linear discriminant analysis - wiki

Not to be confused with latent Dirichlet allocation.

Linear discriminant analysis (**LDA**) is a generalization of **Fisher's linear discriminant**, a method used in

Ronald Fisher

statistics, pattern recognition and machine learning to find a linear combination of features that characterizes or separates two or more classes of objects or events. The resulting combination may be used as a linear classifier, or, more commonly, for dimensionality reduction before later classification.

LDA is closely related to analysis of variance (ANOVA) and regression analysis, which also attempt to express one dependent variable as a linear combination of other features or measurements.[1][2] However, ANOVA uses categorical independent variables and a continuous dependent variable, whereas discriminant analysis has continuous independent variables and a categorical dependent variable (*i.e.* the class label).[3] Logistic regression and probit regression are more similar to LDA than ANOVA is, as they also explain a categorical variable by the values of continuous independent variables. These other methods are preferable in applications where it is not reasonable to assume that the independent variables are normally distributed, which is a fundamental assumption of the LDA method.

LDA is also closely related to principal component analysis (PCA) and factor analysis in that they both look for linear combinations of variables which best explain the data.[4] LDA explicitly attempts to model the difference between the classes of data. PCA on the other hand does not take into account any difference in class, and factor analysis builds the feature combinations based on differences rather than similarities. Discriminant analysis is also different from factor analysis in that it is not an interdependence technique: a distinction between independent variables and dependent variables (also called criterion variables) must be made.

LDA works when the measurements made on independent variables for each observation are continuous quantities. When dealing with categorical independent variables, the equivalent technique is discriminant correspondence analysis.[5][6]

1 LDA for two classes

Consider a set of observations $\vec{x}$ (also called features, attributes, variables or measurements) for each sample of an object or event with known class y. This set of samples is called the training set. The classification problem is then to find a good predictor for the class y of any sample of the same distribution (not necessarily from the training set) given only an observation $\vec{x}$.[7]:338

LDA approaches the problem by assuming that the conditional probability density functions $p(\vec{x}|y = 0)$ and $p(\vec{x}|y = 1)$ are both normally distributed with mean and covariance parameters $(\vec{\mu}_0, \Sigma_0)$ and $(\vec{\mu}_1, \Sigma_1)$, respectively. Under this assumption, the Bayes optimal solution is to predict points as being from the second class if the log of the likelihood ratios is below some threshold T, so that:

$$(\vec{x} - \vec{\mu}_0)^T \Sigma_0^{-1} (\vec{x} - \vec{\mu}_0) + \ln|\Sigma_0| - (\vec{x} - \vec{\mu}_1)^T \Sigma_1^{-1} (\vec{x} - \vec{\mu}_1) - \ln|\Sigma_1| \ > \ T$$

Without any further assumptions, the resulting classifier is referred to as QDA (quadratic discriminant analysis).

LDA instead makes the additional simplifying homoscedasticity assumption (*i.e.* that the class covariances are identical, so $\Sigma_0 = \Sigma_1 = \Sigma$) and that the covariances have full rank. In this case, several terms cancel:

$$\vec{x}^T \Sigma_0^{-1} \vec{x} = \vec{x}^T \Sigma_1^{-1} \vec{x}$$
$$\vec{x}^T {\Sigma_i}^{-1} \vec{\mu_i} = \vec{\mu_i}^T {\Sigma_i}^{-1} \vec{x} \text{ because } \Sigma_i \text{ is Hermitian}$$

and the above decision criterion becomes a threshold on the dot product

$$\vec{w} \cdot \vec{x} > c$$

for some threshold constant c, where

$$\vec{w} = \Sigma^{-1} (\vec{\mu}_1 - \vec{\mu}_0)$$

$$c = \frac{1}{2} (T - \vec{\mu_0}^T \Sigma_0^{-1} \vec{\mu_0} + \vec{\mu_1}^T \Sigma_1^{-1} \vec{\mu_1})$$

This means that the criterion of an input $\vec{x}$ being in a class y is purely a function of this linear combination of the known observations.

It is often useful to see this conclusion in geometrical terms: the criterion of an input $\vec{x}$ being in a class y is purely a function of projection of multidimensional-space point $\vec{x}$ onto vector $\vec{w}$ (thus, we only consider its direction). In other words, the observation belongs to y if corresponding $\vec{x}$ is located on a certain side of a hyperplane perpendicular to $\vec{w}$. The location of the plane is defined by the threshold c.

2 Canonical discriminant analysis for *k* classes

Canonical discriminant analysis (CDA) finds axes (*k* - 1 canonical coordinates, *k* being the number of classes) that best separate the categories. These linear functions are uncorrelated and define, in effect, an optimal $k-1$ space through the *n*-dimensional cloud of data that best separates (the projections in that space of) the k groups. See "Multiclass LDA" for details below.

