Машинное обучение

Минкин Даниэль

28 июля 2025 г.

Содержание

1 Функции потерь

1.1 Задача регрессии

1.1.1 MSE

$$MSE(f, X, y) = \frac{1}{N}\|f(X) - y\|^2_{euclidean} \tag{1}$$

Градиент

Раскроем наше выражение

$$\|f(X) - y\|^2 = \langle f(X) - y, f(X) - y\rangle = \langle f(X), f(X)\rangle - 2\langle f(X), y\rangle + \langle y, y\rangle \tag{2}$$

Возьмем дифференциал от данного выражения

$$[D_{f(X)}(\|f(X)-y\|)] = [D_{f(X)}(\langle f(X), f(X)\rangle)]-2[D_{f(X)}(\langle f(X), y\rangle)]+[D_{f(X)}(\langle y, y\rangle)] \tag{3}$$

Мы можем исключить последее слагаемое

$$[D_{f(X)}(\langle f(X), f(X)\rangle)] - 2[D_{f(X)}(\langle f(X), y\rangle)] \tag{4}$$

Для удобства заменим $f(X)$ на f. Мы получим выражение равное:

$$[D_f(\langle f, f\rangle)] - 2[D_f(\langle f, y\rangle)] \tag{5}$$

Рассмотрим дифференциал второго слагаемого

$$y^T(f + \Delta f) - y^T f = y^T \Delta f = \Delta f^T y \tag{6}$$

Следовательно:

$$[D_f(\langle f, y\rangle)] = y \tag{7}$$

Рассмотрим дифференциал первого выражения, заметим, что по свойству симметричности произвдения и правилу дифференцирования умножения мы получим

$$[D_f(\langle f, f\rangle)] = 2\langle [D_f(f)], f\rangle \tag{8}$$

Легко показать, что

$$[D_f(f)] = E \tag{9}$$

Следовательно наше выражение приводимо к

$$[D_f(\langle f, f\rangle)] = 2\langle E, f\rangle = 2f \tag{10}$$

Приведем все это к одной формуле

$$2f(X) - 2y = 2(f(X) - y) \tag{11}$$

А теперь поделим все на N, где N — размер выборки, так как это монотонное преобразование не зависящее от $f(X)$, оно не влияет на дифференциал

Таким образом:

$$[D_{f(X)}(MSE)] = \frac{2}{N}(f(X) - y) \tag{12}$$

1.1.2 MAE

$$MAE(f, X, y) = \frac{1}{N}\langle \text{sign}(f(X) - y), f(X) - y\rangle \tag{13}$$

Градиент

Уберем деление на N, так как это не влияет на дифференциал, мы добавим его в конце

$$[D_{f(X)}(MAE)] = \frac{1}{N}[D_{f(X)}(\langle \text{sign}(f(X) - y), f(X) - y\rangle)] \tag{14}$$

Как и в прошлом случае заменим $f(X)$ на f и начнем раскрывать выражение. Тогда

$$\begin{gathered}
[D_f(\langle \text{sign}(f - y), f - y\rangle)] \\
= \langle [D_f(\text{sign}(f - y))], f - y\rangle + \langle \text{sign}(f - y), [D_f(f - y)]\rangle \\
= \langle 0, f - y\rangle + \langle \text{sign}(f - y), E\rangle \\
= \text{sign}(f - y)
\end{gathered}$$

Если Δf — вектор строка

Следовательно:

$$[D_{w_0}(MAE)](\Delta f) = \frac{1}{N}\Delta f \cdot \text{sign}(f - y) \tag{15}$$

1.1.3 MAPE

$$MAPE(f, X, y) = \frac{1}{N} \operatorname{sign}(f(X) - y)^* \cdot \operatorname{diag}(\frac{1}{|y_1|}, \frac{1}{|y_2|}, \frac{1}{|y_3|}...) \cdot (f(X) - y) \quad (16)$$

Градиент

По аналогии с MAE

$$[D_{f(X)}(MAPE)] = \frac{1}{N} \cdot \operatorname{diag}(\frac{1}{|y_1|}, \frac{1}{|y_2|}, \frac{1}{|y_3|}...) \cdot \operatorname{sign}(f(X) - y) \quad (17)$$

1.1.4 Huber loss

$$Huber(f, X, y) = \frac{1}{N} \sum_{i=1}^{N} h_\delta(y_i - f(X_i)) \quad (18)$$

при этом

$$h_\delta(z) = \begin{cases} 0.5z^2 & |z| \leq \delta \\ \delta(|z| - 0.5\delta) & |z| > \delta \end{cases}$$

Данная функция потерь более устойчива к выбросам как MAE, но в окрестности нуля ведет себя как MSE.

$$\frac{dh}{dz}(z) = \begin{cases} z & |z| \leq \delta \\ \delta \operatorname{sign}(z) & |z| > \delta \end{cases}$$

Градиент

$$[D_{f(X)}(Huber)] = \frac{1}{N} \frac{dh}{d\delta}^\circ (y - f(X)) \quad (19)$$

Т.е мы поэлементно применяем операцию дифференцирования

2 Метрики

2.1 Задачи классификации

2.1.1 ROC-AUC

ROC-AUC - площадь под ROC кривой, опишем алгоритм ее построения. Это кривая, позволяющая проверять вероятностные модели классификации, по оси Y расположен **true positive rate** или, а по оси X **false positive rate**

Для начала поговорим про TPR и FPR. У нас есть гиперпараметр, меняя который можно регулировать кол-во TP и FP. Данный параметр часто

называется **точкой отсечения** (**cut-off value**) — некая величина p_{limit} такая, что если $p_{predicted} > p_{limit}$ то считается, что объект принадлежит к классу 1

Таблица 1: Исходы предсказания

	Предсказано: Да	**Предсказано: Нет**
Фактически: Да	True Positive (TP)	False Negative (FN)
Фактически: Нет	False Positive (FP)	True Negative (TN)

$$TPR = \frac{TP}{TP + FN} = 1 - \frac{FN}{TP + FN} \tag{20}$$

Мы смотрим распределение правильно угаданных внутри группы, где фактически объекты принадлежат к классу 1

$$FPR = \frac{FP}{FP + TN} \tag{21}$$

Мы смотрим на распределение неправильно угаданных среди группы, где фактически объекты принадлежат к классу 0

Алгоритм построения ROC кривой выглядит следующим образом:

1. Изначально у нас дано неупорядоченное множество пар
   $M = \{(y\_true_0, pred\_prob_0), (y\_true_1, pred\_prob_1)...\}$
2. Введем упорядоченное множество T, состоящее из возможных порогов от 0 до 1, отсортированное по убыванию
3. Для каждого $t \in T$ получим множество
   $M_t = \{(y\_true_0, \mathrm{I}(pred\_prob_0 \geq t)), (y\_true_1, \mathrm{I}(pred\_prob_1 \geq t))...\}$
4. Для каждого M_t рассчитаем TPR и FPR, это и будут точки (x, y) для кривой

Интерпретация ROC AUC. Мы можем считать значение ROC AUC равным вероятности того, что оценка случайно выбранного примера положительного класса будет выше оценки случайно выбранного примера отрицательного класса. Пусть у нас есть функция $g(x) : C \to [0, 1]$, определенная на измеримом множестве C с мерой μ, и два множества A - положительные классы и B - отрицательные классы, такие что $A \cup B = C$ и $A \cap B = \varnothing$. Т.е пространство поделено на области для положительных и отрицательных классов. Пусть есть две случайные величины $A^* \sim \mathrm{Uniform}(A)$ и $B^* \sim \mathrm{Uniform}(B)$. Нам нужно найти $\mathrm{E}[\mathrm{I}(g(A^*) > g(B^*))] = P(g(A^*) > g(B^*))$. Заметим, что $g(A^*)$ и $g(B^*)$ - непрерывные случайные величины, так как являются результатом применения детерминированной функции к случайным величинам. Обозначим их как A_g и B_g. Зададим функции распределения.

