dom}(q)) \setminus \{q = 0\}$$

Утверждение 1. KL-дивергенция обладает следующими свойствами:

1. $D_{KL}(p \parallel q)$ — выпуклая функция на множестве пар вероятностных распределений.
2. $D_{KL}(p \parallel q) = 0 \iff p = q$ почти всюду.
3. $D_{KL}(p \parallel q) \geqslant 0$
4. $D_{KL}(p \parallel q) \neq D_{KL}(q \parallel p)$

Доказательство.

1. Докажем выпуклость множества пар распределений. Для этого достаточно показать замкнутость относительно выпуклых комбинаций. Как известно, плотность распределения — неотрицательная функция с интегралом Лебега 1. Оба этих свойства сохраняются при взятии линейной комбинации. Докажем теперь выпуклость по определению.

$$\begin{aligned}
&D_{KL}(\theta p_1 + (1-\theta)p_2 \parallel \theta q_1 + (1-\theta)q_2) \\
&= \int_D (\theta p_1(x) + (1-\theta)p_2(x)) \log \frac{\theta p_1(x) + (1-\theta)p_2(x)}{\theta q_1(x) + (1-\theta)q_2(x)} dx \\
&= -\int_D (\theta p_1(x) + (1-\theta)p_2(x)) \log \frac{\theta q_1(x) + (1-\theta)q_2(x)}{\theta p_1(x) + (1-\theta)p_2(x)} dx \\
&= -\int_D (\theta p_1(x) + (1-\theta)p_2(x)) \log \frac{\theta p_1(x)\frac{q_1(x)}{p_1(x)} + (1-\theta)p_2(x)\frac{q_2(x)}{p_2(x)}}{\theta p_1(x) + (1-\theta)p_2(x)} dx \\
&\leqslant -\int_D \theta p_1(x) \log \frac{q_1(x)}{p_1(x)} dx - \int_D (1-\theta p_2(x)) \log \frac{q_2(x)}{p_2(x)} dx \\
&= \theta D_{KL}(p_1 \parallel q_1) + (1-\theta) D_{KL}(p_2 \parallel q_2)
\end{aligned}$$

Приём, аналогичный переходу с 3-ей строки на 4-ую, будет использоваться и дальше. Заключительный шаг следует из неравенства Йенсена для логарифма.

2. Воспользуемся выпуклостью и неотрицательностью KL-дивергенции. Пусть $D_{KL}(p_0 \parallel q_0) = 0$. В силу критерия, ограничение $D_{KL}(p \parallel q)$ на прямую $p = p_0$ есть выпуклая функция по q. Из определения ясно, что $D_{KL}(p \parallel p) = 0$. Предположим, что есть $q_0 \neq p_0$ на множестве ненулевой меры такая что $D_{KL}(p \parallel q) = 0$. Рассмотрим отрезок $q_0 + (p_0 - q_0)t,\ t \in [0, 1]$ и скалярную функцию $d(t) := D_{KL}(p_0 \parallel q_0 + (p_0 - q_0)t)$. Поскольку это неотрицательная выпуклая функция, равная нулю на концах отрезка, должно быть $d'(0) = d'(1) = 0$.

$$\begin{aligned}
d(t) &= \int_D p_0(x) \ln \frac{p_0(x)}{tp_0(x) + (1-t)q_0(x)} dx \\
&= -\int_D p_0(x) \ln \frac{tp_0(x) + (1-t)q_0(x)}{p_0(x)} dx \\
&= -\int_D p_0(x) \ln \left(t + (1-t)\frac{q_0(x)}{p_0(x)} \right) dx
\end{aligned}$$

$$
\begin{aligned}
d'(t) &= -\int_D p_0(x)\frac{1-\frac{q_0(x)}{p_0(x)}}{t+(1-t)\frac{q_0(x)}{p_0(x)}}dx \\
&= -\int_D p_0(x)\frac{p_0(x)-q_0(x)}{tp_0(x)+(1-t)q_0(x)}dx \\
&= -\int_D p_0(x)\frac{p_0(x)-q_0(x)}{q_0(x)+(p_0(x)-q_0(x))t}dx \\
d'(1) &= -\int_D p_0(x)\frac{p_0(x)-q_0(x)}{p_0(x)}dx = 0 \\
d'(0) &= -\int_D p_0(x)\frac{p_0(x)-q_0(x)}{q_0(x)}dx \\
&= -\int_D \left(\frac{p_0^2(x)}{q_0} - p_0(x)\right)dx \\
&= 1 - \int_D \frac{p_0^2(x)}{q_0}dx
\end{aligned}
$$

