СТАТЬИ

УДК 004.93

ЛИНЕЙНЫЙ ДИСКРИМИНАНТНЫЙ АНАЛИЗ КАК КОНТРОЛИРУЕМЫЙ ПОДХОД В ЗАДАЧАХ УМЕНЬШЕНИЯ РАЗМЕРНОСТИ ДАННЫХ

Петухов Д.Е., Ткаченко А.В., Белов Ю.С.

Московский государственный технический университет имени Н.Э. Баумана, филиал, Калуга,
e-mail: petuhoff.dmitrij@yandex.ru, fn1-kf@mail.ru

Линейный дискриминантный анализ (LDA) является очень распространенным методом для решения задач уменьшения размерности в качестве шага предварительной обработки для задач машинного обучения и классификации шаблонов. В то же время он обычно используется как черный ящик, но не всегда хорошо понимается. Метод LDA преодолевает ограничение не менее известного метода уменьшения размерности – метода главных компонент (PCA) – путем использования линейного дискриминантного критерия. Этот критерий пытается максимизировать отношение определителя матрицы рассеяния между классами спроецированных образцов к определителю матрицы рассеяния внутри класса спроецированных образцов. Цель этой статьи – предоставить понимание того, что такое LDA, как LDA работает, что даст различным группам читателей возможность лучше понять LDA и знать, как применять эту технику в различных задачах. В статье представлено основное определение метода LDA. Кроме того, подробно описаны два метода вычисления пространства LDA, т. е. зависимые от класса и не зависимые от класса методы. В статье рассматриваются ключевые этапы алгоритма LDA, а также приведен пример использования данного метода. Также был проведен сравнительный анализ LDA с PCA.

Ключевые слова: уменьшение размерности, PCA, LDA, зависимое от класса преобразование, независимое от класса преобразование

LINEAR DISCRIMINANT ANALYSIS AS A SUPERVISED APPROACH TO REDUCING DATA SIZE

Petukhov D.E., Tkachenko A.V., Belov Yu.S.

Bauman Moscow State Technical University, branch, Kaluga,
e-mail: petuhoff.dmitrij@yandex.ru, fn1-kf@mail.ru

Linear discriminant analysis (LDA) is a very common method for dimensionality reduction tasks as a preprocessing step for machine learning and template classification tasks. At the same time, it is usually used as a black box, but is not always well understood. The LDA method overcomes the limitation of an equally well-known method of reducing dimension – the principal component analysis (PCA), using a linear discriminant criterion. This criterion attempts to maximize the ratio of the scattering matrix determinant between the classes of projected samples to the scattering matrix determinant within the class of projected samples. The purpose of this article is to provide an understanding of what the LDA is and how the LDA works, which will provide different groups of readers with the opportunity to better understand the LDA and know how to apply this technique in various tasks. The article gives a basic definition of the LDA method. In addition, two methods for calculating the LDA space are described in detail, i.e., class-dependent and class-independent methods. The article discusses the key stages of the LDA algorithm, as well as an example of the use of this method. Finally, a comparative analysis of LDA with PCA was performed.

Keywords: dimensionality reduction, PCA, LDA, Class-dependent transformation, Class-independent transformation

Методы уменьшения размерности важны во многих приложениях, связанных с машинным обучением, распознаванием образов [1], интеллектуальным анализом данных, биоинформатикой, биометрией и поиском информации. Основная цель методов уменьшения размерности состоит в уменьшении размеров путем удаления избыточных и зависимых объектов посредством преобразования объектов из пространства с более высокой размерностью в пространство с меньшими размерами. Существует два основных подхода по уменьшению размерности: неконтролируемый и контролируемый. При неконтролируемом подходе нет необходимости в маркировке классов данных, в то время как в контролируемом подходе методы уменьшения размерности учитывают метки классов. Существует множество методов сокращения неконтролируемой размерности, например независимый компонентный анализ (**I**CA) и неотрицательная матричная факторизация (NMF), но наиболее известным методом бесконтрольного подхода является анализ главных компонент (PCA) [2]. Этот тип уменьшения данных подходит для многих приложений, таких как визуализация и удаление шума [3]. Контролируемый подход также имеет много методов, таких как смешанный дискриминантный анализ (MDA) и нейронные сети (NN), но наиболее извест-