$$F_{A_g}(k) = \frac{\mu(\{x \in A | g(x) \leq k\})}{\mu(A)} \tag{22}$$

$$F_{B_g}(k) = \frac{\mu(\{x \in B | g(x) \leq k\})}{\mu(B)} \tag{23}$$

Мы можем перейти от функции распределения к PDF. Запишем заданную вероятность, выраженную через PDF.

$$P(A_g > B_g) = \int_0^1 f_{B_g}(b) \int_b^1 f_{A_g}(a)\ da\ db \tag{24}$$

Это равносильно

$$P(A_g > B_g) = \int_0^1 f_{B_g}(b) \cdot (1 - F_{A_g}(b))\ db = 1 - \int_0^1 f_{B_g}(b) \cdot F_{A_g}(b)\ db \tag{25}$$

Теперь посмотрим на ROC-AUC в общем виде. У нас есть кривая (FPR(t), TPR(t)). FPR в общем виде задается как $\frac{\mu(\{x \in B | f(x) \geq t\})}{\mu(B)} = 1 - F_{B_g}(t)$, а TPR как $\frac{\mu(\{x \in A | f(x) \geq t\})}{\mu(A)} = 1 - F_{A_g}(t)$. Следовательно ROC-AUC $= -\int_0^1 TPR(t)FPR^{'}(t)dt$ (по правилу интегрирования параметрически заданной кривой)

$$\text{ROC-AUC} = \int_0^1 (1 - F_{A_g}(t)) f_{B_g}(t) dt = 1 - \int_0^1 F_{A_g}(t) f_{B_g}(t) dt \tag{26}$$

Выражению равносильно тому, что мы вывели выше Ч.Т.Д

2.1.2 CAP curve и CAP-AUC

3 Линейные модели

Линейные модели — класс моделей которые используют линейное преобразование для вектора входных фичей.

3.1 Общее

3.1.1 Регрессия

Формализуем задачу регрессии, пусть у нас есть вектор $\bar{x} \in \mathrm{R}^n$. Тогда предсказание может быть сделано с помощью такой формулы:

$$y_{pred} = \bar{x} \cdot \bar{w} + w_0 \tag{27}$$

Т.е мы ищем такой вектор $\bar{w} \in \mathrm{R}^{n+1}$, который будет выдавать наиболее близкие y_{pred} к y_{true}.

3.1.2 Классификация

В случае решения задачи классификации через линейные модели сначала делается предсказание как в случае регрессии, а после к результату предсказания применяется разделяющее правило (например в случае бинарной классификации правило может задаваться так: если $y_{pred} > 0$ мы относим объект к положительному классу — если нет, то к отрицательному)

В случае бинарной классификации с разделяющим правилом из примера уравнение регрессии задает гиперплоскость, которая разделяет исходное пространство: i.e $\sum_{i=1}^{n} w_i \cdot x_i + c > 0$ — плоскость в n-мерном пространстве, которая делит пространство на положительный и отрицательные классы

3.1.3 Прочее

Несколько фактов:

- **При использовании OneHot Encoding-а мы можем избавиться от одной encoded фичи**. Все просто: пусть у нас есть веса для каждой закодированной фичи $w_1, w_2 ... w_n$, также у нас добавляется константа c к предсказанию по фичам. Давайте удалим последнюю фичу, тогда в случае если $x_1 = 0$ ... $x_{n-1} = 0$ нам нужно добавить к константе еще и w_n, иначе результат изменится, следовательно константа в новой модели должна быть равна $w_n + c$. Однако тогда нам нужно внести поправку в веса в случае если один из $x_i \mid i < n$ равен 1, чтобы результат остался таким же. Мы можем просто вычесть из старых весов w_n, за счет того, что мы добавили его к const ничего не изменится. Таким образом мы успешно исключили одну encoded фичу, оставив результаты предсказаний без изменений. **Важно заметить, что если мы работаем в модели без константы, то тогда данный подход не сработает**

- **Для более сложных зависимостей необходимо использовать новые фичи которые являются функциями от старых**. Т.е мы включаем в модель фичи задаваемые как $f(x_1, x_2 ... x_n)$

- Если между признаками есть приближённая линейная зависимость, коэффициенты в линейной модели могут совершенно потерять физический смысл

3.2 Оценка по МНК

3.2.1 Общее

В первую очередь опустим свободный член, так как можно считать, что у нас просто есть еще один признак, который всегда равен 1. Пусть функция потерь задается как Евклидова норма между предсказанными и истинными значениями. I.e мы решаем следующую задачу:

$$\|Xw - b\|_{euclidean} \to \min_w \tag{28}$$

где X — матрица размера (N, k), N – размер выборки, а k — кол-во фичей, т.е это матрица где в строки записаны вектора, по которым нужно сделать предсказания.

Однако нам также нужно сделать поправку на размер выборки, чтобы значения функции потерь можно было сравнивать между собой для разных выборок. Получается задача выгладит так:

$$\frac{\|Xw - b\|_{euclidean}}{N} \to \min_w \tag{29}$$

Функция потерь является функционалом, так как принимает на вход три значения — матрицу наблюдений, true значения предсказываемой перменной и функцию, которая возвращает некое значение по вектору наблюдений

Значение коэффициентов может быть получено через псевдообратную матрицу, по ее свойству:

$$\|XX^+b - b\|_{euclidean} \leq \|Xw - b\|_{euclidean} \tag{30}$$

для любого w

Так как деление на константу — монотонное преобразование, данное решение минимизирует нашу функцию потерь

> **Важно заметить:** псевдообратная матрица может быть использована и для других норм, однако тогда нам нужно перейти в евклидово пространство A с новым скалярным произведением, тогда псевдообратная матрица будет минимизировать норму задаваемую как $\sqrt{\langle u, u\rangle_A}$. Об этом будет рассказано позже

Таким образом решением данного СЛАУ будет

$$w_* = A^+b \tag{31}$$

В случае если $N \geq k$

$$w_* = (X^TX)^{-1}X^Tb \tag{32}$$

А в противном случае, когда $N < k$

$$w_* = X^T(XX^T)^{-1}b \tag{33}$$

Мы можем считать, что X — матрица полного столбцового или строчного ранга, зачастую у нас не будет полностью ЛЗ столбцов или строк, а даже если они и есть их можно исключить из-за бессмысленности

3.2.2 Невырожденность матрицы X^TX

В реальных задачах матрицах X^TX или XX^T являются невырожденными, однако нам нужно оценить насколько они "невырождены", так как у нас есть погрешность при вычислении детерминанта, например мы могли получить неотрицательное значение из-за логики работы чисел с плавающей точкой или же в изначальных данных содержится погрешность.

