## МСИС | Методы понижения размерности |



Ещё одно решение проблемы мультиколлинеарности состоит в том, чтобы подвергнуть исход. признаки некому линейному преобразованию так, чтобы новые призн. были линейно не зависимы (и даже ортогональны!), а их число бы уменьшилось.

### Метод глав. компонент (PCA)

Пусть есть $k$ исход. числовых признаков $X_1 \dots X_k$ размер. $n$, расcм. матрицу $X = (x_{ij})$, $i=1..n$, $j=1..k$. Пусть кол-во новых признаков – $m$, и они образуют матрицу $Z = (z_{ij})$, $i=1\dots n$, $j=1\dots m$, $m<k$.

И нам хочется, чтобы старые признаки по новым как-то восстанавливались, т.е. $\exists$ матрица $U = (u_{ij})$, $i=1\dots k$, $j=1\dots m$:

$\hat{X}_j = \sum_{\ell=1}^{m} z_\ell \cdot u_{j\ell}$, чтоб и $\hat{X}_j$ был как можно ближе к исходному $X_j$.

Т.е. хотим решить задачу $\Delta^2(Z,U) = \sum_{i=1}^{n} \cancel{\|X_i - \hat{X}_i} \sum_{j=1}^{k} (x_{ij} - \hat{x}_{ij})^2 =$

$= \sum_{j=1}^{k} \|X_j - \hat{X}_j\|^2 = \underbrace{\|ZU^T - X\|^2}_{\text{норма Фробениуса}} \to \min_{Z,U}$. Пусть $Z$ и $U$ ранга $m$, т.е. невырожд.

~~Из~~ **Теорема** Если $m \le \mathrm{rk}\, X$, то минимум $\Delta^2(Z,U)$ достигается, когда столбцы матрицы $U$ есть собств. вектора $X^TX$, соотв. $m$ максимальным собств. значениям. При этом $Z = X \cdot U$, матрицы $U$ и $Z$ ортогональны.

**Св-ва** 1) Матрица $U$ ортонормирована: $U^TU = I_m$

2) Матрица $Z^TZ = \Lambda = \mathrm{diag}(\lambda_1 \dots \lambda_m)$, где $\lambda_1 \ge \dots \ge \lambda_m$ – $m$ максимал. собств. значений матрицы $X^TX$.

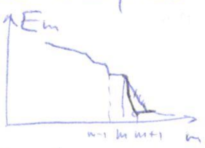
3) $U\Lambda = X^TXU$, $Z\Lambda = XX^TZ$

4) $\|ZU^T - X\|^2 = \|X\|^2 - \mathrm{tr}\,\Lambda = \sum_{j=m+1}^{k} \lambda_j$.

Из св-ва 4 вытекает, что чем меньше $E_m = \frac{\sum_{j=m+1}^{k} \lambda_j}{\sum_{j=1}^{k} \lambda_j}$, тем лучше новые признаки приближ. старые.

Поэтому вводится эффектив. размерность выборки: $\tilde{m} := \min_m \{E_m < \varepsilon\}$.

Кол-во новых признаков можно считать с помощью критерия «крутого склона»:




Находим $m+1$: $E_m \gg E_{m+1}$
В кач. эффектив. размер. берём $m+1$.

Решение задачи лин. регрессии в новых признаках: заменим $X$ на $ZU^T$, получ. след. задачу: $\|y - Z\underbrace{U^T\theta}_{\beta}\|^2 = \|y - Z\beta\|^2 \to \min_\beta$

Тогда $\hat{\beta} = D^{-1}V^Ty$, $\hat{\theta} = UD^{-1}V^Ty = \sum_{j=1}^{m} \frac{1}{\sqrt{\lambda_j}} u_j (v_j^T y)$, где $D = \sqrt{\Lambda}$,

а $V$ – связана с сингулярным разложением, теперь поговорим о нём.

## Сингулярное разложение (SVD)

$\forall$ матрица размера $n\times k$ $n>k$ представима в виде $X = VDU^T$, где

1) $n\times n$ матрица $V$ ортогональна, $V^TV = I_n$, её столбцы — собств. векторы матрицы $XX^T$

2) $n\times k$ матрица $D$ диагональна, $D = diag(\sqrt{\lambda_1}, \dots \sqrt{\lambda_k})$, где $\lambda_j$ — соб. знач. матриц $X^TX$ и $XX^T$.