Заметим, что:

$$
\begin{aligned}
q_0(x) &= \frac{(q_0(x)-p_0(x))^2}{q_0} + 2p_0(x) - \frac{p_0^2(x)}{q_0(x)} \\
\frac{p_0^2(x)}{q_0(x)} &= \frac{(q_0(x)-p_0(x))^2}{q_0(x)} + 2p_0(x) - q_0(x) \\
\int_D \frac{p_0^2(x)}{q_0(x)}dx &= \int_D \frac{(q_0(x)-p_0(x))^2}{q_0(x)}dx + 1 > 1
\end{aligned}
$$

Где последнее неравенство верно в силу неотрицательности подынтегральной функции. Отсюда получаем противоречие: $d'(0) \neq 0$, хотя 0 — точка минимума функции f. Противоречие. Значит, $D_{KL}(p \parallel q) \implies p = q$ почти всюду.
Доказательство в обратную сторону очевидно просто из того, что подынтегральная функция почти всюду равна нулю.

3.

$$
\begin{aligned}
D_{KL}(p \parallel q) &= \int_D p(x)\log\frac{p(x)}{q(x)}dx \\
&= -\int_D p(x)\log\frac{q(x)}{p(x)}dx \\
&\geqslant -\log\int_D q(x)dx = 0
\end{aligned}
$$

Переход к неравенству корректен по неравенству Йенсена для матожидания.

4.

$$
\begin{aligned}
&p(x) := \lambda e^{-\lambda x} I(x > 0),\ \lambda > 0;\ q(x) := I(x \in [0,1]) \\
&D_{KL}(p \parallel q) = \int_{[0,1]} \lambda e^{-\lambda x}\log \lambda e^{-\lambda x}dx \\
&= \log\lambda(1-e^{-\lambda}) - \lambda\int_{[0,1]} \lambda x e^{-\lambda x}dx \\
&= \log\lambda(1-e^{-\lambda}) + (1-(\lambda+1)e^{-\lambda}) \\
&= 1 + \log\lambda + (1 - \log\lambda + \lambda)e^{-\lambda}
\end{aligned}
$$

Так как $\int_{[0,1]} \lambda x e^{-\lambda x} dx = \frac{1}{\lambda}\int_0^\lambda t e^{-t} dt = \frac{1}{\lambda}(e^{-t}(-1-t))|_0^\lambda = \frac{1}{\lambda}(e^{-\lambda}(-1-\lambda)+1)$

$$\begin{aligned} D_{KL}(q \parallel p) &= \int_{[0,1]} q(x) \log \frac{q(x)}{p(x)} dx \\ &= \int_{[0,1]} \log \frac{1}{\lambda e^{-\lambda x}} dx \\ &= -\log \lambda + \frac{\lambda}{2} \end{aligned}$$

Видно, что в этом случае $D_{KL}(p \parallel q) \neq D_{KL}(q \parallel p)$.

$\square$

2 t-SNE

t-SNE — *t-distributed stochastic neighbor embedding* — алгоритм визуализации многомерных данных, сохраняющий их внутреннюю структуру. В двух словах, суть в минимизации различий между распределением точек до и после проецирования: кластеры должны сохраняться.
Физическая аналогия: Алгоритм определяет силы взаимного притяжения и отталкивания между частицами-данными и задаёт уравнения движения, которые затем интегрируются для минимизации энергии системы. Потенциалы подобраны таким образом, что минимальной энергией обладает конфигурация, максимально похожая на ту, которая была в исходных данных.
Основные шаги:

1. Вычислить попарное подобие точек в исходных данных.
   Формализация: подсчитать условные вероятности вида
   $$p_{i|j} := \frac{\exp(-\|x_i - x_j\|/2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|/2\sigma_i^2)}$$
   Смысл: p_{ij} — степень уверенности в том, что точку x_j можно назвать "соседней"с x_i, если считать, что эта вероятность задаётся нормальным распределением с центром в x_i и дисперсией σ_i^2. Т.е. каждая точка порождает своё распределение, на основе которого вычисляется её близость к другим точкам. Поскольку вообще говоря $p_{i|j} \neq p_{j|i}$, коэффициенты подобия симметризуют:
   $$p_{ij} := \frac{p_{i|j} + p_{j|i}}{2}$$
2. Вычислить попарное подобие точек после проецирования.
   Формализация:
   $$q_{i|j} := \frac{f(-\|y_i - y_j\|)}{\sum_{k \neq i} f(-\|y_i - y_k\|)}, \; f(z) := \frac{1}{1+z^2}$$
   где принцип подбора коэффициентов σ_i будет описан ниже. Смысл тот же, но вместо нормального распределения теперь t-распределение (распределение Стьюдента) с одной степенью свободы (оно же распределение Коши). В оригинальной статье указано не оно, но промышленные реализации, насколько я понял, сделано именно так. Причину можно проиллюстрировать:

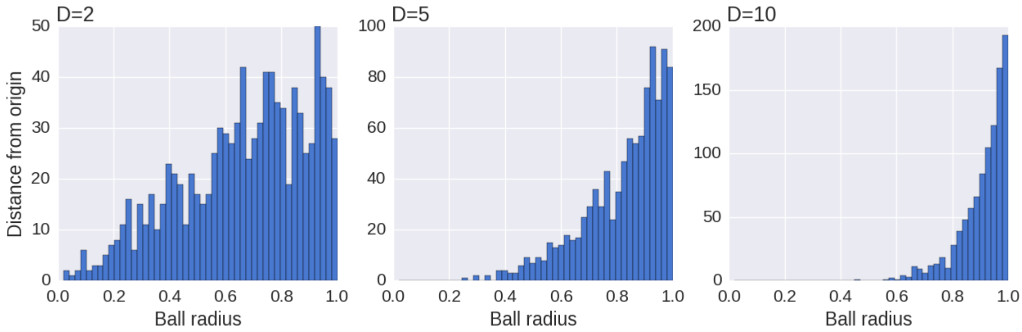

Рис. 1: График из статьи на O'Reilly. Гистограмма отражает распределение расстояний до центра единичного шара в d-мерном пространстве при равномерном распределении точек внутри шара.

Видно, что в малых размерностях распределение имеет тяжёлый хвост, и нормальное распределение такое поведение моделирует плохо, нужна огромная дисперсия. Распределение Стьюдента решает именно эту проблему: оно похоже на нормальное, но вероятность не так сосредоточена относительно среднего.

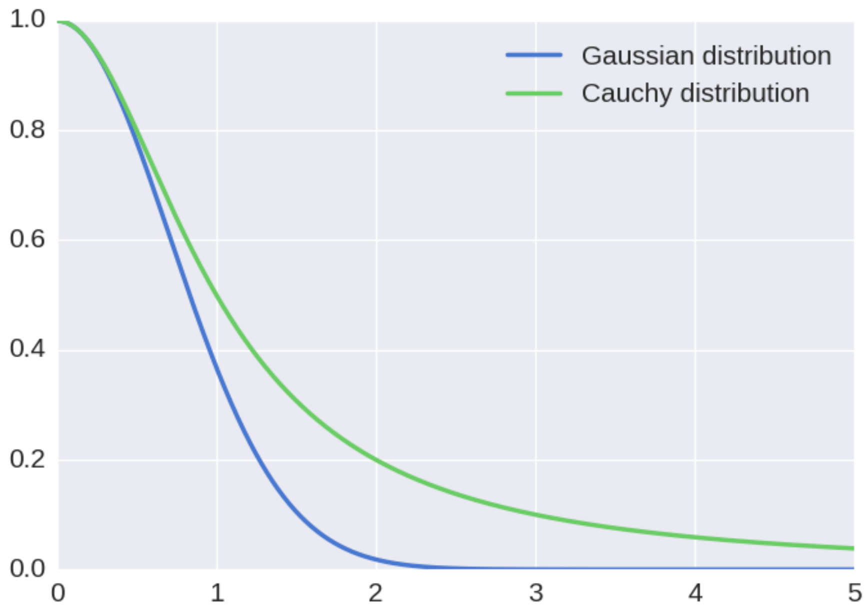

Рис. 2: Сравнение нормального распределения и распределения Стьюдента

Коэффициенты симметризуются тем же способом:

$$q_{ij} := \frac{q_{i|j} + q_{j|i}}{2}$$

3. Позиции точек в маломерном пространстве изменяются таким образом, чтобы максимизировать подобие матриц $P := [p_{ij}]$ и $Q := [q_{ij}]$. В качестве меры подобия выступает дискретная дивергенция Кульбака-Лейблера:

$$D_{KL}(P \parallel Q) := \sum_{i,j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

Она минимизируется методом градиентного спуска и даже имеет адекватный градиент:

$$\frac{\partial D_{KL}(P \parallel Q)}{dy_i} = 4 \sum_j (p_{ij} - q_{ij}) g(|x_i - x_j|) u_{ij}, \; g(z) := \frac{z}{1+z^2}$$

где u_{ij} — единичный вектор, соединяющий y_i с y_j. Формула взята из той же статьи на O'Reilly, её вид согласуется с приведенным в исходной статье.
Ясно, что итоговый вид формулы зависит от выбора функции $f(z)$.
Как было доказано выше, KL-дивергенция обладает такими хорошими с точки зрения оптимизации свойствами, как выпуклость и равенство нулю как критерий совпадения распределений почти всюду, а интеграл Лебега в дискретном случае записывается суммой и может быть эффективно вычислен. Как следствие, алгоритм получил широкое применение.

Вывод всех использованных формул можно найти в приложении к исходной статье.

3 Перспективная выпуклость

Если $f : \mathbb{R}^n \to \mathbb{R}$ — выпуклая функция, то $g(x,t) := tf(x/t), \; \mathrm{dom}(g) := \{(x,t) \in \mathbb{R}^{n+1} | \; t > 0, x/t \in \mathrm{dom}(f)\}$ — *перспективное преобразование* функции f — тоже выпуклая функция.

Доказательство.

$$\begin{aligned}
&\forall (x_1, t_1), \; (x_2, t_2) \in \mathrm{dom}(g) \; \forall \theta \in [0,1] : \\
&(\theta t_1 + (1-\theta)t_2) f\left(\frac{\theta x_1 + (1-\theta)x_2}{\theta t_1 + (1-\theta)t_2}\right) \\
&= (\theta t_1 + (1-\theta)t_2) f\left(\frac{(\theta t_1)\frac{x_1}{t_1} + ((1-\theta)t_2)\frac{x_2}{t2}}{\theta t_1 + (1-\theta)t_2}\right) \\
&= (\theta t_1 + (1-\theta)t_2) f\left(\frac{(\theta t_1)}{\theta t_1 + (1-\theta)t_2}\frac{x_1}{t_1} + \frac{((1-\theta)t_2)}{\theta t_1 + (1-\theta)t_2}\frac{x_2}{t2}\right) \\
&= (\theta t_1 + (1-\theta)t_2) f\left(c_1 \frac{x_1}{t_1} + c_2 \frac{x_2}{t_2}\right) \\
&\leqslant (\theta t_1 + (1-\theta)t_2)\left(c_1 f\left(\frac{x_1}{t_1}\right) + c_2 f\left(\frac{x_2}{t_2}\right)\right) \text{[Неравенство Йенсена]} \\
&= \theta t_1 f\left(\frac{x_1}{t_1}\right) + (1-\theta) t_2 f\left(\frac{x_2}{t_2}\right) \\
&= \theta g(x_1, t_1) + (1-\theta) g(x_2, t_2) \\
&c_1 := \frac{\theta t_1}{\theta t_1 + (1-\theta)t_2}, \; c_2 := \frac{(1-\theta)t_2}{\theta t_1 + (1-\theta)t_2}, \; c1, c2 \geqslant 0, \; c_1 + c_2 = 1
\end{aligned}$$