Число обусловленности

Число обусловленности — максимальное значение отношения относительного изменения функции и относительного измененииия арргмента

$$\mu(f,x) = \max_{\Delta x} \frac{\frac{\|f(x+\Delta x)-f(x)\|}{\|f(x)\|}}{\frac{\|\Delta x\|}{\|x\|}} \tag{34}$$

в малой окрестности Δx

Легко заметить, что чем больше значение $\mu(f,x)$, тем хуже так как наше решение может очень сильно измениться от малого приращения аргумента, а мы бы хотели иметь "стабильное" решение

Покажем на зависимость числа обусловленности задачи $f(A) = (A^TA)^{-1}$ от матрицы корреляции. Мы воспользуемся определением числа обусловленности:

$$\mu(f,A) = \max_{\Delta A} \frac{\frac{\|f(A+\Delta A)-(A^TA)^{-1}\|}{\|(A^TA)^{-1}\|}}{\frac{\|\Delta A\|}{\|A\|}} \tag{35}$$

Тут нам понадобится матричное дифференцирование, нам нужно найти дифференциал $(A^TA)^{-1}$, заметим, что это композиция двух функций. Во-первых, вспомним определение дифференциала, пусть задана функция $f : R^n \to R^m$ Тогда ее дифференциал может быть найден так:

$$f(x_0 + \Delta x) - f(x_0) = [D_{x_0} f](\Delta x) + o(\|\Delta x\|) \tag{36}$$

При этом дифференциал $[D_{x_0} f]$ — линейное отображение из R^n в R^m, т.е мы находим линейную часть приращения отображения

Найдем дифференциал $f(X) = X^TX$ где X - матрица (n,m) и $n > m$. Тогда:

$$\begin{aligned} f(X+\Delta X) - f(X) &= (X+\Delta X)^T(X+\Delta X) - X^TX \\ &= (X^T + (\Delta X)^T)(X+\Delta X) - X^TX \\ &= X^TX + X^T\Delta X + (\Delta X)^TX + (\Delta X)^T\Delta X - X^TX \\ &= X^T\Delta X + (\Delta X)^TX + (\Delta X)^T\Delta X \end{aligned}$$

Можно легко показать, что $[D_X f](\Delta X) = X^T \Delta X + (\Delta X)^T X$ является линейным оператором по ΔX. При этом $(\Delta X)^T \Delta X$ и есть $o(\|\Delta X\|)$.

Найдем дифференциал $g(X) = X^{-1}$. Мы будем использовать равенство $E = X^{-1}X$. Продифференцируем выражение с обеих сторон.

$$0 = [D_{X_0}(X^{-1}X)](H) \tag{37}$$

По правилам дифференцирования умножения мы получаем

$$0 = [D_{X_0}(X^{-1})](H) \cdot X_0 + X_0^{-1} \cdot [D_{X_0}(X)](H) \tag{38}$$

Таким образом

$$[D_{X_0}(X^{-1})](H) \cdot X_0 = -X_0^{-1} \cdot [D_{X_0}(X)](H) \tag{39}$$

Т.е все сводится к формуле

$$[D_{X_0}(X^{-1})](H) = -X_0^{-1} \cdot [D_{X_0}(X)](H) \cdot X_0^{-1} \tag{40}$$

При этом $[D_{X_0}(X)](H) = H$, таким образом итоговая формула имеет вид

$$[D_{X_0}(X^{-1})](H) = -X_0^{-1} \cdot H \cdot X_0^{-1} \tag{41}$$

Мы имеем дело с функцией $g(f(X))$, дифференциал такой функции представим как

$$[D_{f(X_0)}(g)]([D_{X_0}(f)](\Delta X)) \tag{42}$$

Следовательно $[D_{X_0}((X^T X)^{-1})](\Delta X)$, может быть найдено так

$$(X_0^T X_0)^{-1} \cdot (X_0^T \Delta X + (\Delta X)^T X_0) \cdot (X_0^T X_0)^{-1} \tag{43}$$

Таким образом в изначальной формуле числа обусловленности мы можем записать приращение функции как

$$\mu(A) = \max_{\Delta A} \frac{\frac{\|(A^TA)^{-1} \cdot (A^T \Delta A + (\Delta A)^T A) \cdot (A^TA)^{-1}\|}{\|(A^TA)^{-1}\|}}{\frac{\|\Delta A\|}{\|A\|}} \tag{44}$$

Оценим наше выражение сверху, сделаем несколько предположений на будущее, **пусть мы используем одно и тоже семейство норм для входного и выходного пространства и для этого семейства выполняется свойство субмультипликативности**, тогда для $(A^TA)^{-1}$ и $(A^T \Delta A + (\Delta A)^T A)$ можно будет использовать следующее мажорирование

$$\|(A^TA)^{-1}(A^T \Delta A + (\Delta A)^T A)(A^TA)^{-1}\| \leq \|(A^TA)^{-1}\|^2 \|A^T \Delta A + (\Delta A)^T A\| \tag{45}$$

В итоге мы можем мажорировать наше выражение так:

$$\frac{\frac{\|(A^TA)^{-1}\cdot(A^T\Delta A+(\Delta A)^TA)\cdot(A^TA)^{-1}\|}{\|(A^TA)^{-1}\|}}{\frac{\|\Delta A\|}{\|A\|}} \leq$$
$$\frac{\|(A^TA)^{-1}\|\cdot\|A^T\Delta A+(\Delta A)^TA\|}{\frac{\|\Delta A\|}{\|A\|}}$$

Продолжая мажорирование мы получим следующую ситуацию

$$\frac{\|(A^TA)^{-1}\|\cdot\|A^T\Delta A+(\Delta A)^TA\|}{\frac{\|\Delta A\|}{\|A\|}} \leq$$
$$\frac{\|(A^TA)^{-1}\|\cdot(\|A^T\Delta A\|+\|(\Delta A)^TA\|)}{\frac{\|\Delta A\|}{\|A\|}}$$

Предположим, что норма выходного пространства устойчива к транспонированию, также вспомним условие про субмультипликативность, тогда наше выражение примет вид

$$\frac{\|(A^TA)^{-1}\|\cdot(\|A^T\Delta A\|+\|(\Delta A)^TA\|)}{\frac{\|\Delta A\|}{\|A\|}} =$$
$$2\cdot\frac{\|(A^TA)^{-1}\|\cdot\|A^T\Delta A\|}{\frac{\|\Delta A\|}{\|A\|}} \leq$$
$$2\cdot\frac{\|(A^TA)^{-1}\|\cdot\|A^T\|\cdot\|\Delta A\|}{\frac{\|\Delta A\|}{\|A\|}} =$$
$$2\cdot\frac{\|(A^TA)^{-1}\|\cdot\|A^T\|}{\frac{1}{\|A\|}} =$$
$$2\cdot\|(A^TA)^{-1}\|\cdot\|A\|^2$$

Таким образом, при определенных требованиях к семейству норм мы получаем, что

$$\max_{\Delta A}\frac{\frac{\|(A^TA)^{-1}\cdot(A^T\Delta A+(\Delta A)^TA)\cdot(A^TA)^{-1}\|}{\|(A^TA)^{-1}\|}}{\frac{\|\Delta A\|}{\|A\|}} \leq$$
$$2\cdot\|(A^TA)^{-1}\|\cdot\|A\|^2$$

Вместо точного максимума мы можем использовать оценку сверху для всех значений.