3) $k\times k$ матрица $U$ ортогон., $U^TU = I_k$, столбцы $u_j$ — соб. векторы матр. $X^TX$.

Тогда решение задачи наим. квадратов вычисл. так: $D^{-1} = diag(\frac{1}{\sqrt{\lambda_1}}, \dots \frac{1}{\sqrt{\lambda_k}})$

$$\hat\theta = \underbrace{(X^TX)^{-1}X^T}_{\text{псевдообр. матрица}}y = (UD^T\underbrace{V^TV}_{I_n}DU^T)^{-1}UD^TV^Ty = UD^{-1}V^Ty = \sum_{j=1}^{k}\frac{1}{\sqrt{\lambda_j}}u_j(V_j^Ty)_j$$

$diag(1\dots1, 0\dots0)$

$$X\hat\theta = (VDU^T)U(D^T)^{-1}V^Ty = VV^Ty^k, \text{ ну и т.д.}$$

Для гребневой регрессии решение будет таким: $\hat\theta_{ridge} = \sum_{j=1}^{k}\frac{\sqrt{\lambda_j}}{\lambda_j+\tau}u_j(V_j^Ty)_j$.

Кстати, регуляризация сокращает эффектив. размерность (а мы знаем, что это ~~приводит~~ делает нашу модель более устойчивой), т.к.

$$tr(X(X^TX+\tau I_n)^{-1}X^T) = tr\left(diag\left(\frac{\lambda_j}{\lambda_j+\tau}\right)\right) = \sum_{j=1}^{k}\frac{\lambda_j}{\lambda_j+\tau} < n \quad (\text{это когда мы сделали } \sum\lambda_i = n)$$

## Связь сингуляр. разложения и метода глав. компонент

Если $k=m$ (т.е. кол-во признаков не уменьшаем), то $Z = V\cdot\sqrt{\Lambda}$.

Замечание Главные компоненты вычисляются по $X^TX$ и поэтому зависят от масштаба признаков $\Rightarrow$ перед их вычислением данные надо отнормировать.

## Ещё свойства метода глав. компонент

5) Проекции объектов на I глав. компоненту $c_1$ (соотв. соб. знач. $\lambda_1$, считаем, что $\lambda_1 \geq \lambda_2 \dots$) имеют наибольшую выбороч. дисперсию среди проекций на всевозможные направл. $d$ в пр-ве $\mathbb{R}^k$.

Далее, $\forall j\geq 2$ $c_j$ — $j$-тая глав. компонента, т.е. соб. вектор матрицы $X^TX$ с соб. знач. $\lambda_j$, — направление с наибольшей выбороч. дисперсией проекций объектов среди направл., $\perp$ векторам $c_1 \dots c_{j-1}$.

6) На пред. св-ве основан степенной метод вычисления глав. компонент

Пусть есть произвол. симм. матрица $A$, и её соб. числа $|\lambda_1| > |\lambda_2| \geq \dots \geq |\lambda_n|$ $c_1$ — соб. вектор, соотв. $\lambda_1$ и для нек. $x_0 \in \mathbb{R}^n$ $c_1^T\cdot x_0 \neq 0$.

Положим $x_{i+1} = Ax_i$

**Теорема** $b_{i+1} = \frac{x_i^Tx_{i+1}}{|x_i|^2} \to \lambda_1$ со скоростью геометрич. прогрессии со знаменателем $\gamma = |\lambda_2/\lambda_1| < 1$

Док-во: Пусть $x_0 = b_1\cdot c_1 + \dots + b_nc_n$. Поскольку $Ac_i = \lambda_ic_i$, то $A^ix_0 = x_i = \lambda_1^ib_1c_1 + \dots + \lambda_n^ib_nc_n = \lambda_1^i(b_1c_1 + \delta_i)$, где $\delta_i = b_2(\lambda_2/\lambda_1)^ic_2 + \dots$, причём $|\delta_i| = O(\gamma^i) \xrightarrow[i\to+\infty]{} 0$

Замечание Если $|\lambda_1| > 1$, то $|x_i| \to +\infty$, если $|\lambda_i| < 1$, то $|x_i| \to 0$ при $i \to \infty$

Поэтому на каждом шаге надо $x_i$ нормировать.