ным методом этого подхода является линейный дискриминантный анализ (LDA). Эта категория методов уменьшения размерности используется в биометрии, биоинформатике и химии. Метод LDA разработан для преобразования объектов в пространство более низкой размерности, которое максимизирует отношение дисперсии между классами к дисперсии внутри классов, тем самым гарантируя максимальную отделимость классов. Существует два типа метода LDA для работы с классами: зависящий от класса и независимый от класса. В классозависимом LDA для каждого класса вычисляется одно отдельное пространство нижних измерений для проецирования на него своих данных, тогда как в классонезависимом LDA каждый класс будет рассматриваться как отдельный класс по отношению к другим классам. В этом типе существует только одно пространство нижнего измерения для всех классов, чтобы проецировать на него свои данные. Хотя метод LDA считается наиболее широко используемым методом сокращения данных, он страдает от ряда проблем. Первой проблемой является то, что LDA не может найти пространство с меньшей размерностью, если размеры намного превышают число выборок в матрице данных. Таким образом, матрица внутри класса становится сингулярной, что известно как проблема малой выборки (SSS). Существуют различные подходы, которые предлагаются для решения этой задачи. Первый подход заключается в удалении нулевого пространства матрицы внутри класса, второй подход использует промежуточное подпространство (например, PCA) для преобразования матрицы внутри класса в матрицу полного ранга; таким образом, она может быть инвертирована. Третий подход, хорошо известное решение, заключается в использовании метода регуляризации для решения сингулярных линейных систем. Во второй задаче – задаче линейности, если различные классы нелинейно разделимы, LDA не может различать эти классы. Одним из решений этой проблемы является использование функций ядра [4].

Цель исследования: рассмотреть, что представляет собой LDA и как он работает, изучить алгоритм и методы вычисления пространства LDA, сравнить LDA с PCA.

Линейный дискриминантный анализ (LDA) – обобщение линейного дискриминанта Фишера – метода, используемого в статистике, распознавании образов и машинном обучении для поиска линейной комбинации признаков, которая характеризует или разделяет два или более классов объектов или событий. Этот метод проецирует набор данных на пространство с более низкой размерностью с хорошей разделимостью классов, чтобы избежать перенапряжения («проклятие размерности») и снизить вычислительные затраты. Полученная комбинация может быть использована в качестве линейного классификатора или чаще – для уменьшения размерности перед последующей классификацией [5].

Линейный дискриминантный анализ – контролируемый алгоритм, который учитывает меченые данные при использовании метода уменьшения размерности. Этот метод применяется для поиска нового пространства признаков, которое максимизирует разделимость классов путем использования подхода, очень похожего на тот, который применяется в анализе главных компонентов (PCA).

PCA – статистическая процедура, которая преобразует набор возможных коррелированных переменных в набор линейно некоррелированных признаков, называемых главными компонентами. Он существенно отбрасывает наименее важные переменные, сохраняя при этом ценные, находя главные компонентные оси, вдоль которых дисперсия данных высока [2].

В LDA мы намерены максимизировать разделение между классами путем максимизации расстояния между центроидами классов и в то же время минимизировать дисперсию внутри класса, чтобы образовались хорошо разделенные неперекрывающиеся кластеры. Минимизация внутриклассовой дисперсии приводит к созданию компактных, менее распространенных классов. Эти функции называются дискриминантными функциями.

Различные подходы к LDA

Наборы данных могут быть преобразованы, и тестовые векторы могут быть классифицированы в преобразованном пространстве двумя различными подходами [6].

Преобразование, зависящее от класса: этот тип подхода предполагает максимизацию отношения дисперсии между классами к дисперсии внутри классов. Основная цель состоит в том, чтобы максимизировать это соотношение так, чтобы была получена адекватная сепарабельность классов. Подход к классовому типу предполагает использование двух критериев оптимизации для независимого преобразования наборов данных.