Покажем влияние на данный результат близости матрицы к "вырожденной". Применим сингулярное разложение к матрице A, если она имеет размеры (n,m), где $n>m$

$$A = U_{(n,n)}\Sigma_{(n,m)}V^T_{(m,m)} \tag{46}$$

Σ — диагональная матрица где на диагонали находятся сингулярные числа матрицы, а U и V — унитарные матрицы, тогда A^TA равно
$V\Sigma^TU^TU\Sigma V^T = V\Sigma^TE\Sigma V^T$. Тогда:

$$(A^TA)^{-1} = (V\Sigma^TE\Sigma V^T)^{-1} \tag{47}$$

При этом $\Sigma^TE\Sigma$ — матрица (m, m), если матрица A имеет размер (n, m), где $n > m$. Так как матрица A имеет m сингулярных значений и является матрицей полного столбцового ранга — матрица $\Sigma^T\Sigma$ является обратимой. Продолжим раскрывать наше выражение

$$(A^TA)^{-1} = V(\Sigma^T\Sigma)^{-1}V^T \tag{48}$$

Мажорируем наше старое выражение, оно будет равно

$$2 \cdot \|(A^TA)^{-1}\| \cdot \|A\|^2 \leq 2 \cdot \|(\Sigma^T\Sigma)^{-1}\| \cdot \|V\|^2 \cdot \|A\|^2 \tag{49}$$

$\Sigma^T\Sigma$ — квадратная матрица на диагонали у которой находятся квадраты сингулярных значений матрицы, следовательно, обратная матрица к ней будет иметь вид $\text{diag}(\frac{1}{\sigma_1^2},\ \frac{1}{\sigma_2^2}\ ...\ \frac{1}{\sigma_m^2})$

$$2 \cdot \|(\Sigma^T\Sigma)^{-1}\| \cdot \|V\|^2 \cdot \|A\|^2 = 2 \cdot \|\,\text{diag}(\frac{1}{\sigma_1^2},\ \frac{1}{\sigma_2^2}\ ...\ \frac{1}{\sigma_m^2})\| \cdot \|V\|^2 \cdot \|A\|^2 \tag{50}$$

При этом сингулярные числа показывают близость матрицы к невырожденной, или же близость к матрице полного столбцового/строкового ранга, если матрица A имеет сильную зависимость между столбцами то $\min\{\sigma_1, \sigma_2...\} \to 0$, что ведет к увеличению числа обусловленности (сингулярные числа рассмотрю позже)

Второй подход к числу обусловленности

Мы можем сделать проще и рассмотреть число обусловленности матрицы A как линейного оператора. Рассмотрим уравнение:

$$Ax = b \tag{51}$$

Тогда решением в общем виде является

$$x = A^+b \tag{52}$$

Рассмотрим число обусловленности

$$\mu(A, b) = \max_{\Delta b} \frac{\frac{\|A^+(b+\Delta b) - A^+b\|}{\|A^+b\|}}{\frac{\|b+\Delta b-b\|}{\|b\|}} = \max_{\Delta b} \frac{\frac{\|A^+\Delta b\|}{\|A^+b\|}}{\frac{\|\Delta b\|}{\|b\|}} \tag{53}$$

Перенеся деление

$$\max_{\Delta b} \frac{\|A^+ \Delta b\| \cdot \|b\|}{\|\Delta b\| \cdot \|A^+ b\|} = \max_{\Delta b}(\frac{\|A^+ \Delta b\|}{\|\Delta b\|}) \cdot (\frac{\|b\|}{\|A^+ b\|}) \tag{54}$$

Аналогично прошлому случаю затребуем свойство субмультипликативности от нормы, тогда мы можем мажорировать первый множитель с помощью $\|A^+\|$. Следовательно:

$$\mu(A, b) = \|A^+\| \frac{\|b\|}{\|A^+ b\|} = \|A^+\| \frac{\|Ax + \epsilon\|}{\|x\|} \tag{55}$$

где ϵ - вектор отклонений

$$\|A^+\| \frac{\|Ax + \epsilon\|}{\|x\|} \leq \|A^+\| \frac{\|Ax\| + \|\epsilon\|}{\|x\|} \tag{56}$$

Таким образом мы получаем

$$\|A^+\| \frac{\|Ax\|}{\|x\|} + \|A^+\| \frac{\|\epsilon\|}{\|x\|} \leq \|A^+\| \frac{\|A\|x\|}{\|x\|} + \|A^+\| \frac{\|\epsilon\|}{\|x\|} \tag{57}$$

Финальное выражение выглядит так:

$$\|A^+\|\|A\| + \|A^+\| \frac{\|\epsilon\|}{\|x\|} \tag{58}$$

В случае если используется сингулярная норма: Сингулярная норма матрицы — ее максимальное сингулярное значение, а псевдообратная матрица имеет обратные сингулярные значения к исходной матрице (легко показать), таким образом:

$$\|A\|_{spec} = \sigma_{max}(A) \tag{59}$$

и

$$\|A^+\|_{spec} = \frac{1}{\sigma_{min}(A)} \tag{60}$$

Таким образом мы получаем:

$$\mu(A, b) = \frac{\sigma_{max}(A)}{\sigma_{min}(A)} + \|A^+\| \frac{\|\epsilon\|}{\|x\|} \tag{61}$$

Если пренебречь вторым слагаемым, то мы получим то, что описано в учебнике ШАДА: “Пожертвовав математической строгостью, мы можем считать, что число обусловленности матрицы X – это корень из отношения наибольшего и наименьшего из собственных чисел матрицы $X^T X$“.

Пояснение: Сингулярные значения матрицы A — это квадратные корни из собственных значений матрицы $A^T A$

Intuition проблемы числа обусловленности

Матрица A — линейное отображение из одного пространства в другое, если в новом пространстве есть вектора с сильной линейной связью, то оно становится более сжатым, т.е непохожие вектора в исходном пространстве могут стать сильно более похожими в новом пространстве за счет того, что оно сжато в размерах, таким образом если мы немного изменим целевой вектор (из нового пространства), то вектор который его образовал может сильно отличаться от того, что мы получили в прошлый раз, так как за счет сжатия они лежат рядом в новом пространстве, однако не в старом, таким образом: число обусловленности — лишь следствие того, что новое пространство сжато

Для наглядности можно рассмотреть, двумерный случай, рассмотрим матрицу перехода в новое пространство

$$\begin{pmatrix} 1 & 1 \\ 1 & 1+\epsilon \end{pmatrix}$$

где $\epsilon > 0$

Мы видим, что $\left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]$ отобразится в $\left[\begin{smallmatrix}1\\1\end{smallmatrix}\right]$, а $\left[\begin{smallmatrix}0\\1\end{smallmatrix}\right]$ в $\left[\begin{smallmatrix}1\\1+\epsilon\end{smallmatrix}\right]$. При маленьких ϵ матрица остается невырожденной, однако пространство "сжимается" делая $\left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]$ и $\left[\begin{smallmatrix}0\\1\end{smallmatrix}\right]$ все менее отличимыми в новом пространстве

3.3 Сложность точного решения

Алгоритмическая сложность точного решения для матрицы A размерами (N, D) задается как:

$$O(D^2N + D^3 + DN + D^2) \tag{62}$$

Из них мы тратим D^2N на перемножение матриц A^T и A, а D^3 — на обращение данной матрицы, DN — на умножение A^T на b и также тратим D^2 на умножение обратной матрицы на вектор A^Tb

Рассмотрим способы ускорения данных расчетов

- Во-первых легко заметить, что обращение матрицы имеет кубическую сложность, что плохо для задач с большим количеством фичей, рассмотрим итерационный алгоритм Шульца. Тогда $X_{k+1} = 2X_k - X_kAX_k = X_k(2E - AX_k)$, где A — исходная матрица, нужно заметить, что умножение A на X_k требует $O(D^3)$ итераций, однако на практике это не равносильно такому же количеству итераций, как и при обращении матрицы за счет возможности использования векторизации и распараллеливания расчетов, после нам требуется еще D^2 операций на вычитание и еще D^3 на перемножение итоговой матрицы с X_k. Также доп итерации тратятся на проверку сходимости, которая проводится через такую норму $\|X_kA - E\|$. Таким образом, при долгой сходимости данный способ может проигрывать точным алгоритмам по итеряциям

- Для симметричных матриц хорошо применимы методы Крылова, такие как Conjugate Gradient, MINRES, SYMMLQ, которые дают квадратичную сложность (слишком сложная тема, чтобы тут подробно раскрывать)
- Умножение матриц может быть распараллелено, например на GPU