3 Fisher's linear discriminant

The terms *Fisher's linear discriminant* and *LDA* are often used interchangeably, although Fisher's original article[1] actually describes a slightly different discriminant, which does not make some of the assumptions of LDA such as normally distributed classes or equal class covariances.

Suppose two classes of observations have means $\vec{\mu}_0, \vec{\mu}_1$ and covariances Σ_0, Σ_1 . Then the linear combination of features $\vec{w} \cdot \vec{x}$ will have means $\vec{w} \cdot \vec{\mu}_i$ and variances $\vec{w}^T \Sigma_i \vec{w}$ for $i = 0, 1$. Fisher defined the separation between these two distributions to be the ratio of the variance between the classes to the variance within the classes:

$$S = \frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{within}}} = \frac{(\vec{w} \cdot \vec{\mu}_1 - \vec{w} \cdot \vec{\mu}_0)^2}{\vec{w}^T \Sigma_1 \vec{w} + \vec{w}^T \Sigma_0 \vec{w}} = \frac{(\vec{w} \cdot (\vec{\mu}_1 - \vec{\mu}_0))^2}{\vec{w}^T (\Sigma_0 + \Sigma_1) \vec{w}}$$

This measure is, in some sense, a measure of the signal-to-noise ratio for the class labelling. It can be shown that the maximum separation occurs when

$$\vec{w} \propto (\Sigma_0 + \Sigma_1)^{-1} (\vec{\mu}_1 - \vec{\mu}_0)$$

When the assumptions of LDA are satisfied, the above equation is equivalent to LDA.

Be sure to note that the vector $\vec{w}$ is the normal to the discriminant hyperplane. As an example, in a two dimensional problem, the line that best divides the two groups is perpendicular to $\vec{w}$.

Generally, the data points to be discriminated are projected onto $\vec{w}$; then the threshold that best separates the data is chosen from analysis of the one-dimensional distribution. There is no general rule for the threshold. However, if projections of points from both classes exhibit approximately the same distributions, a good choice would be the hyperplane between projections of the two means, $\vec{w} \cdot \vec{\mu}_0$ and $\vec{w} \cdot \vec{\mu}_1$. In this case the parameter c in threshold condition $\vec{w} \cdot \vec{x} > c$ can be found explicitly:

$$c = \vec{w} \cdot \frac{1}{2}(\vec{\mu}_0 + \vec{\mu}_1) = \frac{1}{2} \vec{\mu}_1^T \Sigma^{-1} \vec{\mu}_1 - \frac{1}{2} \vec{\mu}_0^T \Sigma^{-1} \vec{\mu}_0$$

Otsu's Method is related to Fisher's linear discriminant, and was created to binarize the histogram of pixels in a grayscale image by optimally picking the black/white threshold that minimizes intra-class variance and maximizes inter-class variance within/between grayscales assigned to black and white pixel classes.

4 Multiclass LDA

In the case where there are more than two classes, the analysis used in the derivation of the Fisher discriminant can be extended to find a subspace which appears to contain all of the class variability. This generalization is due to C.

R. Rao.[8] Suppose that each of C classes has a mean μ_i and the same covariance Σ . Then the scatter between class variability may be defined by the sample covariance of the class means

$$\Sigma_b = \frac{1}{C}\sum_{i=1}^{C}(\mu_i - \mu)(\mu_i - \mu)^T$$

where μ is the mean of the class means. The class separation in a direction $\vec{w}$ in this case will be given by

$$S = \frac{\vec{w}^T \Sigma_b \vec{w}}{\vec{w}^T \Sigma \vec{w}}$$

This means that when $\vec{w}$ is an eigenvector of $\Sigma^{-1}\Sigma_b$ the separation will be equal to the corresponding eigenvalue.

If $\Sigma^{-1}\Sigma_b$ is diagonalizable, the variability between features will be contained in the subspace spanned by the eigenvectors corresponding to the $C - 1$ largest eigenvalues (since Σ_b is of rank $C - 1$ at most). These eigenvectors are primarily used in feature reduction, as in PCA. The eigenvectors corresponding to the smaller eigenvalues will tend to be very sensitive to the exact choice of training data, and it is often necessary to use regularisation as described in the next section.