Преобразование, не зависящее от класса: этот подход предполагает максимизацию отношения общей дисперсии к дисперсии внутри класса. Он использует только один критерий оптимизации для преобразования

наборов данных, поэтому все точки данных независимо от их классовой принадлежности преобразуются с помощью этого преобразования. В этом типе LDA каждый класс рассматривается как отдельный класс по сравнению со всеми другими классами.

Алгоритм LDA

LDA основан на концепции поиска линейной комбинации переменных (предикторов), которая наилучшим образом разделяет два класса (цели). Чтобы захватить понятие отделимости, Фишер определил следующую функцию оценки [4].

$$Z = \beta_1 x_1 + \beta_2 x_2 + ... + \beta_d x_d ,$$

$$S(\beta) = \frac{\beta^T \eta_1 - \beta^T \eta_2}{\beta^T C \beta} \text{ – функция оценки,}$$

$$S(\beta) = \frac{\overline{Z}_1 - \overline{Z}_2}{\text{дисперсия } Z \text{ внутри групп}}.$$

С учетом функции оценки задача состоит в том, чтобы определить линейные коэффициенты, максимизирующие оценку. Эта задача может быть решена с помощью следующих уравнений.

$\beta = C^{-1}(\mu_1 - \mu_2)$ – коэффициенты модели;

$C = \frac{1}{n_1 + n_2}(n_1 C_1 + n_2 C_2)$ – объединенная ковариационная матрица,

где β – коэффициент линейной модели;
C_1, C_2 – ковариационные матрицы;
μ_1, μ_2 – средние векторы.

Одним из способов оценки эффективности дискриминации является расчет расстояния Махаланобиса между двумя группами [7]. Расстояние больше 3 свидетельствует, что в двух средних различаются более чем на 3 стандартных отклонения. Это означает, что перекрытие (вероятность неправильной классификации) довольно мало.

$$\Delta^2 = \beta^T (\mu_1 - \mu_2),$$

Δ – расстояние Махаланобиса между двумя группами.

Наконец, новая точка классифицируется с проецированием ее на максимально разделяющее направление и классификацией ее как C1, если:

$$\beta^T \left(x - \left(\frac{\mu_1 + \mu_2}{2} \right) \right) > -\log \frac{p(c_1)}{p(c2)},$$

где β^T – вектор коэффициентов;
x – векторные данные;

$\frac{\mu_1 + \mu_2}{2}$ – средний вектор;

$\log \frac{p(c_1)}{p(c2)}$ – вероятность класса.

Пример:

Предположим, что мы получили набор данных от банка относительно его клиентов малого бизнеса, которые не выполнили свои обязательства (красный квадрат), и тех, кто не выполнил (синий круг), разделенных просроченными днями (DAYSDELQ) и количеством месяцев в бизнесе (BUSAGE). Мы используем LDA для поиска оптимальной линейной модели, которая наилучшим образом разделяет два класса (default и non-default) (рис. 1).

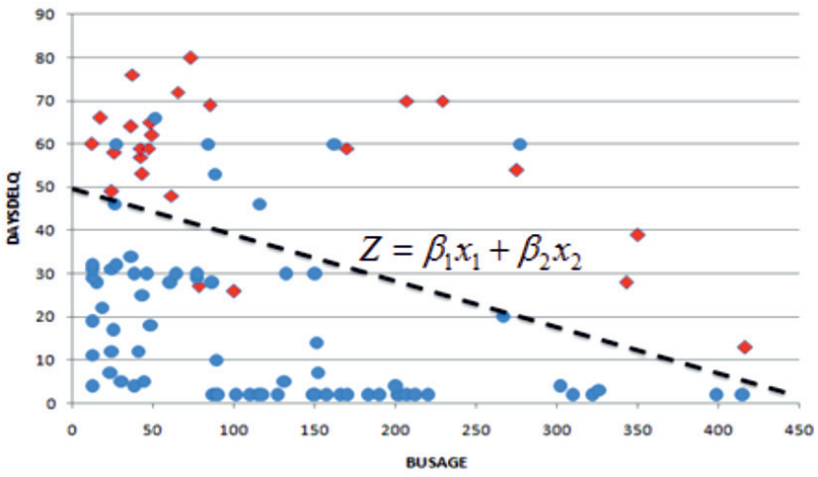

Рис. 1. График зависимости

Первым шагом является вычисление средних векторов, ковариационных матриц и вероятностей классов (табл. 1).

