YSDA ML HW #6

Тихон Волжин

22 октября 2023 г.

Задача 3

Сначала предположим, что объекты центрированные, и речь идет о проекции на несмещенное подпространство. Пусть есть фиксированный ОНБ базис в $\mathbb{R}^D$, и рассмотрим произвольное подпространство размерности d. Параметризуем множество этих подпространств ортонормированными базисами в координатах исходного, при этом подпространство определяется базисом с точностью до действия группы $O(d)$. Запишем минимизируемую сумму в матричном виде: представим ее как сумму квадратов всех элементов матрицы, являющейся поэлементной разницей между $X \in M_{N\times D}(\mathbb{R})$ и проекцией на d-мерное пространство $Y \in M_{N\times D}(\mathbb{R})$. То есть в пространстве матриц $N \times D$ нужно взять норму матрицы в виде $Tr(H \times H^T) = \sum_{i,j} h_{ij}^2$:

$$\sum_{i=1}^{n}(X_i - Y_i)^2 = ||X - Y|| = Tr((X - Y)(X - Y)^T)$$

Имея d ортонормированных векторов $E = (e_1, \ldots, e_d) \in M_{D\times d}$ можно записать проекцию на пространство $Y_{proj} = X\times E$, а затем записать в исходных координатах через "обратное"преобразование $Y = XEE^T$ (то есть $E \times E^T$ - проектор на d-мерное пространство заданное ОНБ $e_1, \ldots, e_d$ в исходных координатах). То есть минимизируемый функционал имеет вид:

$$F_X(E) = Tr((X - XEE^T)(X - XEE^T)^T) \longrightarrow min$$

Из условия ортонормированности базиса $e_1, \ldots, e_d$ можно записать (I - единичная матрица):

$$E^T E = I \in M_{d\times d}$$

Итоговый Лагранжиан может быть записан в виде:

$$L_X(E) = Tr((X - XEE^T)(X - XEE^T)^T) + \sum_{i,j}\lambda_{ij}(e_i e_j - \delta_{ij}) \longrightarrow min$$

Можно ввести матрицу $\Lambda = \{\lambda_{ij}\} \in M_{d\times d}$ - симметричная матрица коэффициентов Лагранжа, и записать условный вклад в трейсово-матричном виде:

$$L_X(E) = Tr((X - XEE^T)(X - XEE^T)^T) + Tr(\Lambda(E^T E - I)) \longrightarrow min$$

$$E^T E = I$$

Распишем Лагранжиан в более удобном виде:

$$Tr(XX^T - 2XEE^T X^T + XEE^T EE^T X^T) = Tr(XX^T - 2XEE^T X^T + XEI_{d\times d}E^T X^T) =$$

$$= Tr(XX^T - XEE^T X^T) = Tr(XX^T - XEE^T X^T) = Tr(X(I_{D\times D} - EE^T)X^T) =$$

$$= Tr(X^T X(I_{D\times D} - EE^T)$$

Наконец, возьмем производную Лагранжиана (везде используем линейность при протаскивании производной через линейные операции - линейность Tr, линейность при перемножении матриц итп):

$$[D_E F_X(E)](H) = -Tr(X^T X[D_E E](H)E^T) - Tr(X^T XE[D_E E^T](H)) = -Tr(X^T XHE^T) - Tr(X^T XEH^T) =$$

$$= -Tr(E^T X^T X H) - Tr(X^T X E H^T) = -Tr(H^T X^T X E) - Tr(X^T X E H^T) = -Tr(X^T X E H^T) - Tr(X^T X E H^T) =$$
$$= -Tr((X^T X E + X^T X E) H^T) = -2Tr(X^T X E H^T) = -2\langle X^T X E, H \rangle$$

Отсюда градиент:

$$\nabla_E F_X(E) = -2X^T X E$$

Аналогично можно получить выражение для градиента условной части Лагранжиана:

$$Tr(\Lambda([D_E E^T](H)E))Tr(\Lambda(E^T[D_E E](H))) = Tr(\Lambda H^T E) + Tr(\Lambda E^T H) = 2Tr(E\Lambda H^T) = 2\langle E\Lambda, H \rangle$$