If classification is required, instead of dimension reduction, there are a number of alternative techniques available. For instance, the classes may be partitioned, and a standard Fisher discriminant or LDA used to classify each partition. A common example of this is "one against the rest" where the points from one class are put in one group, and everything else in the other, and then LDA applied. This will result in C classifiers, whose results are combined. Another common method is pairwise classification, where a new classifier is created for each pair of classes (giving $C(C - 1)/2$ classifiers in total), with the individual classifiers combined to produce a final classification.

5 Incremental LDA

The typical implementation of the LDA technique requires that all the samples are available in advance. However, there are situations where the entire data set is not available and the input data are observed as a stream. In this case, it is desirable for the LDA feature extraction to have the ability to update the computed LDA features by observing the new samples without running the algorithm on the whole data set. For example, in many real-time applications such as mobile robotics or on-line face recognition, it is important to update the extracted LDA features as soon as new observations are available. An LDA feature extraction technique that can update the LDA features by simply observing new samples is an *incremental LDA algorithm*, and this idea has been extensively studied over the last two decades.[9] Catterjee and Roychowdhury proposed an incremental self-organized LDA algorithm for updating the LDA features.[10] In other work, Demir and Ozmehmet proposed online local learning algorithms for updating LDA features incrementally using error-correcting and the Hebbian learning rules.[11] Later, Aliyari *et a*l. derived fast incremental algorithms to update the LDA features by observing the new samples.[9]

6 Practical use

In practice, the class means and covariances are not known. They can, however, be estimated from the training set. Either the maximum likelihood estimate or the maximum a posteriori estimate may be used in place of the exact value in the above equations. Although the estimates of the covariance may be considered optimal in some sense, this does not mean that the resulting discriminant obtained by substituting these values is optimal in any sense, even if the assumption of normally distributed classes is correct.

Another complication in applying LDA and Fisher's discriminant to real data occurs when the number of measurements of each sample exceeds the number of samples in each class.[4] In this case, the covariance estimates do not have full rank, and so cannot be inverted. There are a number of ways to deal with this. One is to use a pseudo inverse instead of the usual matrix inverse in the above formulae. However, better numeric stability may be achieved by first projecting the problem onto the subspace spanned by Σ_b .[12] Another strategy to deal with small sample size is to use a shrinkage estimator of the covariance matrix, which can be expressed mathematically as

$$\Sigma = (1 - \lambda)\Sigma + \lambda I$$

where I is the identity matrix, and λ is the *shrinkage intensity* or *regularisation parameter*. This leads to the framework of regularized discriminant analysis[13] or shrinkage discriminant analysis.[14]

Also, in many practical cases linear discriminants are not suitable. LDA and Fisher's discriminant can be extended for use in non-linear classification via the kernel trick. Here, the original observations are effectively mapped into a higher dimensional non-linear space. Linear classification in this non-linear space is then equivalent to non-linear classification in the original space. The most commonly used example of this is the kernel Fisher discriminant.

LDA can be generalized to multiple discriminant analysis, where c becomes a categorical variable with N possible states, instead of only two. Analogously, if the class-conditional densities $p(\vec{x}|c = i)$ are normal with shared covariances, the sufficient statistic for $P(c|\vec{x})$ are the values of N projections, which are the subspace spanned by the N means, affine projected by the inverse covariance matrix. These projections can be found by solving a generalized eigenvalue problem, where the numerator is the covariance matrix formed by treating the means as the samples, and the denominator is the shared covariance matrix.

7 Applications

In addition to the examples given below, LDA is applied in positioning and product management.

7.1 Bankruptcy prediction

In bankruptcy prediction based on accounting ratios and other financial variables, linear discriminant analysis was the first statistical method applied to systematically explain which firms entered bankruptcy vs. survived. Despite limitations including known nonconformance of accounting ratios to the normal distribution assumptions of LDA, Edward Altman's 1968 model is still a leading model in practical applications.

7.2 Face recognition

In computerised face recognition, each face is represented by a large number of pixel values. Linear discriminant analysis is primarily used here to reduce the number of features to a more manageable number before classification. Each of the new dimensions is a linear combination of pixel values, which form a template. The linear combinations obtained using Fisher's linear discriminant are called *Fisher faces*, while those obtained using the related principal component analysis are called *eigenfaces*.

7.3 Marketing

In marketing, discriminant analysis was once often used to determine the factors which distinguish different types of customers and/or products on the basis of surveys or other forms of collected data. Logistic regression or other methods are now more commonly used. The use of discriminant analysis in marketing can be described by the following steps:

1. Formulate the problem and gather data — Identify the salient attributes consumers use to evaluate products in this category — Use quantitative marketing research techniques (such as surveys) to collect data from a sample of potential customers concerning their ratings of all the product attributes. The data collection stage is usually done by marketing research professionals. Survey questions ask the respondent to rate a product from one to five (or 1 to 7, or 1 to 10) on a range of attributes chosen by the researcher. Anywhere from five to twenty attributes are chosen. They could include things like: ease of use, weight, accuracy, durability, colourfulness, price, or size. The attributes chosen will vary depending on the product being studied. The same question is asked about all the products in the study. The data for multiple products is codified and input into a statistical program such as R, SPSS or SAS. (This step is the same as in Factor analysis).