Затем мы вычисляем объединенную ковариационную матрицу и, наконец, коэффициенты линейной модели.

$$C = \frac{1}{n_1 + n_2}(n_1 C_1 + n_2 C_2) = \begin{bmatrix} 10495 & -718 \\ -718 & 322 \end{bmatrix},$$

$$\beta = C^{-1}(\mu_1 - \mu_2) = \begin{bmatrix} -0,0095 & -0,1408 \end{bmatrix},$$

$$Z = -0,0095 BUSAGE - 0,1408 DAYSDELQ.$$

Расстояние Махаланобиса 2,32 показывает небольшое перекрытие между двумя группами, что означает хорошее разделение между классами по линейной модели.

$$\Delta^2 = \beta^T(\mu_1 - \mu_2) = 5,40,$$

$$\Delta = 2,32.$$

В следующей таблице мы рассчитываем Z-балл, используя приведенное выше уравнение Z. Однако оценка сама по себе не может быть применена для прогнозирования результата. Нам также нужно уравнение в столбце 5, чтобы выбрать класс N или Y. Мы прогнозируем класс N, если вычисленное значение больше –1,1, в противном случае класс Y. Как показано ниже, модель LDA сделала две ошибки (табл. 2).

$$\beta^T\left(x - \left(\frac{\mu_1 + \mu_2}{2}\right)\right) > -\log\frac{p(c_1)}{p(c2)} =$$

$$= -\log\left(\frac{0,75}{0,25}\right) = -1,1.$$

Сравнение LDA с PCA

Как линейный дискриминантный анализ (LDA), так и анализ главных компонент (PCA) являются методами линейного преобразования, которые обычно используются для уменьшения размерности (оба метода служат методами факторизации матрицы данных). Наиболее важным отличием между обоими методами является то, что PCA может быть описан как «неконтролируемый» алгоритм, поскольку он «игнорирует» метки классов и его цель – найти направления (так называемые главные компоненты), которые максимизируют дисперсию в наборе данных, в то время как LDA является «контролируемым» алгоритмом, который вычисляет направления («линейные дискриминанты»), представляющие оси, которые максимизируют разделение между несколькими классами [8].

Таблица 1

Средние векторы, ковариационные матрицы и вероятности классов

Класс	Счет	Вероятность	Статистика	BUSAGE	DAYSDELQ
N	$n_1 = 75$	$p(c_1) = 0,75$	Средние векторы (μ)	116,23	16,89
			Ковариационная матрица	$\begin{bmatrix} 9323 & -607 \\ -607 & 333 \end{bmatrix}$	
Y	$n_2 = 25$	$p(c_2) = 0,25$	Средние векторы (μ)	115,04	55,32
			Ковариационная матрица	$\begin{bmatrix} 14009 & -1053 \\ -1053 & 287 \end{bmatrix}$	

Таблица 2

Вычисление Z-оценки и прогнозирование

BUSAGE	DAYSDELQ	Класс	Z-оценка	$\beta^T\left(x - \left(\frac{\mu_1 + \mu_2}{2}\right)\right)$	Прогнозирование
87	2	N	–1,1081	5,0752	N
89	2	N	–1,1271	5,0562	N
90	2	N	–1,1366	5,0466	N
116	46	N	–7,5788	–1,.3965	Y
88	53	N	–8,2984	–2,1155	Y
42	57	Y	–8,4246	–2,2405	Y
26	58	Y	–8,4134	–2,2289	Y
42	59	Y	–8,7062	–2,5221	Y

