Задание №3.1. Расхождение Кульбака-Лейблера и максимальное правдоподобие

Расхождение Кульбака-Лейблера (KL divergence) является мерой, показывающей, как одно распределение вероятностей отличается от другого. Оно появилось в теории информации, но нашло свое применение во многих других областях, таких как, в частности, статистика, машинное обучение, информационная геометрия и многие другие. В машинном обучении КЛ-расхождение играет существенную роль, объединяя различные концепции, которые на первый взгляд могут показаться несвязанными.

В этом задании мы рассмотрим расхождение Кульбака-Лейблера для дискретных распределений, попрактикуемся в некоторых простых манипуляциях с ним и посмотрим на его связь с оценкой максимального правдоподобия.

КЛ-расхождение между двумя дискретными распределениями $P(X)$ и $Q(X)$, заданными над множеством элементарных исходов $\mathcal{X}$, определяется следующим образом[1]:

$$D_{KL}(P||Q) = \sum_{x\in\mathcal{X}} P(x)\log\frac{P(x)}{Q(x)}.$$

Для удобства рассуждений будем полагать, что $Q(x) > 0, \forall x$. Еще одно часто используемое соглашение, это то, что $0\log 0 = 0$.

Чтобы понять, что означает КЛ-расхождение, достаточно вспомнить определения энтропии:

$$H(P) = -\sum_{x\in\mathcal{X}} P(x)\log P(x)$$

и перекрестной энтропии:

$$H(P,Q) = -\sum_{x\in\mathcal{X}} P(x)\log Q(x).$$

Тогда становится понятно, что КЛ-расхождение – это разница между ними:

$$D_{KL}(P||Q) = \sum_{x\in\mathcal{X}} P(x)\log\frac{P(x)}{Q(x)} = \sum_{x\in\mathcal{X}} P(x)\log P(x) - \sum_{x\in\mathcal{X}} P(x)\log Q(x) = H(P,Q) - H(P).$$

То есть это разница в среднем количестве бит на символ, которую мы платим, если будем кодировать сообщения, распределенные по $P(X)$, с помощью оптимальной схемы кодирования, построенной для распределения $Q(X)$.

Если перекрестная энтропия между P и Q равна $H(P)$ (и как следствие $D_{KL}(P||Q) = 0$), то это означает, что $P = Q$. В машинном обучении часто бывает, что нам надо найти распределение Q, которое будет как можно «ближе» к другому распределению P. Для этих целей часто пользуются расхождением Кульбака-Лейблера в качестве функции потерь. Как мы увидим в этом упражнении, оценка максимального правдоподобия, которая часто является целью оптимизации, оказывается эквивалентной минимизации КЛ-расхождения между учебной выборкой (то есть эмпирическим распределением) и моделью.

Теперь покажем основные свойства КЛ-расхождения.

---

[1] Если P и Q являются плотностями непрерывных распределений, то в определении KL-расхождения суммирование заменяется на интегрирование. Все остальное остается без изменений.

Вопрос №1. Неотрицательность [1 балл]

Докажите, что $\forall P, Q$

$$D_{KL}(P||Q) \geq 0$$

и что

$$D_{KL}(P||Q) = 0 \text{ тогда и только тогда, когда } P = Q.$$

Подсказка. Вы можете воспользоваться неравенством Йенсена: если f – это выпуклая функция, а X – случайная величина, то $E[f(X)] \geq f(E[X])$. Более того, если f строго выпуклая (f выпуклая, если ее Гессиан $H \geq 0$; она строго выпуклая, если $H > 0$, например, $f(x) = -\log x$ – это строго выпуклая функция), тогда $E[f(X)] = f(E[X])$ подразумевает, что $X = E[X]$, т.е. что X – это константа.

Вопрос №2. Цепное правило для КЛ-расхождения [2 балла]

КЛ-расхождение между двумя условными распределениями определяется следующим образом:

$$D_{KL}\big(P(X|Y) \mid\mid Q(X|Y)\big) = \sum_y P(y) \left(\sum_{x \in \mathcal{X}} P(x|y) \log \frac{P(x|y)}{Q(x|y)} \right).$$

Его можно интерпретировать как ожидаемое КЛ-расхождение между соответствующими условными распределениями по x (то есть между $P(X|Y = y)$ и $Q(X|Y = y)$), где матожидание берется по y.

Докажите следующее цепное правило для КЛ-расхождения:

$$D_{KL}\big(P(X|Y) \mid\mid Q(X|Y)\big) = D_{KL}(P(X) \mid\mid Q(X)) + D_{KL}\big(P(Y|X) \mid\mid Q(Y|X)\big).$$

Вопрос №3. КЛ-расхождение и максимальное правдоподобие [2 балла]

Рассмотрим задачу определения плотности распределения и предположим, что нам дана учебная выборка $\{x^{(i)}; i = 1, \dots, m\}$. Пусть эмпирическое распределение есть $\widehat{\mathrm{P}}(x) = \frac{1}{m}\sum_{i=1}^{m} 1\{x^{(i)} = x\}$. $\hat{P}$ это просто равномерное распределение на учебной выборке.

Предположим, что у нас есть семейство распределений P_θ, параметризованное θ. Если хотите, можете считать P_θ альтернативным обозначением для $P(x; \theta)$. Докажите, что нахождение оценки максимального правдоподобия для θ эквивалентно нахождению P_θ с минимальным расхождением Кульбака-Лейблера от $\hat{P}$, то есть:

$$\operatorname*{argmin}_\theta D_{KL}(\hat{P} \mid\mid P_\theta) = \operatorname*{argmax}_\theta \sum_{i=1}^{m} \log P_\theta\big(x^{(i)}\big).$$

Замечание. Посмотрите на то, как связаны подзадачи 2 и 3 и оценка параметров модели наивного Байеса. В модели наивного Байеса мы предполагали следующую форму P_θ:

$$P_\theta(x, y) = p(y) \prod_{i=1}^{d} p(x_i|y).$$

Используя цепное правило для КЛ-расхождения, имеем:

$$D_{KL}\big(\hat{P} \mid\mid P_\theta\big) = D_{KL}\left(\hat{P}(y) \mid\mid p(y)\right) + \sum_{i=1}^{d} D_{KL}\left(\hat{P}(x_i|y) \mid\mid p(x_i|y)\right).$$

Это означает, что поиск максимального правдоподобия/минимума КЛ-расхождения параметров раскладывается на $2n + 1$ независимых оптимизационных задач: одну для априорного распределения классов $p(y)$ и по одной для каждого условного распределения $p(x_i|y)$ для каждого признака x_i при одном из двух возможных значений y. В частности, независимый поиск оценок максимального правдоподобия для каждой из этих задач приводит также к максимизации правдоподобия совместного распределения.