Если же вдруг $c_1^Tx_0 = 0$ (мы же заранее не знаем $c_1$), то стоит рассм. $n$ лин. незав. векторов — на одном из них $c_1^Tx_{0i} \neq 0$.

Алгоритм выч. $\lambda_k$ и $c_k$ (в предположении, что $\lambda_j$ и $c_j$ известны $\forall j \le k-1$) (относ. быстрый)

1) $i=0$, $t_0=0$, выбираем произвол. $x_0 \in \mathbb{R}^n$
2) Ортогонализируем $x_i$ по векторам $c_1, \ldots c_{k-1}$ (если $k=1$, то пропуск. этот шаг): $y_i = x_i - (c_1^T x_i) c_1 - \ldots - (c_{k-1}^T, x_i) c_{k-1}$
3) Нормируем $y_i$: $e_i = y_i / |y_i|$
4) Вычисляем $x_{i+1} = A e_i$ и $t_{i+1} = e_i^T x_{i+1}$
5) Нормируем $x_{i+1}$: $z_{i+1} = x_{i+1} / |x_{i+1}|$
6) Если $|t_{i+1} - t_i| \le \varepsilon$, то положим $\lambda_k = t_{i+1}$, $c_k = z_{i+1}$ и закончим процесс. Если нет, то выбираем $x_{i+1} := z_{i+1}$, $i := i+1$ и возвращ. к шагу 2.

Собственно, сингулярное разложение делается через QR-разложение матрицы $A = QR$, где $Q$ - ортогон., $R$ - верхнетреуг., которое, в свою очередь, получается похожей процедурой (методом ортогонал. Грама-Шмидта). А этот алгоритм нужен в том случае, если мы не хотим сразу все глав. компоненты искать, а только первые несколько.

Замечание Понижать размерность можно и с помощью сингулярного разложения, взяв в кач. нового пр-ва $m<k$ соб. векторов матрицы $X^T X$

МСПС | Нелинейные методы понижения размерности | (1)

Метод главных компонент — это прекрасно, но слепка прошлый век.
Его недостатки:
1) Если $|\lambda_i| = |\lambda_{i+1}|$ для какого-то $i$, то степенной метод не работает.
2) ~~Метод~~ PCA способен находить только лин. подпространства исход. пр-ва, которые „объясняют" данные с высокой точностью. На практике поверхность, вдоль кот. располагаются данные, может ~~быть~~ существенно отлич. от линейной.
3) PCA явл. инвариантным относ. поворота коорд. в пр-ве признаков ⇒ ⇒ восстановл. значений признаков может быть неоднозначным. Иногда это губит весь метод.

Многомер. шкалирование (локальное линейное погружение, LLE) local linear embedding

Цель, как и в ∀ методе пониж. размерности, миним. функ. некую ф-цию $F$, выражающее суммарное расхождение между заданными расст. между объектами $d_{ij}$ (где $d_{ij} = \rho(x_i, x_j)$, $\{x_i\}$ - объекты, ну или признаки) и расст. $\delta_{ij}$ между образами объектов в подпр-ве небольшой размер.

Например, $F_0 = \sum_{i<j} (\delta_{ij} - d_{ij})^2$ - станд. метод многомер. шкалирования.

Или метод Сэммона $F_1 = \frac{1}{\sum_{i<j} d_{ij}} \sum_{i<j} \frac{(\delta_{ij} - d_{ij})^2}{d_{ij}}$ (более точно передает небольшие различия и менее точно — большие, т.к. при отображ. бо́льших расстояний допустимы большие ошибки)

Теперь как искать $\min F$? Надо задать начал. конфигурацию:
а) можно проект. наши данные в некое подпр-во размер. $m$
б) или $m$ глав. комп. взять, в) или $m$ случ. векторов

Метод сопряжённых градиентов

1) Определяем $\vec{p_t}$ в пр-ве $\mathbb{R}^{nm}$ по ф-лам: $\vec{x_t} \in \mathbb{R}^{n \cdot m}$