2. Estimate the Discriminant Function Coefficients and determine the statistical significance and validity — Choose the appropriate discriminant analysis method. The direct method involves estimating the discriminant function so that all the predictors are assessed simultaneously. The stepwise method enters the predictors sequentially. The two-group method should be used when the dependent variable has two categories or states. The multiple discriminant method is used when the dependent variable has three or more categorical states. Use Wilks's Lambda to test for significance in SPSS or F stat in SAS. The most common method used to test validity is to split the sample into an estimation or analysis sample, and a validation or holdout sample. The estimation sample is used in constructing the discriminant function. The validation sample is used to construct a classification matrix which contains the number of correctly classified and incorrectly classified cases. The percentage of correctly classified cases is called the hit ratio.

3. Plot the results on a two dimensional map, define the dimensions, and interpret the results. The statistical program (or a related module) will map the results. The map will plot each product (usually in two-dimensional space). The distance of products to each other indicate either how different they are. The dimensions must be labelled by the researcher. This requires subjective judgement and is often very challenging. See perceptual mapping.

7.4 Biomedical studies

The main application of discriminant analysis in medicine is the assessment of severity state of a patient and prognosis of disease outcome. For example, during retrospective analysis, patients are divided into groups according to severity of disease – mild, moderate and severe form. Then results of clinical and laboratory analyses are studied in order to reveal variables which are statistically different in studied groups. Using these variables, discriminant functions are built which help to objectively classify disease in a future patient into mild, moderate or severe form.

In biology, similar principles are used in order to classify and define groups of different biological objects, for example, to define phage types of Salmonella enteritidis based on Fourier transform infrared spectra,[15] to detect animal source of Escherichia coli studying its virulence factors[16] etc.

7.5 Earth Science

This method can be used to separate the alteration zones. For example, when different data from various zones are available, discriminate analysis can find the pattern within the data and classify it effectively.[17]

8 See also

- Data mining
- Decision tree learning
- Factor analysis
- Kernel Fisher discriminant analysis
- Logit (for logistic regression)
- Multidimensional scaling
- Pattern recognition
- Perceptron
- Preference regression
- Quadratic classifier

9 References

[1] Fisher, R. A. (1936). "The Use of Multiple Measurements in Taxonomic Problems". *Annals of Eugenics* **7** (2): 179–188. doi:10.1111/j.1469-1809.1936.tb02137.x. hdl:2440/15227.

[2] McLachlan, G. J. (2004). *Discriminant Analysis and Statistical Pattern Recognition*. Wiley Interscience. ISBN 0-471-69115-1. MR 1190469.

[3] Analyzing Quantitative Data: An Introduction for Social Researchers, Debra Wetcher-Hendricks, p.288

[4] Martinez, A. M.; Kak, A. C. (2001). "PCA versus LDA" (PDF). *IEEE Transactions on Pattern Analysis and Machine Intelligence* **23** (=2): 228–233. doi:10.1109/34.908974.

[5] Abdi, H. (2007) "Discriminant correspondence analysis." In: N.J. Salkind (Ed.): *Encyclopedia of Measurement and Statistic*. Thousand Oaks (CA): Sage. pp. 270–275.

[6] Perriere, G.; & Thioulouse, J. (2003). "Use of Correspondence Discriminant Analysis to predict the subcellular location of bacterial proteins", *Computer Methods and Programs in Biomedicine*, 70, 99–105.

[7] Venables, W. N.; Ripley, B. D. (2002). *Modern Applied Statistics with S* (4th ed.). Springer Verlag. ISBN 0-387-95457-0.

[8] Rao, R. C. (1948). "The utilization of multiple measurements in problems of biological classification". *Journal of the Royal Statistical Society, Series B* **10** (2): 159–203.

[9] Aliyari Ghassabeh, Youness; Rudzicz, Frank; Moghaddam, Hamid Abrishami (2015-06-01). "Fast incremental LDA feature extraction". *Pattern Recognition* **48** (6): 1999–2012. doi:10.1016/j.patcog.2014.12.012.

[10] Chatterjee, C.; Roychowdhury, V.P. (1997-05-01). "On self-organizing algorithms and networks for class-separability features". *IEEE Transactions on Neural Networks* **8** (3): 663–678. doi:10.1109/72.572105. ISSN 1045-9227.

