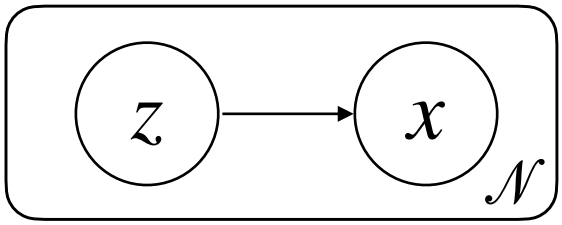
Функция потерь вариационного автоэнкодера в деталях

Вероятностный подход

Давайте для начала определим вероятностную графическую модель для описания наших данных. За x обозначим переменную, которая представляет наши данные, и будем предполагать, что x генерируется из скрытой переменной z (закодированное представление), которая напрямую не наблюдается. Таким образом, для каждой точки данных предполагается следующий процесс генерации, состоящий из двух шагов:

- сначала латентное представление z выбирается из предварительного распределения $p(z)$;
- затем данные x выбираются из распределения условного правдоподобия $p(x \mid z)$.



Имея в виду такую вероятностную модель, мы можем рассмотреть вероятностные версии энкодера и декодера. «Вероятностный декодер» естественным образом определяется как $p(x \mid z)$, то есть описывает распределение декодированной переменной с учетом закодированной. Тогда как «вероятностный энкодер» определяется как $p(z \mid x)$, то есть описывает распределение закодированной переменной с учетом декодированной.

На этом этапе мы уже можем заметить, что регуляризация скрытого пространства, которой нам не хватало в простых автокодировщиках, естественным образом проявляется здесь в определении процесса генерации данных: предполагается, что закодированные представления z в скрытом пространстве соответствуют априорному распределению $p(z)$. В противном случае мы можем использовать теорему Байеса, которая устанавливает связь между априорным распределением $p(z)$, правдоподобием $p(x \mid z)$ и апостериорным распределением $p(z \mid x)$:

$$p(z \mid x) = \frac{p(x \mid z)p(z)}{p(x)}.$$

Теперь предположим, что $p(z)$ - стандартное гауссовское распределение, а $p(x|z)$ - гауссовское распределение, среднее значение которого определяется детерминированной функцией f от переменной z, а ковариационная матрица имеет вид cI, где c – положительная константа, I – единичную матрицу. Предполагается, что функция f принадлежит семейству функций, обозначенному F. Таким образом, мы имеем

$$p(z) = \mathcal{N}(0, I)$$
$$p(x|z) = \mathcal{N}(f(z), cI),\ f \in F,\ c > 0$$

Вариационный вывод

В статистике вариационный вывод **(VI)** - это метод аппроксимации сложных распределений. Идея состоит в том, чтобы установить параметризованное семейство распределений (например, семейство гауссиан, параметры которого являются средним значением и ковариацией) и найти наилучшее приближение к нашему целевому распределению среди этого семейства.

Давайте будем аппроксимировать $p(z|x)$ гауссовым распределением $q_x(z)$, среднее значение и ковариация которого определяются двумя функциями, μ и σ. Предполагается, что эти две функции принадлежат, соответственно, к семействам функций M и Σ. Таким образом, мы можем обозначить

$$q_x(z) = \mathcal{N}(\mu(x), \sigma(x)),\ \ \mu \in M,\ \ \sigma(x) \in \Sigma.$$

Итак, мы определили семейство кандидатов для вариационного вывода, и теперь нам нужно найти наилучшее приближение среди этого семейства путем оптимизации функций μ и σ. (фактически, их параметров), чтобы минимизировать расхождение Кульбака-Лейблера между приближением $q_x(z)$ и $p(z|x)$. Другими словами, мы ищем оптимальные μ^* и σ^* такие, что

$$(\mu^*, \sigma^*) = \underset{(\mu,\sigma)\in M\times\Sigma}{\operatorname{argmin}}\ KL(q_x(z), p(z|x)) =$$

$$= \underset{(\mu,\sigma)\in M\times\Sigma}{\operatorname{argmin}}\ (E_{z\sim q_x}(\log q_x(z)) - E_{z\sim q_x}(\log \frac{p(x|z)p(z)}{p(x)}))$$

$$= \underset{(\mu,\sigma)\in M\times\Sigma}{\operatorname{argmin}} (E_{z\sim q_x}(\log q_x(z)) - E_{z\sim q_x}(\log p(x\,|\,z)) - E_{z\sim q_x}(\log p(z)) + E_{z\sim q_x}(\log p(x))$$
$$= \underset{(\mu,\sigma)\in M\times\Sigma}{\operatorname{argmax}} (E_{z\sim q_x}(\log p(x\,|\,z)) - KL(q_x(z), p(z)))$$