$p_1 = -g_1$, $p_t = -g_t + \beta_t p_{t-1}$ при $t \geq 2$, где

$g_t = \left(\frac{\partial F}{\partial x_{11}}, \dots \frac{\partial F}{\partial x_{nm}}\right)$, $p_{t-1}$ - направл. на пред. шаге, $\beta_t = \frac{\|g_t\|^2}{\|g_{t-1}\|^2}$

2) Производится перемещ. до точки min по выбр. направлению:

$x_{t+1} = x_t + \alpha_t p_t$, где $F(x_t + \alpha_t p_t) = \min_{\alpha > 0} F(x_t + \alpha p_t)$. Можно брать и $\alpha_t = \frac{(x_t - x_{t-1})^T (g_t - g_{t-1})}{\|g_t - g_{t-1}\|^2}$

3) Если $|g_{t+1}| \leq \varepsilon$, где $\varepsilon$ мало, то заканчиваем.

Замечания: 1) Почему просто не взять $p_t = -g_t$? (Метод наискорейшего спуска)
Плохо работает, если ф-ция - „овраг", не останавливается.

2) $\frac{\partial F_1}{\partial x_{il}} = \frac{2}{c_1} \sum_{j=1, j \neq i}^{n} \left(\frac{1}{d_{ij}} - \frac{1}{\delta_{ij}}\right)(x_{il} - x_{jl})$, если метрика евклидова

3) Метод стохастич. градиента: случайно выбираем произвольный $\frac{\partial F}{\partial x_{ij}}$ и по нему идём. Сходится медленнее, зато не требует больших выч. затрат.

4) Метод Сэммона хорош тем, что позволяет работать с матрицами различий с пропусками. Для этого суммируем в $F_1$ только по тем парам объектов, по кот. пропусков нет. Хорошо работает, даже если доля пропусков 30%.

## Метод t-SNE.

Разработан в 2008. Намного лучше PCA, но работает медленнее.
Пусть $\mathcal{X} \subset \mathbb{R}^k$ — выбороч. пр-во векторов признаков наших изуч. объектов.
Пусть объекты $\mathcal{X}$ подчиняются гипотезе многообразия: $\exists f: \mathbb{R}^d \to \mathbb{R}^k$ — гладк.
$\forall x \in \mathcal{X} \ \exists z^* \in \mathbb{R}^d: \ x = f(z^*) + \varepsilon(z^*)$, где $\varepsilon(z^*)$ — центриров. сл. вектор с конеч. матрицей ковариаций. $d$ будем называть эфф. размерностью $\mathcal{X}$.
Но, увы, $d$ не знаем. Будем искать поиск реш. в пр-ве $\mathbb{R}^\ell$, $\ell < k$.
Рассм. выборку из $n$ объектов $x_1 \dots x_n \subset \mathcal{X}$, и $d_{ij} = P(x_i, x_j)$ и $\delta_{ij} = Q(x_i, x_j)$ — вер. меры сходства объектов в $\mathbb{R}^k$ и $\mathbb{R}^\ell$:

$$d_{i|j} = \frac{\exp\left(-\frac{1}{2\sigma_i^2}\|x_i - x_j\|^2\right)}{\sum\limits_{m \neq i} \exp\left(-\frac{1}{2\sigma_i^2}\|x_i - x_m\|^2\right)}, \quad d_{ij} = \frac{d_{i|j} + d_{j|i}}{2n};$$

$$\delta_{ij} = \frac{(1 + \|z_i - z_j\|^2)^{-1}}{\sum\limits_{m \neq i} (1 + \|z_i - z_m\|^2)^{-1}}, \quad \delta_{ii} = 0. \quad z_i \text{ — образы } x_i \text{ в } \mathbb{R}^\ell.$$

$\sigma_i$ — это bandwidth, и выбирается соответственно.

Далее минимизируем $F(d_{ij}, \delta_{ij})$, и $z_{min} = \underset{Z \in \mathbb{R}^{n \times \ell}}{\arg\min} F(d_{ij}, \delta_{ij})$, где

$$F(d_{ij}, \delta_{ij}) = \sum_{i \neq j} d_{ij} \cdot \ln \frac{d_{ij}}{\delta_{ij}}$$ — расстояние Кульбака-Лейблера.

Ну и задача решается градиентными методами.