3.4 Использование разложений для решения задачи

3.4.1 Построим QR разложение матрицы A

В данном разложении столбцы матрицы Q отнормированны и имеют единичную длину, i.e $Q^TQ = E$, а R — верхнетреугольная квадратная матрица, тогда:

$$w = (R^TQ^TQR)^{-1}R^TQ^Tb = (R^TR)^{-1}R^TQ^Tb = R^{-1}R^{T^{-1}}R^TQ^Tb \tag{63}$$

Таким образом итоговая формула будет иметь вид

$$w = R^{-1}Q^Tb \tag{64}$$

Для матрицы A размером (N, D) QR разложение будет иметь сложность сложность $O(ND^2 - \frac{D^3}{3})$, также мы затратим $O(D^3)$ на обращение R (мы можем сократить кол-во итераций за счет того, что R — верхнетреугольная матрица). Плюсами такого решения является то, что мы снижаем кол-во операций перемножения матриц, что повышает численную стабильность и снижает общее кол-во итераций. Суммарное кол-во итераций задается так:

$$O(ND^2 + D^3\frac{2}{3} + DN + D^2) \tag{65}$$

Мы получаем $O(DN)$ за счет перемножения Q^Tb и $O(D^2)$ за счет финального перемножения вектора R^{-1} и Q^Tb. В итоге сложность примерно такая же как и в прошлом случае, однако часть операций структурно отличается от него, что может позволить ускорить алгоритм.

3.4.2 Построим сингулярное разложение матрицы A

Про сингулярное разложение уже было сказано выше. Однако в данном случае мы будем использовать усеченное сингулярное разложение, где матрица с сингулярными значениями является квадратной, а U и V — ортогональны по столбцам (i.e $U^TU = E$ и $V^TV = E$) тогда $A = U\Sigma V^T$. Тогда:

$$w = (V\Sigma U^TU\Sigma V^T)^{-1}V\Sigma U^Tb = (V\Sigma\Sigma V^T)^{-1}V\Sigma U^Tb \tag{66}$$

Раскроем обратную матрицу

$$w = V^{T^{-1}}\Sigma^{-1}\Sigma^{-1}V^{-1}V\Sigma U^Tb = V^{T^{-1}}\Sigma^{-1}U^Tb = V\Sigma^{-1}U^Tb \tag{67}$$

Данное решение хорошо себя ведет в случае плохой обусловленности матрицы A

3.4.3 Заключение про точное решение

Можно увидеть, что точное решение является достаточно вычислительно сложным, как минимум из-за обращения матрицы, которое при большом количестве фичей будет вносить значительный вклад в итоговую сложность. Так же матрица A очень часто является плохо обусловленной в практических задачах.

3.5 Регуляризация

Зачем нам регуляризация? Рассмотрим влияние числа обусловленности на модель. Мы уже знаем, что число обусловленности влияет на стабильность модели. Однако

Во-первых, при добавлении $L2$ регуляризации в модель регрессии, она будет называться гребневая регрессия (Ridge regression). А в случае $L1$ — название будет лассо регрессия (Lasso regression).

Также заметим, что мы не будем регуляризовать вес, который соответствует константе, пока что опустим это, однако будем помнить об этом. Такая регуляризация просто не имеет смысла

3.5.1 Аналитическое решение задачи MSE с L2 регуляризацией

Нам нужно решить задачу

$$\min_{w}(X,y)\{\|Xw-y\|_2^2+\lambda\|w\|_2^2\} \tag{68}$$

Продифференцируем функцию $\|Xw-y\|_2+\lambda\|w\|_2^2$. Для этого мы представим нормы как скалярные произвдения

$$\langle Xw-y,Xw-y\rangle+\lambda\langle w,w\rangle \tag{69}$$

Мы уже знаем дифференциал первой части

$$2A^T(Aw-b)+[D_w(\langle w,w\rangle)] \tag{70}$$

$$[D_w(\langle w,w\rangle)]=2w \tag{71}$$

Следовательно нам нужно найти такой w, что

$$A^T(Aw-b)+\lambda w=0 \tag{72}$$

Это равносильно

$$A^TAw-A^Tb+\lambda w=0 \tag{73}$$

$$(A^TA+\lambda E)w-A^Tb=0 \tag{74}$$

Следовательно

$$(A^TA + \lambda E)w = A^Tb \tag{75}$$

Благодаря этому мы получаем, что

$$w = (A^TA + \lambda E)^{-1}A^Tb \tag{76}$$

Заметим, что для всех собственных значений $A^TA + \lambda E$ выполняется $\lambda_i(A^TA+\lambda E) \geq \lambda$. Так как для любого i верно $\lambda_i(A^TA) \geq 0$. Следовательно, данная матрица обратима

Заметим, что для $L1$ аналитического решения не существует, так как:

$$\begin{gathered}[D_w(\|w\|_1)] = [D_w(\langle w, \text{sign}(w)\rangle)] \\ = \langle [D_w(w)], \text{sign}(w)\rangle + \langle w, [D_w(\text{sign}(w))]\rangle \\ = \langle [D_w(w)], \text{sign}(w)\rangle + \langle w, 0\rangle \\ = \langle [D_w(w)], \text{sign}(w)\rangle = \langle 1, \text{sign}(w)\rangle\end{gathered}$$

3.5.2 Интуиция

Как мы видим на примере задачи с минимизацией MSE с L2 регуляризацией, ее добавление делает матрицу A^TA более "невырожденной", что повышает численную стабильность обращения матрицы и как следствие всего решения в целом. Предполагая, что целевая переменная декомпозируется на истинные значения и случайный шум, мы можем сказать, что наша модель получила лучшую обобщающую способность. Теперь измененииe случайного шума в таргете приведет к меньшему изменению вектора весов, так как снизилось число обусловленности. Следовательно вектор весов более стабилен и менее чувствителен к шуму в данных.

3.5.3 Разреживание весов

Давайте рассмотрим, почему для регуляризации используется также и L1 норма. Очень полезным свойством является, то что при ее использовании наиболее несущественные признаки склонны получать веса равные нулю. Таким образом L1 норма может использоваться для отбора фичей. Нестрого покажем почему это так. Нам нужно найти минимум функции

$$f(w) = \text{Loss}(w) + a \cdot \|w\|_1 \tag{77}$$

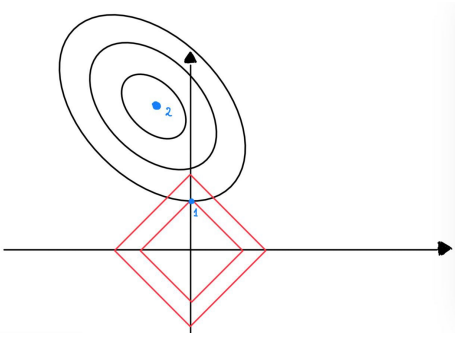
тут Loss — некая выпуклая функция потерь

Пусть мы знаем значение $\text{Loss}(w)$ в оптимуме $f(x)$ и это значение равно t. Мы возьмем линию уровня для t это $\{w \ : \ \text{Loss}(w) = t\}$, тогда нашей целью является решение такой системы

$$\begin{cases} A = \{w \ : \ \text{Loss}(w) = t\} \\ a \cdot \|w\|_1 = k \mid w \in A \\ k \to \min \end{cases}$$

Грубо говоря, нужно найти такую линию уровня нормы, которая пересекается с линией уровня функции потерь и при этом соответствует минимальному значению

Линии уровня L1 нормы это N-мерные октаэдры. Линия уровня функции потерь зафиксирована, меняя k мы можем сжимать и разжимать октаэдры. Мы будем снижать k настолько насколько возможно и скорее всего финальное пересечение линиий уровня функции потерь и нормы будет находиться на грани размерности, так как данные точки линии уровня L1 “выпирают“. Это можно показать визуально:



