Теория статистического обучения

Н. К. Животовский
nikita.zhivotovskiy@phystech.edu

23 мая 2016 г.

Материал находится в стадии разработки, может содержать ошибки и неточности. Автор будет благодарен за любые замечания и предложения, направленные по указанному адресу

1 Снижение размерности

Мотивации к снижению размерностей:

- Высокие размерности требуют больших вычислительных затрат.
- Классификация и регрессия в задачах высокой размерности ведет к низкой обобщающей способности.
- Данных обычно нужно гораздо больше чем размерность пространства.
- Малоразмерные данные иногда проще интерпретировать.

Самые простые способы снижения размерностей основаны на линейных преобразованиях. В случае конечномерных пространств для столбца $x \in \mathbb{R}^d$ соответствующее преобразование можно задать прямоугольной матрицей $W \in \mathbb{R}^{n,d}$.

§1.1 Метод главных компонент

Пусть даны векторы $x_1, \dots, x_m$ размерности d. С помощью матрицы $W \in \mathbb{R}^{n,d}$ отображаем вектор x_i в вектор $y_i = Wx_i$. Рассмотрим матрицы W_1, U_1, такие что

$$(W_1, U_1) = \arg \min_{W \in \mathbb{R}^{n,d}, U \in \mathbb{R}^{d,n}} \sum_{i=1}^{m} \|x_i - UWx_i\|_2^2. \tag{1.1}$$

В этом случае преобразование, соответствующее матрице W_1, осуществляет снижение размерности *методом главных компонент*. Говоря неформально, линейное преобразование W_1 отображает набор n-мерных векторов в набор d мерных векторов таким образом, что в сочетании с некоторым "обратным"линейным преобразованием U_1 сумма квадратов евклидовых норм восстановленных векторов меняется наименьшим возможным образом.

Лемма 1.1. *Пусть (W, U) — решение 1.1. Если матрица UW имеет ранг n, то столбцы матрицы U ортогональны (U^TU единичная матрица) и $W = U^T$.*

Рассмотрим подход без использования SVD-разложения.

Доказательство.

Пусть U, W — некоторые матрицы, такие, что UW имеет ранг n. Рассмотрим линейное отображение $x \to UWx$. Рассмотрим также образ этого отображения $R = \{UWx : x \in \mathbb{R}^d\}$. Он является не более чем n мерным подпространством в $\mathbb{R}^d$. Пусть $V \in \mathbb{R}^{d,n}$ матрица, столбцы которой образуют ортонормальный базис в данном подпространстве. Как и ранее V^TV — единичная матрица. Каждый вектор в образе может быть разложен по этому базису: R состоит из элементов вида Vy, $y \in \mathbb{R}^n$. Имеем,

$$\|x - Vy\|_2^2 = \|x\|^2 + y^TV^TVy - 2y^TV^Tx = \|x\|^2 + \|y\|^2 - 2y^T(V^T)x$$

Беря градиент по y получаем, что $y = V^Tx$. Таким образом, для всех x мы имеем

$$VV^Tx = \arg\min_{\tilde{x} \in R} \|x - \tilde{x}\|^2.$$

Так как данное соотношение выполнено для всех x_i-ых векторов, то заменяя UW на VV^T, мы не увеличиваем слагаемых. В итоге,

$$\sum_{i=1}^m \|x_i - UWx_i\|_2^2 \geq \sum_{i=1}^m \|x_i - VV^Tx_i\|_2^2$$

■

Таким образом решение ищется среди ортогональных матриц максимального ранга. Задача переписывается в виде

$$\arg \min_{U \in \mathbb{R}^{d,n}, U^TU=I} = \sum_{i=1}^m \|x_i - UU^Tx_i\|_2^2.$$

Далее,

$$\|x - UU^Tx\|_2^2 = \|x\|^2 - x^TUU^Tx = \|x\|^2 - tr(U^Txx^TU).$$

Минимизация последнего выражения по U эквивалентна поиску матрицы

$$U_1 = \arg \max_{U \in \mathbb{R}^{d,n}, U^TU=I} tr(U^T \sum_{i=1}^m x_ix_i^TU), \tag{1.2}$$

где в посленем равенстве мы используем линейность следа и равенство $tr(AB) = tr(BA)$.

Лемма 1.2. *Решением задачи 1.2 является матрица U, столбцы которой есть $u_1, \ldots, u_n$ — собственные векторы, соответсвующие наибольшим собственным значениям матрица $A = \sum_{i=1}^m x_ix_i^T$.*

Доказательство.