В последнем уравнении мы можем наблюдать следующий компромисс: мы хотим максимизировать вероятность «наблюдений» (максимизация ожидаемого логарифмического правдоподобия для первого члена) и в то же время сохранить близкое к априорному распределение (минимизация расхождения KL между $q_x(z)$ и $p(z)$ для второго члена).

Давайте убедимся, что, мы можем получить для любой функции f из F (каждая из которых определяет свой вероятностный декодер $p(x\,|\,z)$) наилучшее приближение к $p(z\,|\,x)$, обозначенное $q_x^*(z)$. Несмотря на ее вероятностный характер, мы ищем как можно более эффективную схему кодирования-декодирования, а затем мы хотим выбрать функцию f, которая максимизирует ожидаемую логарифмическую правдоподобие x при заданном z, когда z выбирается из $q_x^*(z)$.

Другими словами, для заданного входа x мы хотим максимизировать вероятность, чтобы получить оценку $\hat{x} = x$, когда мы выбираем z из распределения $q_x^*(z)$, а затем выбираем $\hat{x}$ из распределения $p(x\,|\,z)$. Таким образом, мы ищем оптимальное $f*$ такое, что

$$f* = \underset{f\in F}{\operatorname{argmax}}(E_{z\sim q_x^*}(\log p(x\,|\,z))$$
$$= \underset{f\in F}{\operatorname{argmax}}(E_{z\sim q_x^*}(-\frac{||x - f(z)||^2}{2c})).$$

Настало время, собрать все вместе! Мы ищем такие оптимальные $f*, \mu*, \sigma*$, что

$$(f*, \mu*, \sigma*) = \underset{(f,\mu,\sigma)\in F\times M\times\Sigma}{\operatorname{argmax}} (E_{z\sim q_x^*}(-\frac{||x - f(z)||^2}{2c}) - KL(q_x(z), p(z))).$$

Перевод на язык Машинного Обучения

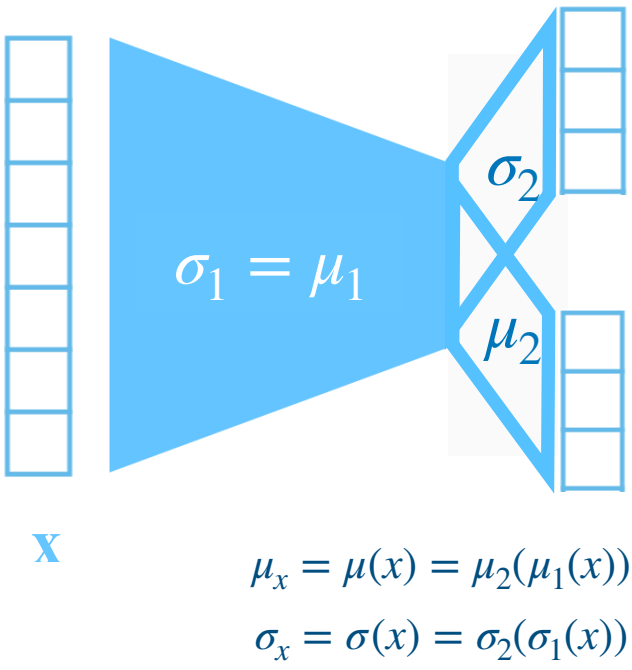
Итак, мы создали вероятностную модель, которая зависит от трех функций, f, μ, σ. Поскольку мы не можем легко оптимизировать все пространство функций, мы

ограничиваем область оптимизации и выражаем f, μ, σ как нейронные сети. Таким образом, F, M, Σ соответствуют семействам функций, определяемых архитектурами сетей, а оптимизация выполняется по параметрам этих сетей.
На практике μ и σ не определяются двумя полностью независимыми сетями, а разделяют одну архитектуру, так что мы имеем

$$\mu(x) = \mu_2(\mu_1(x)) \quad \sigma(x) = \sigma_2(\sigma_1(x)) \quad \mu_1(x) = \sigma_1(x)$$