3.6 Линейная классификация

3.6.1 Разделяющий подход (SVM)

Во-первых, далее классы принадлежат множеству $\{-1,\ 1\}$, пока не указано иное. В данной ситуации мы пытемся найти разделяющую плоскость, которая будет делить пространство на два класса. Есть две ситуации:

- **Выборка линейно разделима**. В данном случае существует плоскость, которая идеально разделит исходные данные на два класса, так что с одной стороны плоскости будут только самплы одного класса, а с другой - другого
- **Выборка не линейно разделима**. В данном случае такой гиперплоскости не существует

Итоговое предсказание может быть задано как

$$y = \text{sign}(\langle w, X_i \rangle) \tag{78}$$

Рассмотрим как работает классификация в целом, пусть задана решающая плоскость:

$$\langle w, x \rangle + b = 0 \tag{79}$$

(В нашем случае все точно также, только 1 в конце вшита в X_i, а b в w). w перпендикулярна заданной плоскости. Это легко показать так: получим вектор лежащий на плоскости, для этого рассмотрим две точки x_1 и x_2 обе из которых лежат на плоскости i.e

$$\begin{cases} \langle w, x_1 \rangle + b = 0 \\ \langle w, x_2 \rangle + b = 0 \end{cases}$$

Рассмотрим вектор $x^* = x_1 - x_2$, заметим, что данный вектор лежит на плоскости, мы можем проверить его перпендикулярность w.

$$\langle w, x^* \rangle = \langle w, x_1 - x_2 \rangle = \langle w, x_1 \rangle - \langle w, x_2 \rangle = -b - (-b) = 0 \tag{80}$$

Заметим, что w^* соответствует любому вектору лежащему на плоскости, следовательно w параллельна всей плоскости. Рассмотрим геометрический смысл $\langle w, x \rangle + b$, где x — произвольный вектор, в нашей регрессии это просто $\langle w, x \rangle$ для удобства. Мы знаем, что w перепендикулярен плоскости т.е мы можем провести из произвольной точки x перпендикулрную к плоскости прямую, используя w. $x_{new} = x + aw$, где a — произвольный коэффициент. При этом для x_{new} выполняется $\langle w, x_{new} \rangle + b = 0$. Найдем коэффициент a.

$$\langle w, x + aw \rangle + b = 0 \tag{81}$$

Это равносильно

$$\langle w, x \rangle + a\langle w, w \rangle + b = 0 \tag{82}$$

Следовательно

$$a = -\frac{\langle w, x \rangle + b}{\langle w, w \rangle} \tag{83}$$

Нас интересует расстояние от x, до этой новой полученной точки, которое равно

$$|a| \cdot \|w\| = |\frac{\langle w, x \rangle + b}{\|w\|^2}| \cdot \|w\| \tag{84}$$

Следовательно искомое расстояние раввно

$$\frac{|\langle w, x \rangle + b|}{\|w\|} \tag{85}$$

Получается мы имеем дело с числителем расстояния. Заметим, что максимизация расстояния равносильна

$$\begin{cases} |\langle w, x \rangle + b| \to \max \\ \|w\| \to \min \end{cases}$$

Мы воспользуемся этим далее

Продолжим. Мы хотим, чтобы

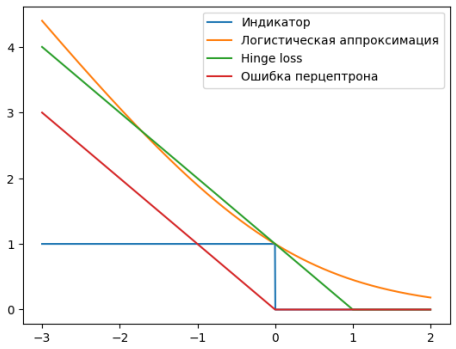
$$\sum_i \mathrm{I}[y_i \neq \mathrm{sign}(\langle w, X_i \rangle)] \to \min \tag{86}$$

Или же

$$\sum_i \mathrm{I}[y_i \cdot \langle w, X_i \rangle < 0] \to \min \tag{87}$$

Обозначии $\mathrm{M} = y_i \cdot \langle w, X_i \rangle$, это называется отступом (margin) классификатора. Можно легко заметить, что чем больше margin, тем лучше решается задача классификации. Отступ положителен, когда $\mathrm{I}[y_i = \mathrm{sign}(\langle w, X_i \rangle)]$, то есть класс угадан верно, при этом чем больше отступ, тем больше расстояние от сампла до разделяющей гиперплоскости, то есть «уверенность классификатора». Таким же образом, если отступ отрицателен, то класс угадан неверно, при этом, чем больше по модулю отступ, тем более сокрушительно ошибается классификатор. Нам нужно аппроксимировать индикатор непрерывной функцией

Варианты аппроксимации:



Распишем все представленные варианты

- **Логистическая**. Задается формулой $\log(1 + e^{-M})$, в примере делилась на $\log(2)$, чтобы проходить через точку $(0, 1)$, но это не необходимо для задачи оптимизации
- **Ошибка перцептрона**. Задается как $\max(0, -M)$
- **Hinge loss**. Задается как $\max(0, 1 - M)$

Можно заметить, что последние два лосса очень похожи. Их разница лишь в том, что Hinge loss продолжает штрафовать даже правильно классифицированные точки, если они приближаются к разделяющей плоскости

слишком близко, а именно если они находятся на расстоянии $\frac{1}{\|w\|}$ и ближе от разделяющей плоскости. Итоговый функционал потерь задается как:

$$\sum_i \text{loss}(y_i, X_i) + \lambda \|w\|^2 \tag{88}$$

На выходе мы получаем строго выпуклую функцию, что гарантирует уникальность найденного минимума. Это достигается за счет того, что норма — строго выпуклая функция, а все три перечисленных выше аппроксимации функции потерь являются выпуклыми, а сумма выпуклой и строго выпуклой функции дает строго выпуклую функцию.

Как мы видим, данная задача оптимизации эквивалентна задаче

$$\begin{cases} |\langle w, x \rangle + b| \to \max \quad \text{(При определенных условиях)} \\ \|w\| \to \min \end{cases}$$

Именно поэтому мы можем использовать $\langle w, x \rangle$ вместо расстояния до решающей плоскости при оптимизации. Вот цитата из учебника ШАДа об SVM:

> Итоговое положение плоскости задаётся всего несколькими обучающими примерами. Это ближайшие к плоскости правильно классифицированные объекты, которые называют опорными векторами или support vectors. Весь метод, соответственно, зовётся методом опорных векторов, или support vector machine, или сокращённо SVM. Начиная с шестидесятых годов это был сильнейший из известных методов машинного обучения. В девяностые его сменили методы, основанные на деревьях решений. Почему же SVM был столь популярен? Из-за небольшого количества параметров и доказуемой оптимальности. Сейчас для нас нормально выбирать специальный алгоритм под задачу и подбирать оптимальные гиперпараметры для этого алгоритма перебором, а когда-то трава была зеленее, а компьютеры медленнее, и такой роскоши у людей не было. Поэтому им нужны были модели, которые гарантированно неплохо работали бы в любой ситуации. Такой моделью и был SVM. Другие замечательные свойства SVM: существование уникального решения и доказуемо минимальная склонность к переобучению среди всех популярных классов линейных классификаторов. Кроме того, несложная модификация алгоритма, ядровый SVM, позволяет проводить нелинейные разделяющие поверхности.