Приведем симметричную матрицу A к диагональному виду: $A = VDV^T$. Считаем, что собственные значения на диагонали расположены по убыванию. Тогда

$$U^TVDV^TU = B^TDB,$$

где $B = V^T U$. Прямой подсчет показывает, что

$$tr(B^T D B) = \sum_{j=1}^{d} D_{j,j} \sum_{i=1}^{n} B_{j,i}^2.$$

Но матрица $B^T B$ — единичная, поэтому $\sum\limits_{j=1}^{d}\sum\limits_{i=1}^{n} B_{j,i}^2 = n$. Достроим квадратную ортогональную матрицу $\tilde{B}$ посредством присоединения к B дополнительных столбцов. Отсюда легко понять, что $\sum\limits_{i=1}^{n} B_{j,i}^2 \leq n$. Таким образом,

$$tr(U^T A U) \leq \max_{\beta \in [0,1]^d, \|\beta\|_1 \leq n} \sum_{j=1}^{d} D_{j,j}\beta_j. \tag{1.3}$$

В решении оставляем n самых больших компонент $D_{i,j}$. Одновременно, задав в качестве компонент матрицы U n собственных векторов, соответствующих наибольшим собственным значениям, получим достижение равенства в 1.3. ■

§1.2 Случайные проекции

Случайные проекции еще один способ снижения размерности. Идея возникла в 70-ые годы с помощью случайных линейных преобразований удалось доказать существование преобразований сохраняющих расстояния и снижающих размерности. Будем неформально говорить, что линейное преобразование, соответствующее W слабо изменяет расстояния, если соотношение

$$\frac{\|Wx_1 - Wx_2\|}{\|x_1 - x_2\|}$$

близко к единице для любой пары векторов x_1, x_2 из заданного множества.

Лемма 1.3. *Пусть $Z \sim \chi_n^2$, тогда для $\varepsilon \in (0, 3)$*

$$\Pr\{(1-\varepsilon)n \leq Z \leq (1+\varepsilon)n\} \geq 1 - 2\exp(\varepsilon^2 n/6).$$

Лемма 1.4. *Пусть $x \in \mathbb{R}$, а $W \in \mathbb{R}^{\times,}$ — случайная матрица, состоящая из независимых стандартных нормальных величин. Тогда для всех $\varepsilon \in (0, 3)$*

$$\Pr\left\{\left|\frac{\|(1/\sqrt{n})Wx\|^2}{\|x\|^2} - 1\right| \geq \varepsilon\right\} \leq 2\exp(\varepsilon^2 n/6).$$

Без ограничения общности считаем, что $\|x\|^2 = 1$. Тогда эквивалентая формулировка последнего условия имеет вид

$$\Pr\left\{(1-\varepsilon)n \leq \|Wx\|^2 \leq (1+\varepsilon)n\right\} \geq 1 - 2\exp(\varepsilon^2 n/6). \tag{1.4}$$

Легко видеть, что $\|Wx\|^2$ имеет распределение χ_n^2.

Лемма 1.5 (Лемма Джонсона–Линденштраусса). *Пусть дано множество $X \subset \mathbb{R}^d$ из m векторов. Для $\delta \in (0,1)$ фиксируем*

$$\varepsilon = \sqrt{\frac{6\log(2m/\delta)}{n}} \leq 3 \tag{1.5}$$

Если матрица W состоит из независимых нормальных величин с математическим ожиданием 0 и дисперсией $\frac{1}{n}$, то с вероятностью $1-\delta$ одновременно для всех $x \in X$:

$$\left| \frac{\|Wx\|^2}{\|x\|^2} - 1 \right| \leq \varepsilon.$$

Доказательство.

Доказательство получается комбинацией неравенства Буля и неравенства 1.4.■

Замечание 1.1. В лемме Джонсона–Линденштраусса нет зависимости от размерности исходного пространства!

§1.3 Compressed Sensing

Еще одним хорошо изученным способом сжатия данных / снижения размерности задачи является так называемый Compressed Sensing. Как и ранее в этом методе сжатие производится с помощью линейного преобразования. Однако в данном случае предполагается, что исходные векторы в многомерном пространстве имеют лишь небольшое количество ненулевых компонент. Такие векторы часто именуются sparse-векторами.