□

4 Обратное неравенство Йенсена

Если f — выпуклая функция, а $\lambda_1 > 0, \; \forall i > 2 : \lambda_i \leqslant 0, \; \sum_{i=1}^n \lambda_i = 1$, то

$$\forall x_1, \ldots, x_n \in \mathbb{R} : f(\lambda_1 x_1 + \ldots + \lambda_n x_n) > \lambda_1 f(x_1) + \ldots \lambda_n f(x_n)$$

Доказательство.

$$\lambda_1 = 1 - \sum_{i=2}^n \lambda_i, \; y := \lambda^T \mathbf{x}$$

$$\implies x_1 = \frac{1}{\lambda_1} y - \sum_{i=2}^n \frac{\lambda_i}{\lambda_1} x_i$$

Заметим, что в полученном выражении все коэффициенты неотрицательные и суммируются к единице, потому можно применить обычное неравенство Йенсена.

$$f(x_1) < \frac{1}{\lambda_1} f(y) - \sum_{i=2}^{n} \frac{\lambda_i}{\lambda_1} f(x_i)$$

Откуда домножением на λ_1 и переносом суммы в левую часть получается неравенство из условия. $\square$

5 Логарифмический барьер для конуса второго порядка

Доказать, что функция

$$f(\mathbf{x}, t) := -\log(t^2 - \mathbf{x}^T\mathbf{x})$$

выпуклая на

$$X := \{(\mathbf{x}, t) \in \mathbb{R}^n \times \mathbb{R}^+ \mid \|\mathbf{x}\|_2 < t\}$$

Доказательство. Во первых, заметим, что $t^2 - \mathbf{x}^T\mathbf{x}$ — выпуклая на X функция. В самом деле, её матрица Гессе имеет вид $\operatorname{diag}(-2, \dots, -2, 2)$. Она положительно определена на X, т.к. $2(t^2 - \mathbf{x}^T\mathbf{x}) > 0$ на X, а значит выпукла там по дифференциальному критерию второго порядка. К сожалению, функция $-\log(\cdot)$ — выпуклая невозрастающая, а композиция такой функции с выпуклой не обязательно выпуклая.

Scalar composition

We first consider the case $k = 1$, so $h : \mathbf{R} \to \mathbf{R}$ and $g : \mathbf{R}^n \to \mathbf{R}$. We can restrict ourselves to the case $n = 1$ (since convexity is determined by the behavior of a function on arbitrary lines that intersect its domain).

To discover the composition rules, we start by assuming that h and g are twice differentiable, with $\mathbf{dom}\, g = \mathbf{dom}\, h = \mathbf{R}$. In this case, convexity of f reduces to $f'' \geq 0$ (meaning, $f''(x) \geq 0$ for all $x \in \mathbf{R}$).

The second derivative of the composition function $f = h \circ g$ is given by

$$f''(x) = h''(g(x))g'(x)^2 + h'(g(x))g''(x). \tag{3.9}$$

Now suppose, for example, that g is convex (so $g'' \geq 0$) and h is convex and nondecreasing (so $h'' \geq 0$ and $h' \geq 0$). It follows from (3.9) that $f'' \geq 0$, *i.e.*, f is convex. In a similar way, the expression (3.9) gives the results:

$$\begin{array}{l} f \text{ is convex if } h \text{ is convex and nondecreasing, and } g \text{ is convex,} \\ f \text{ is convex if } h \text{ is convex and nonincreasing, and } g \text{ is concave,} \\ f \text{ is concave if } h \text{ is concave and nondecreasing, and } g \text{ is concave,} \\ f \text{ is concave if } h \text{ is concave and nonincreasing, and } g \text{ is convex.} \end{array} \tag{3.10}$$