3.6.2 Kernel-ы в SVM

Интуиция

Пусть в нашем пространстве данные не являются линейно разделимыми или они трудноразделимы. Мы хотим отобразить наши данные в новое

пространство, где будем искать разделяющую плоскость. Мы предполагаем, что в этом новом пространстве использование разделяющей плоскости будет более эффективно, чем в исходном. Однако напрямую отображать данные в необходимое пространство вычислительно трудно, так же как и считать скалярное произведение между ними в нём. Однако мы можем вычислить скалярное произвдение в целевом пространстве не переходя в него,это так называемый kernel trick. Вместо использования обычного скалярного произведения, мы используем kernel, который возвращает скалярное произведение репрезентаций двух переданных векторов в новом пространстве. Таким образом мы можем улавливать нелинейные зависимости работая при этом в исходном пространстве.

(Cooming soon)

3.6.3 Вероятностный подход

4 Эмбеддеры в Word2Vec

Цель эмбедера - отобразить некий нечисловой объект в многомерное числовое пространство, так чтобы похожие объекты в старом пространстве получали как можно более сонаправленные вектора в новом.

Существует два основных подхода к созданию эмбедеров в контексте Word2Vec, это

- CBOW — предсказывает слово по контексту. Она работает быстрее, но, в силу своей логики, хуже улавливает семантику редких слов.
- Skip-Gram — предсказывает контекст по слову. Из-за необходимости обрабатывать несколько выходов для каждого слова эта архитектура работает медленнее, зато лучше захватывает семантику редких слов.

Примечание Обе архитектуры основаны на концепции Bag of Words, которая утверждает, что вся информация содержится в совокупности слов без учёта их порядка. Это справедливо только для небольших окон (эмпирическое наблюдение). Проявление данной концепции можно будет наблюдать в используемых функциях потерь

Обе архитектуры изначально использовали softmax выходной слой для предсказния вероятности по всем словам и log-loss функцию потерь, предсказывая вероятности по всем словам словаря, однако более эффективным способом является Negative Sampling, который является надстройкой над оргинальной моделью

4.1 CBOW

Пусть в словаре n слов, каждое из которых имеет векторное представление. Рассмотрим множество окон K, полученных из текстового корпуса.

Каждое окно соответствует целевому слову w_{target}. Для каждого $k \in K$ мы стремимся максимизировать:

$$P(w_{target} \mid w_{context}) \to \max$$

Здесь $w_{context}$ определяется как $\text{mean}_{w \in k}(w)$. Этот вектор аккумулирует информацию о контексте.

Вероятность $P(w_m \mid w_{context})$ наблюдать слово w_m в контексте слов из $w_{context}$ вычисляется как $\sigma(w_{target} \cdot w_{context})$.

Однако, мы не можем использовать такую функцию потерь, так как модель будет просто пытаться сделать векторы слова и контекста сонаправленными для каждого окна. К функции максимизации добавляются потери на случайных словах, не входящих в текущее окно контекста, это и есть Negative Sampling. Мы минимизируем вероятность для таких слов, что приводит к итоговой формуле:

$$P(w_{target} \mid w_{context}) - \prod_{w_{rand} \notin k \mid w_{rand} \neq w_{target}} P(w_{rand} \mid w_{context}) \to \max$$

Это грубая аппроксимация log-loss функции, учитывающей все слова словаря в оригинальной реализации.

С точки зрения оптимизации, формула эквивалентна:

$$\log(P(w_{target} \mid w_{context})) + \sum_{w_{rand} \notin k \mid w_{rand} \neq w_{target}} \log(P(w_{rand} \mid -w_{context})) \to \max$$

CBOW использует две матрицы: одна кодирует контекст, другая — целевое слово и негативные примеры. Это реализует принцип разделения ответственности: матрица контекста учится понимать контекст слова, а матрица целевых слов учится отражать семантику. Обоснование использования двух матриц можно найти в статьях по теме:

> Предполагается, что слова и контексты принадлежат разным словарям. Например, вектор слова "собака"отличается от вектора контекста "собака". Это связано с тем, что слова редко встречаются в своих собственных контекстах, следовательно модель должна присваивать низкую вероятность $P(\text{собака} \mid \text{собака})$, что требует малой величины $w \cdot w$. В оптимуме $\|w\| \to 0$, что неправильно.

Эта проблема скорее относится к оригинальной реализации. Negative Sampling частично решает данную проблему, так как минимизация вероятности выплняется не по всем словам вне окна, а по случайным словам, не входящим в окно, также нужно учитывать, что при семплировании мы не берем целевое слово, что положительно сказывается на решении данной проблемы

4.2 Skip-Gram

Логика этой модели похожа на CBOW, но с несколькими ключевыми отличиями. Мы работаем с вероятностями $P(w_{context} \mid w_{target})$, где $w_{context}$ — множество векторов контекста. С добавлением негативного семплирования задача оптимизации принимает вид:

$$\sum_{w_i \in k} \log(P(w_i \mid w_{target})) + \sum_{w_{rand} \notin k | w_{rand} \neq w_{target}} \log(P(w_{rand} \mid -w_{target})) \to \max$$

В Skip-Gram используется одна матрица для всех векторов. Это связано с тем, что эмбеддинг должен аккумулировать информацию о контексте: предсказание вероятности контекста следует непосредственно из вектора слова. Проблема описанная выше, которая обосновывает использование двух матриц, решается с помощью Negative Sampling, который не позволяет минимизировать $\|w\|$

4.3 Выравнивание эмбеддингов

Допустим мы используем эмбеддеры для перевода слов с одного языка на другой. У нас могут быть размечены слова переводы для различных слов. Мы хотим отобразить наши вектора в единое векторное пространство, чтобы семантически похожие слова внутри разных языков имели как можно более сонаправленные вектора в этом новом пространстве, так как это работает внутри одного языка. Есть два основных способа:

- MUSE (Multilingual Unsupervised and Supervised Embeddings)
- VecMap

Пусть у нас есть два словаря для двух разных языков W_c — словарь нашего языка, а W_f — словарь иностранных слов. Формально можно считать, что существует два линейных пространства: L_c, L_f, таких, что $\forall w_c \in W_c$: $w_c \in L_c$ и $\forall w_f \in W_f : w_f \in L_f$. Т.е каждый словарь опредлен над L_c и L_f соответственно. Цель выравнивания можно определить как $L_c \to L_f$.

Нам нужно привести словари к единому виду, так как они обучены на разных корпусах. Можно использовать общее подмножество слов. Например через явный перевод, или нахождение общих слов, таких как заимствованные. Мы создаем перекресный словарь по ним. Некоторые методы концентрируются на распределении векторов, не используя перекрестный словаь.

4.3.1 MUSE

Вот статья: https://arxiv.org/abs/1710.04087

Исходя из названия поддерживает как supervised так и unsupervised режимы

Unsupervised

Мы используем GAN для того, чтобы получить вектора с таким же распределением как и в целевом словаре. Однако в данном случае в качестве генератора выступает обыкновенная ортогональная матрица, которая учится обманывать дискриминатор. При этом пары "Слово-Перевод" семплируются случайно, без смыслового сопоставления, так как наша цель — именно имитация распределения. Авторы оригинальной статьи использовали 50 тысяч самых часто встречающихся слов для обучения генератора для словарей размерами по 200 тысяч. Однако использование всех слов не приводило к значимому ухудшению результатов. Слова выбирались равномерно, так как семплирование пропорционально частоте не дало улучшений. После того как мы обучили GAN, мы приступаем к Procrustes analysis (анализ Прокруста), предварительно применив полученную матрицу ко всем эмбэддингам.