Итоговый градиент:

$$\nabla_E \Lambda_E = 2E\Lambda$$

В итоге необходимое условие экстремума:

$$\nabla_E L_X(E) = -2X^T X E + 2E\Lambda = 0$$

$$X^T X E = E\Lambda$$

Так как Λ - симметричная матрица, то у нее есть диагональный вид $\Lambda_{diag} = diag(\lambda'_1, \ldots, \lambda'_d)$, к которому можно прийти ортогональной заменой O:

$$X^T X E = EO\Lambda_{diag} O^{-1}$$

$$X^T X (EO) = (EO)\Lambda_{diag}$$

Замена $E \longrightarrow EO$ оставляет то же подпространство размерности d, но с другим базисом, тк ортогональная замена оставляет базис ортогональным и легко видно, что проекция не изменяется (тк подпространство то же):

$$XEE^T = XEOO^{-1}E^T = \{O^{-1} = O^T\} = X(EO)O^T E^T = X(EO)(EO)^T$$

Поэтому сразу запишем измененный Лагранжиан с заменой $E' = EO$:

$$L_X(E') = Tr(X^T X(I_{D\times D} - E'E'^T)) + Tr(\Lambda_{diag}(E'^T E' - I)) \longrightarrow min$$
$$E'^T E' = I$$

Полученное необходимое условие $\qquad X^T X E' = E'\Lambda_{diag}$

Из полученного необходимого условия видно, что векторы из матрицы $E' = (e'_1, \ldots, e'_d)$ должны быть собственными векторами матрицы $X^T X$ - для значения e'_i собственное значение:

$$\Lambda_{diag}(i, i) = \lambda'_i$$

Это мы получаем непосредственно решая полученные выше уравнения Лагранжа. Резюмируя: мы получили вид Лагранжиана, в котором матрица Λ диагональна, воспользовавшись неоднозначностью выбора ортонормированного базиса в исследованном подпространстве размерности d. То есть условия экстремума выполняются для всех матриц E, которые ортогональной заменой базиса O(в проекционном подпространстве) можно привести к матрицей, состоящей из координат d собственных векторов $X^T X$.В таком виде мы сразу получаем, что для экстремума необходимо, чтобы вектора были собственными векторами матрицы $X^T X$ и элементы матрицы условных коэффициентов Лагранжа Λ_{diag} в этом базисе - соответствующие собственные значения. На эту задачу нужно смотреть как на нахождение минимума функционала на грассманиане $\mathbf{Gr}(\mathbf{d}, \mathbf{D})$, который является компактным многообразием без края. Функционал на нем непрерывно дифференцируем, и точки экстремума - подпространства натянутые на d собственных векторов $X^T X$, их конечное число и из компактности следует, что глобальный минимум достигается на одном из таких подпространств (если на компактном без края многообразии есть дифференцируемый функционал, то его глобальный минимум или максимум должен достигаться в одной из точек экстремума, это я к тому, что каких-то минимумов/максимумов на бесконечности не будет, как

это было бы для какого-нибудь функционала на некомпактном $\mathbb{R}^n$; и необходимости анализировать гессиан нет, достаточно рассмотреть значения функционала в экстремальных точках и выбрать наименьшее). Смотря на функционал в этих точках:

$$F_X(E) = Tr(X^T X(I_{D\times D} - EE^T)) = Tr(X^T X) - Tr(X^T XEE^T) = Tr(X^T X) - Tr(X^T XEE^T) =$$

$$= Tr(X^T X) - Tr(E\Lambda E^T) = Tr(X^T X) - Tr(\Lambda E^T E) = \{E^T E = I\} = Tr(X^T X) - Tr(\Lambda) =$$

$$= Tr(X^T X) - Tr(O\Lambda O^{-1}) = Tr(X^T X) - Tr(\Lambda_{diag}) = Tr(X^T X) - \sum_{i=1}^{d} \lambda'_i$$

Где, как мы уже выяснили, $\sum_{i=1}^{d} \lambda'_i$ - сумма из какой-то выборки собственных значений $X^T X$. Переобозначив выборку в терминах собственных значений $X^T X$:

$$\lambda_1^{X^T X}, \ldots, \lambda_D^{X^T X}$$

Тогда итоговый функционал имеет вид:

$$F_X(E) = Tr(X^T X) - \sum_{i=1}^{d} \lambda_{n_i}^{X^T X} = \sum_{i=1}^{D} \lambda_i^{X^T X} - \sum_{i=1}^{d} \lambda_{n_i}^{X^T X} = \sum_{i=1}^{D-d} \lambda_{d_i}^{X^T X}$$

где набор $\{n_i\}$ - выборка из d элементов из $\{1, \ldots, n\}$, а $\{d_i\}$ - ее дополнение в $\{1, \ldots, n\}$ из D - d элементов. Кароче говоря, уже очевидно, что нужно выбирать подпространство из первых d главных компонент, тк функционал - это сумма значений оставшихся собственных чисел. Значит выбрав первые d главных компонент с наибольшими собственными значениями (и кстати стоит напомнить что все они неотрицательны, тк $X^T X$ - положительно полуопределенная матрица) сумма оставшихся будет наименьшая.
В конце стоит отметить, что изначально предполагалось выше что выборка центрирована. В случае, если это не так, то легко увидеть, что функционал от этого не изменяется, ведь если $X_i = X'_i + \bar{X}$ (X'_i - центрированный объект, $\bar{X}$ - признаковое среднее), то его проекция на главные компоненты описывается также в терминах сдвига $Y_i = Y'_i + \bar{X}$ (проекция на смещенное подпространство):

$$\sum_{i=1}^{n} (X_i - Y_i)^2 = \sum_{i=1}^{n} (X'_i - \bar{X} - (Y'_i - \bar{X}))^2 = \sum_{i=1}^{n} (X'_i - \bar{X} - Y'_i + \bar{X})^2 = \sum_{i=1}^{n} (X'_i - Y'_i)^2$$