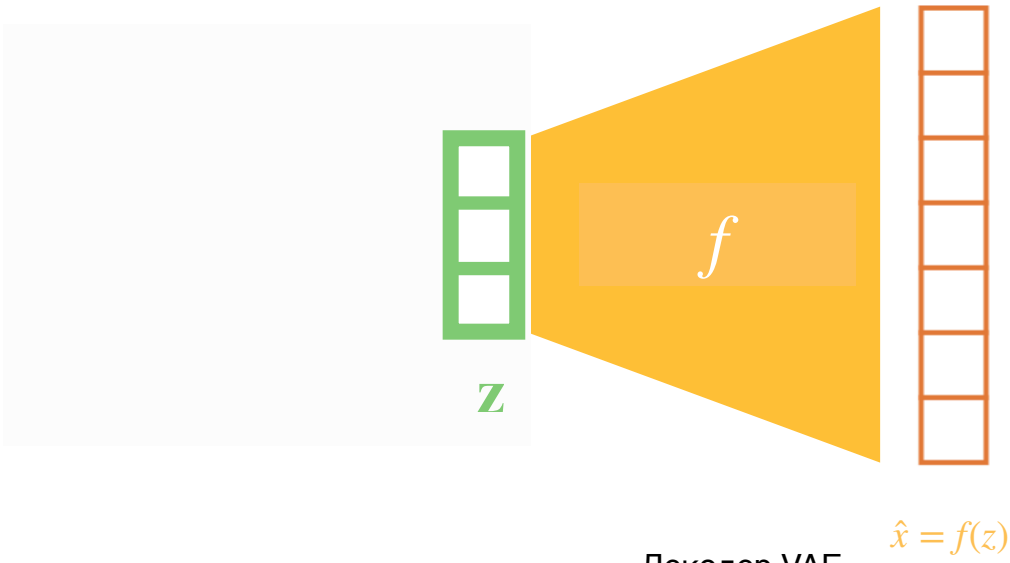
Поскольку $\sigma(x)$ определяет ковариационную матрицу $q_x(z)$, предполагается, что $\sigma(x)$ является квадратной матрицей. Однако, чтобы упростить вычисления и уменьшить количество параметров, мы делаем дополнительное предположение, что наша аппроксимация $q_x(z)$ является многомерным распределением Гаусса с диагональной ковариационной матрицей. При таком предположении $\sigma(x)$ - это просто вектор диагональных элементов ковариационной матрицы, и тогда он имеет тот же размер, что и $\mu(x)$.

Однако мы сокращаем таким образом семейство распределений, которое мы рассматриваем для вариационного вывода, и, таким образом, полученная аппроксимация для $p(z \mid x)$ может быть менее точной.



Энкодер VAE

В отличие от кодирующей части, которая моделирует $p(z|x)$ и для которой мы рассматривали гауссиан со средним значением и ковариацией как функциями от x и (μ, σ), наша модель предполагает для $p(x|z)$ гауссиан с фиксированной ковариацией. Функция $f(z)$, определяющая среднее значение этого гауссовского процесса, моделируется нейронной сетью и может быть представлена следующим образом:



Декодер VAE

Общая архитектура затем получается путем объединения энкодера и декодера. Однако нам по-прежнему нужно быть очень осторожными с тем, как мы сэмплируем из распределения, возвращаемого кодировщиком во время обучения. Процесс сэмплирования должен быть выражен таким образом, чтобы ошибка могла распространяться по сети. Чтобы сделать градиентный спуск возможным, несмотря на случайную выборку, которая происходит на полпути архитектуры, используется простой трюк, называемый **трюком репараметризации**, и заключается в использовании того факта, что если z - случайная величина, следующая гауссовскому распределению со средним $\mu(x)$ и с ковариацией $\sigma(x)$, то z можно выразить так:

$$z = \mu(x) + \sigma(x)\zeta, \;\; \zeta \sim \mathcal{N}(0, I)$$

Следовательно, функция потерь VAE примет вид:

$$loss = C||x - \hat{x}||^2 + KL[N(\mu_x, \sigma_x), N(0,1)] = C||x - f(z)||^2 + KL[N(\mu(x), \sigma(x)), N(0,1)]$$