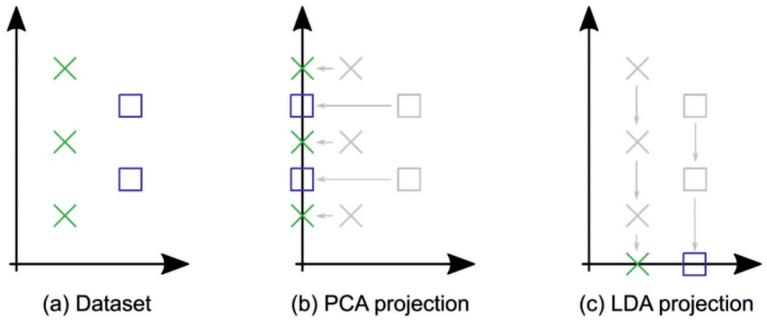

Рис. 2. Пример различного поведения подходов LDA и PCA для одного и того же набора данных. Набор данных представлен в (a). В (b) показан набор данных, спроецированный на основе того, что выбрал бы PCA, поскольку это максимизирует дисперсию независимо от меток классов (синие и зеленые цвета в этом случае). В (c) показан набор данных, спроецированный на основе того, что выбрал бы LDA, поскольку это максимизирует разделение классов, в данном случае зеленые и синие записи [9]

Они преследуют две совершенно разные цели (рис. 2):

– PCA ищет направления (компоненты), которые максимизируют дисперсию в наборе данных, поэтому ему не нужно рассматривать метки классов;

– LDA ищет направления (компоненты), которые максимизируют разделение классов, и для этого ему нужны метки классов.

Заключение

В статье были представлены и объяснены основные сведения о LDA. Авторы работы стремились дать общие сведения о том, что такое LDA, описать подходы к LDA, основной алгоритм LDA, а также было представлено небольшое сравнение алгоритмов LDA и PCA.

Список литературы

1. Гришанов К.М., Белов Ю.С. Методы выделения признаков для распознавания символов // Электронный журнал: наука, техника и образование. 2016. № 1 (5). С. 110–119. [Электронный ресурс]. URL: http://nto-journal.ru/catalog/informacionnye-texnologii/109/ (дата обращения: 17.01.2020).

2. Tharwat A., Principal component analysis-a tutorial. International Journal of Applied Pattern Recognition. 2016. Vol. 3. No. 3. P. 197–240. DOI: 10.1504/IJAPR.2016.079733.

3. Белов Ю.С., Нифонтов С.В., Азаренко К.А. Применение вейвлет-фильтрации для шумоподавления в речевых сигналах // Фундаментальные исследования. 2017. № 4–1. С. 29–33.

4. Alaa Tharwat, Abdelhameed Ibrahim, Aboul Ella Hassanien, Tarek Gaber, Linear discriminant analysis: A detailed tutorial. Ai Communications. May 2017. vol. 30. no. 2. P. 169–190. [Electronic resource]. URL: https://www.researchgate.net/publication/316994943_Linear_discriminant_analysis_A_detailed_tutorial (date of access: 17.01.2020).

5. Leo M.L. Nollet, Linear Discriminant Analysis. Hyperspectral Imaging Analysis and Applications for Food Quality, 2018. P. 115–118.

6. Davidson T., Thileepavathi D. Face and Expression Recognition under Illumination and Occlusion Using GSRRR and ICP Framework. International Journal of Advanced Research in Science, Engineering and Technology. February 2015. Vol. 2. Issue 2. P. 457–462. [Электронный ресурс]. URL: http://ijarset.com/upload/2015/february/11_IJARSET_titus.pdf (дата обращения: 17.01.2020).

7. Lahav A, Talmon R, Kluger Y, Mahalanobis distance informed by clustering. A Journal of the IMA. June 2019. Vol. 8. Issue 2. P. 377–406. DOI: 10.1093/imaiai/iay011.

8. Martinez A.M., Kak A.C., PCA versus LDA. IEEE Transactions on Pattern Analysis and Machine Intelligence. February 2001. Vol. 23. Issue 2. P. 228–233. DOI: 10.1109/34.908974.

9. Alberti M., Seuret M., Pondenkandath V., Ingold R., Liwicki M., Historical Document Image Segmentation with LDA-Initialized Deep Neural Networks.HIP2017: Proceedings of the 4th International Workshop on Historical Document Imaging and Processing. 2017. P. 95–100. DOI: 10.1145/3151509.3151519.