Поэтому алгоритм действий не меняется, просто нужно перейти к центрированным объектам. чтд

P.S. для большей строгости, следует кое-что упомянуть... Наш функционал определен на пространстве матриц E, то есть в $R^{D\times d}$.На нем он очевидно непрерывен и дифференцируем, тк состоит из линейных и прочих дифференцируемых операций (перемножение компонент итп) и их композиций. Затем можно сузить подмножество матриц, удовлетворяющих условию $E^T E = I$, и при ограничении на подмножество - дифференцируемость остается. Затем, главный момент в том, мы хотим взять фактор по действию группы ортогональных матриц (тк функционал принимает на нем одинаковые значения на орбите каждого элемента E под действием $O(d)$: $F_X(E) = F_X(E \times O), O \in O(d)$. После факторизации мы уже переходим к функционалу на компактном многообразии со всеми хорошими свойствами (т е к грассманиану), где не нужно рассматривать и анализировать экстремумы, доказывать существования минимумов/максимумов, искать гессиан и прочее, тут автоматом получаем, что минимум где-то в конечном числе экстремумов полученных из уравнений Лагранжа. Для строгого доказательства, не хватало, собственно, доказательства дифференцируемости функционала и того, что и того, что экстремумам в исходном пространстве соответствуют экстремумы на грассманиане. Второе +- очевидно, тк в одну сторону градиент перпендикулярен линии уровня, а орбита O(d) очевидно содержится в линии уровня, тк значение функционала на ней постоянно. Поэтому направление роста функционала не определяется орбитой, и при факторизации если градиент был нулевым, то после также останется нулем, а если есть нулевой градиент на грассманиане, значит и на исходном многообразии он был нулем, то есть существует взаимно однозначное соответствие экстремумов (ну можно представить градиент как вектор, у которого в определенных локальных координатах в соответствии

с орбитой (то есть d(d-1)/2 компонент направлены вдоль орбиты), для компонент орбиты стоят нули, а для других компонент координат возможно не нули, и при факторизации от этого вектора отсекаются d(d-1)/2 нулей (размерность O(d)), а при обратном переходе добавляются, и в общем если отсекать и добавлять нули к нулевому вектору повышая или понижая его размерность, то будет нулевой вектор просто другой размерности, надеюсь, что достаточно строго). И главный вопрос, касающийся дифференцируемости самого функционала на грассмановом многообразии: если проекция $p : X \longrightarrow Y$ исходного многообразия на фактор многообразие (в нашем случае многообразия, задаваемого условием $X = \{E : EE^T = I, E \in \mathbb{R}^{D\times d}\}$, на $Y = \mathbf{Gr}(\mathbf{d}, \mathbf{D})$), дифференцируема, тогда любой дифференцируемый функционал на X дифференцируем на Y. Дифференцируемость проекции очевидна, так как это фактор многообразия по подмногообразиям, то есть в координатах с соответствием многообразию орбит можно отсечь $d(d-1)/2$ компонент, что очевидно дифференцируемое отображение, ну или с помощью . Теорема взята отсюда отсюда (утверждение на стр 19 пар 3.1, группы Ли тут не при чем, хоть и фактор по группе Ли, важно что проекция дифференцируема и все это гладкие многообразия). А значит

P.P.S. Хотя я так подумал, что про грассманиан можно было не писать, тк $M = \{E : EE^T = I, E \in \mathbb{R}^{D\times d}\}$ - итак компактное (замкнутое, как прообраз замкнутого множества для непрерывного отображения EE^T и ограниченное, тк все компоненты матрицы можно оценить сверху по модулю как 1, чтобы могло выполнится условие ортонормированности, а значит можно заключить множество в ограниченный шар конечного размера например - $\sqrt(D)$, а в евклидовом пространстве замкнутость и огранниченность влечет компактность) гладкое многообразие, то есть можно было не рассматривать фактор, чтобы доказать, что минимум находтся именно среди этих экстремумов, но ладно в принципе от интерпритации грассманиана хуже не стало, зато теперь мы рассматриваем конечное число C_D^d изолированных точек на грассманиане, а не подмногообразия в M.