Опр. 1.1. *Матрица $W \in \mathbb{R}^{n,d}$ обладает свойством RIP(ε, s), если для любого вектора x, $\|x\|_0 \leq s$ (s-sparse вектор)*

$$\left| \frac{\|Wx\|_2^2}{\|x\|_2^2} - 1 \right| \leq \varepsilon.$$

Из определния следует, что матрица, обладающая таким свойством, незначительно изменяет нормы sparse векторов, то есть на соответствующих векторах преобразование практически ортогонально.

Теорема 1.6 (Точное восстановление вектора). *Пусть $\varepsilon < 1$, а матрица W является $(\varepsilon, 2s)$-RIP матрицей. Пусть ненулевой вектор x обладает свойством $\|x\|_0 \leq s$, а $y = Wx$. Тогда решение задачи*

$$\min_{\tilde{x}: y = W\tilde{x}} \|\tilde{x}\|_0 \tag{1.6}$$

в точности совпадает с x.

Доказательство.

Пусть найдется $\tilde{x} \neq x$. Но мы имеем $\|\tilde{x}\|_0 \leq \|x\|_0 \leq s$, тогда $\|\tilde{x} - x\|_0 \leq 2s$. Из линейности задачи следует, что $W(\tilde{x} - x) = 0$. А это противоречит свойству RIP. ■

Проблема оптимизационной задачи 1.6 заключается в том, что на практике сложно найти эффективный неперебoрный алгоритм ее решения. Следующий нетривиальный результат существенно упрощает эту задачу.

Теорема 1.7 (сведение к выпуклой задаче). *Предположим, что* $\varepsilon < \frac{1}{1+\sqrt{2}}$. *Тогда*

$$x = \arg\min_{\tilde{x}:y=W\tilde{x}} \|\tilde{x}\|_0 = \arg\min_{\tilde{x}:y=W\tilde{x}} \|\tilde{x}\|_1.$$

Задача поиска вектора $x = \arg\min_{\tilde{x}:y=W\tilde{x}} \|\tilde{x}\|_1$ хорошо изучена. Можно доказать, что эта задача сводится к задаче линейного программирования вида $u_1 + \ldots + u_n$ при условиях $Ax = b, -u \leq x \leq u, x, u \in \mathbb{R}^n, u \geq 0$. Эффективный метод решения этой задачи носит название basis pursuit algorithm.

Следующая теорема указывает способ построения матриц, с большой вероятностью обладающих RIP свойством.

Теорема 1.8. *Пусть* $\varepsilon, \delta \in (0,1)$, $s \leq d$ *и*

$$n \geq 100\frac{s\log(40d/\delta\varepsilon)}{\varepsilon^2}.$$

Если матрица $W \in \mathbb{R}^{n,d}$ состоит из независимых нормальных величин с математическим ожиданием 0 и дисперсией $\frac{1}{n}$, то с вероятностью $1-\delta$ матрица W является (ε, s)-RIP матрицей. Итак,

- Есть надежда на точное восстановление вектора, если матрица W обладает свойством RIP. Это свойство как раз заключается в том, что соответствующее линейное преобразование мало изменяет норму вектора, который имеет разреженное представление, то есть имеет небольшое число ненулевых компонент.
- Вектор может быть эффективно восстановлен.
- Случайные гауссовские матрицы специальных размеров с большой вероятностью обладают свойством RIP.

§1.4 Сравнение Compressed Sensing с PCA

- Матрица в Compressed Sensing не зависит от наблюдений и является одной и той же одновременно для всех векторов нужной степенью разреженности.
- Рассмотрим множество векторов в $\mathbb{R}^d$, представляющее стандартный базис в этом пространстве. Для этих векторов $s = 1$ и размерность сжатого пространства n имеет порядок $\log(d)$. Одновременно нет надежды на хорошее восстановление с помощью PCA так как векторы живут в многомерном подпространстве.
- Пусть в PCA все векторы лежат в точности в некотором n–мерном подпространстве d–мерного пространства, тогда PCA даст нам тривиальное снижение размерности, которое имеет точное восстановление. Легко показать, что Compressed sensing в этом случае является менее эффективным способ сжатия.

Список литературы

[1] *Candes E., Tao T.* Decoding by linear programming // IEEE Trans. Info. Th., vol. 51, no. 12, pp. 4203–4215, Probability and Statistics, 2005.

[2] *Donoho D.* Compressed sensing // IEEE Trans. Info. Th., vol. 52, no. 4, pp. 1289–1306, 2006.

[3] *Shalev-Shwartz S., Ben-David S.* Understanding Machine Learning: From Theory to Algorithms // Cambridge University Press, 2014