Рис. 3: Boyd, 'Convex Optimization', p.84

Можно проверить, что градиент и матрица Гессе для барьера запишутся так:

$$\nabla f(\mathbf{x}, t) = \frac{2}{t^2 - \mathbf{x}^T\mathbf{x}} \begin{bmatrix} \mathbf{x} \\ -t \end{bmatrix}$$

$$\nabla^2 f(\mathbf{x}, t) = \frac{2}{(t^2 - \mathbf{x}^T\mathbf{x})^2} \begin{bmatrix} (t^2 - \mathbf{x}^T\mathbf{x})I + 2\mathbf{x}\mathbf{x}^T & -2t\mathbf{x} \\ -2t\mathbf{x}^T & t^2 + \mathbf{x}^T\mathbf{x} \end{bmatrix}$$

Вывод простой, но технический, поэтому я его опустил.
Тем не менее, ни первый, ни второй дифференциальный критерий у меня применить не вышло (по крайней мере, на данный момент). Я смог доказать выпуклость в частных случаях малой размерности вектора $\mathbf{x}$, явным образом проверив положительную определённость матрицы Гессе методом Сильвестра, но не более того. Было бы интересно узнать, как эту задачу предполагалось решать.
Я также выяснил, что в общем виде, когда логарифмический барьер задан как

$$\varphi(\mathbf{x}, t) := -\log(-h(\mathbf{x}, t))$$

где функция h выпукла, он, очевидно, является выпуклой функцией (как композиция выпуклой невозрастающей и вогнутой, см. Бойда, стр. 84), а его градиент и матрица Гессе запишутся так:

$$\nabla\varphi(\mathbf{x}, t) = \frac{1}{h(\mathbf{x}, t)}\nabla h(\mathbf{x}, t)$$
$$\nabla^2\varphi(\mathbf{x}, t) = \frac{1}{h^2(\mathbf{x}, t)}\nabla h(\mathbf{x}, t)\nabla h(\mathbf{x}, t)^T - \frac{1}{h(\mathbf{x}, t)}\nabla^2 h(\mathbf{x}, t)$$

Из этих формул очевидно доказательство теоремы о композиции из Бойда, но в данном случае формула для матрицы Гессе не помогает, т.к. $\nabla h(\mathbf{x}, t)\nabla h(\mathbf{x}, t)^T$ — положительно полуопределённая матрица, а $\nabla^2 h(\mathbf{x}, t)$ — отрицательно определённая, и почему их сумма будет положительно определённой неясно. □

6 Кратчайший путь в графе

Исследовать на выпуклость/вогнутость функцию $p_{ij}(\mathbf{c})$ — длину кратчайшего пути между вершинами i и j во взвешенном ориентированном графе G с вектором весов $\mathbf{c} \in \mathbb{R}^{|E(G)|}$.

Решение. Рассмотрим все простые пути между вершинами i и j в графе G и сопоставим k-му из них функцию $\varphi_k(\mathbf{c})$ — его стоимость при векторе весов $\mathbf{c}$. Каждая такая функция линейна по $\mathbf{c} : \varphi_k(\alpha\mathbf{c}) = \alpha\varphi_k(\mathbf{c})$. $p_{ij}(\mathbf{c})$, в свою очередь, естественно выражается через них:

$$p_{ij}(\mathbf{c}) = \min_k \varphi_k(\mathbf{c})$$

Иными словами, подграфик $p_{ij}(\mathbf{c})$ есть пересечение выпуклых множеств, задаваемых подграфиками линейных функций φ_k. Поскольку пересечение сохраняет выпуклость, p_{ij} **вогнутая**. Можно было сказать и короче: операция взятия поточечного минимума сохраняет вогнутость.
Заметим, что выпуклой p_{ij} быть не может, что демонстрирует следующий пример: пусть $|V(G)| = 4,\ E(G) = \{(1,2),(1,3),(2,4),(3,4)\},\ \mathbf{c}_1 = (0,2,0,2),\ \mathbf{c}_2 = (2,0,2,0),\ \theta = \frac{1}{2}$. Тогда

$$p_{ij}(\theta\mathbf{c}_1 + (1-\theta)\mathbf{c}_2) = 2 > \theta p_{ij}(\mathbf{c}_1) + (1-\theta)p_{ij}(\mathbf{c}_2) = 0$$

□