Рассмотрим задачу более подробно:
Наша цель — найти такую матрицу W, что

$$\begin{cases} \|WX - Y\|_F^2 \to \min \\ WW^T = E \\ W^TW = E \end{cases}$$

Аналитически задача решается так:
Нам дано $\text{trace}((WX - Y)^T(WX - Y))$, расскроем транспонирование и получаем $\text{trace}((X^TW^T - Y^T)(WX - Y)) = \text{trace}(X^TW^TWX - X^TW^TY - Y^TWX + Y^TY)$. Так как $W^TW = E$. Расскроем выражение $\text{trace}(X^TX - X^TW^TY - Y^TWX + Y^TY)$. Так как оптимизация выполняется по W, это равносильно $\text{trace}(-X^TW^TY - Y^TWX)$. Следовательно

$$\begin{cases} \text{trace}(-X^TW^TY - Y^TWX) \to \min \\ WW^T = E \\ W^TW = E \end{cases}$$

Мы можем преобразовать $\text{trace}(-X^TW^TY - Y^TWX) \to \min$ в $\text{trace}(X^TW^TY + Y^TWX) \to \max$. Так как $\text{trace}(A) = \text{trace}(A^T)$. Вся задача оптимизации сводится к $\text{trace}(Y^TWX) \to \max$. Это можно свести к $\text{trace}(WXY^T) \to \max$. Пусть $M = XY^T$. Тогда нам нужно оптмизировать $\text{trace}(WM)$. По теореме Eckart–Young–Mirsky, максимум будет достигаться в $W = UV^T$

Вот как происходит обучение в рамках Procrustes analysis: Во-первых, мы выбираем наиболее частые слова для данного этапа. Во-вторых, мы образуем синтетический словарь, на основе Cross-Domain Similarity Local Scaling (CSLS). Обсудим это подробнее:

CSLS

Мы не можем просто использовать NN, так как с таким подходом часто возникает ситуация, что для пары (x, y), где y — ближайший к x вектор, далеко не всегда выполняется, что x – ближайший к y вектор. Авторы ссылаются на статью, которая показывает, что в высокоразмерных пространствах это приводит к образованию "хабов" и "антихабов" — векторов, которые являются соседями сразу для многих или не практически не являются соседями ни для кого, соответственно. Для решения данной проблемы мы бдуем "штрафовать" точки за то, что они являются хабами. Пусть мы будем использовать для этого K соседей. Определим пару функций: $N_{c'}(x)$ — функция, которая возвращает первые K соседей вектора x из словаря W'_c, который образован путем применения генератора ко всем эмбеддингам из W_c. Аналогичная функция для W_f — $N_f(x)$. Определим еще две функции:

$$r_{c'}(x) = \frac{1}{K} \sum_{y \in N_{c'}(x)} cos(x, y) \tag{89}$$

где cos — косинусная метрика. Также введем аналогчиную функцию для W_f — $r_{f'}(x)$.

Введем финальную функцию:

$$\text{CSLS}(x, y) = 2 \cdot cos(x, y) - r_{c'}(y) - r_f(x) \tag{90}$$

где $x \in W_{c'}$ и $y \in W_f$

Получается для каждого языка мы ищем соседей в противоположном. Такая фнукция сходства штрафует векторы, которые лежат в плотных областях. И по наблюдениям авторов метода использование CSLS дает значительное улучшение показателей.

Мы создаем синтетический словарь, ищя MNN (Mutal Nearest Neigh.) по CSLS на каждой итерации. Далее, для этого словаря мы ищем матрицу W по алгоритму выше и применяем ко всем эмбэддингам. Процесс повторяется итерационно. Согласно эмпирическим данным, обычно хватает от 5 до 10 итераций. Итоговая функция для трансформации вектора в новое пространство имеет вид

$$\prod_{i=0}^{n-1} W_{n-i} \cdot W_G \cdot x \tag{91}$$

где W_G - матрица генератора

Supervised

Если у нас есть перекрестный словарь, то мы переходим сразу к шагу Procrustes анализа. Однако нужно учитывать, что, несмотря на название, метод разрабатывался для unsupervised режима работы

4.3.2 VecMap

В первую очередь мы нормализуем эмбэддинги, а также центрируем каждую координату. После этого мы снова применяем нормализацию. Постром матрицы $M_X = X^T X$ и $M_Y = Y^T Y$ (тут предполагается, что X и Y — словари в которые по столбцам записаны эмбэддинги). Если пространства эмбэддингов полностью изометричны, то существует такая перестановка строк или столбцов что матрицы M_X и M_Y будут равны. Тут предполагается, что размеры словарей по которым будет производиться выравнивание равны, также как и размерности эмбеддингов. На практике изометрия выполняется приближённо, а поиск перестановок NP-полная задача.

Возьмем и независимо отсортируем каждую строку в M_X и в M_Y, получив матрицы sorted(M_X) и sorted(M_Y). При условии идеальной изометрии получившиеся матрицы будут равны с точностью до перестановки строк. Исходя из предположения, что изометрия приближенно выполняется мы можем искать ближайших соседей для строки из одной матрицы среди строк другой матрицы и считать, что найденный ближайший сосед является репрезентацией исходной строки в другом пространстве.

Важное замечание: если мы используем SVD для матриц эбэддингов (т.е $X = U\Sigma V^T$) то получим, что $M_X = V\Sigma U^T U\Sigma V^T = V\Sigma^2 V^T$. Авторы оригинальной статьи заметили, что использование матрицы $\sqrt{M_X} = V\Sigma V^T$ более эффективно (в оригинале использовалась матрица $U\Sigma U^T$, так как эмбеддинги были записаны по строкам).

Процесс обучения мы начинаем с формирования небольшого словаря, например 25 слов. Для его формирования мы используем алгоритм описанный выше (в оригинальной статье использовались 4000 наиболее частых слов для создания начального словаря). Далее по известному алгоритму из Procrustes анализа находим ортгональную матрицу W, используя только эмбеддинги из словаря, которую после применяем ко всем эмбеддингам. Далее мы пытаемся добавить новое слово в словарь, для этого мы используем алгоритм из начала. Для этого мы ищем ближайшего соседа для каждой строки из корня sorted версии матрицы схожести одного языка в такой же матрице другого языка, пару наиболее похожих слов мы добавляем в словарь. После снова переходим к поиску матрицы W, итеративно повторяя весь процесс с новым словарем.

Однако такой сырой алгоритм нуждается в дополнениях. Вот некоторые из них:

- Dropout. Для каждого вектора при поиске его соседей, каждый сосед с вероятностью $1 - p$ игнорируется, т.е мы зануляем веса связи с вероятностью $1 - p$. В предложенной авторами метода реализации мы начинаем с $p = 0.1$ (сильный стохастический шум). Если целевая функция не улучшается 50 итераций подряд, увеличиваем p в 2 раза (аналог simulated annealing).

- Использование наиболее частых слов. Для обучения мы используем лишь первые 20000 наиболее частых слов в каждом языке.

- MNN + CSLS. Как и в MUSE, для поиска нового слова мы используем MNN + CSLS, вместо обычного NN.

5 Контекстуальные эмбэддеры

Нужно заметить, что сейчас статические эмбэддеры не используются, так как на рынке доминируют контекстуальные эмбэддеры, т.е вектор слова меняется в зависимости от контекста в котором оно находится

Примеры контекстуальных эмбэддеров:

- **ELMo**. Использует двунаправленные LSTM
- **BERT**. Использует трансформеры и учится на задачах маскирования. Маскирование похоже на CBOW, часть слов в окне маскируется и заменяется на токен маски. После этого модель пытается предсказать на основе контекста, что за слово скрыто за токеном маски.

(Информация по ним будет внесена позже)

6 ONNX

ONNX (Open Neural Network Exchange) - формат для представления моделей из разных фреймворков в унифицированном виде. Это позволяет передавать модели между ними, а также ONNX удобен для инференса