[11] Demir, G. K.; Ozmehmet, K. (2005-03-01). "Online Local Learning Algorithms for Linear Discriminant Analysis". *Pattern Recogn. Lett.* **26** (4): 421–431. doi:10.1016/j.patrec.2004.08.005. ISSN 0167-8655.

[12] Yu, H.; Yang, J. (2001). "A direct LDA algorithm for high-dimensional data — with application to face recognition", *Pattern Recognition*, 34 (10), 2067–2069

[13] Friedman, J. H. (1989). "Regularized Discriminant Analysis" (PDF). *Journal of the American Statistical Association* (American Statistical Association) **84** (405): 165–175. doi:10.2307/2289860. JSTOR 2289860. MR 0999675.

[14] Ahdesmäki, M.; Strimmer K. (2010) "Feature selection in omics prediction problems using cat scores and false nondiscovery rate control", *Annals of Applied Statistics*, 4 (1), 503–519.

[15] Preisner O, Guiomar R, Machado J, Menezes JC, Lopes JA. Application of Fourier transform infrared spectroscopy and chemometrics for differentiation of Salmonella enterica serovar Enteritidis phage types. Appl Environ Microbiol. 2010;76(11):3538–3544.

[16] David DE, Lynne AM, Han J, Foley SL. Evaluation of virulence factor profiling in the characterization of veterinary Escherichia coli isolates. Appl Environ Microbiol. 2010;76(22):7509–7513.

[17] Tahmasebi, P., Hezarkhani, A., & Mortazavi, M. (2010). Application of discriminant analysis for alteration separation; sungun copper deposit, East Azerbaijan, Iran. Australian Journal of Basic and Applied Sciences, 6(4), 564-576.

10 Further reading

- Duda, R. O.; Hart, P. E.; Stork, D. H. (2000). *Pattern Classification* (2nd ed.). Wiley Interscience. ISBN 0-471-05669-3. MR 1802993.
- Hilbe, J. M. (2009). *Logistic Regression Models*. Chapman & Hall/CRC Press. ISBN 978-1-4200-7575-5.
- Mika, S.; et al. (1999). "Fisher Discriminant Analysis with Kernels". *IEEE Conference on Neural Networks for Signal Processing IX*: 41–48. doi:10.1109/NNSP.1999.788121.
- Mark Burdon and Paul Harpur, 'Re-Conceptualising Privacy and Discrimination in an Age of Talent Analytics' (2014) 37 University of New South Wales Law Journal, 2, 679-712.1
- Miranda Terry and Paul Harpur, 'The New Era of Segmenting Society on Ability Lines: Workplace Analytics and Disability Discrimination' (Society for Disability Studies, Atlanta USA, 10-13 June 2015).

11 External links

- ALGLIB contains open-source LDA implementation in C# / C++ / Pascal / VBA.
- Psychometrica.de open-source LDA implementation in Java
- LDA tutorial using MS Excel
- Biomedical statistics. Discriminant analysis

12 Text and image sources, contributors, and licenses

12.1 Text

- **Linear discriminant analysis** *Source:* https://en.wikipedia.org/wiki/Linear_discriminant_analysis?oldid=720193226 *Contributors:* The Anome, Fnielsen, XJaM, Edward, Michael Hardy, Kku, Den fjättrade ankan~enwiki, Hike395, Jihg, Deus~enwiki, Nonick, Giftlite, Duncharris, Dfrankow, Pgan002, 3mta3, Arcenciel, Forderud, Crackerbelly, RichardWeiss, Qwertyus, Rjwilmsi, Mathbot, Predictor, Adoniscik, YurikBot, Timholy, Doncram, Tcooke, Shawnc, SmackBot, Maksim-e~enwiki, Slashme, Mcld, Memming, Solarapex, Beetstra, Dicklyon, Lifeartist, StanfordProgrammer, Petrus Adamus, Sopoforic, Cydebot, Thijs!bot, AlexAlex, AnAj, Mack2, Cpl Syx, Stephenchou0722, Lwaldron, R'n'B, Bahram.zahir, G.kunter~enwiki, Nechamayaniger, Daviddoria, SieBot, Ivan Štambuk, Mverleg, Jonomillin, OKBot, Melcombe, Produit, Statone, Calimo, Qwfp, Addbot, Mabdul, AndrewHZ, Lightbot, Yobot, Citation bot, Klisanor, Sylwia Ufnalska, Morten Isaksen, Olg wiki, SchnitzelMannGreek, Pcoat, FrescoBot, X7q, Citation bot 1, Wkretzsch, Heavy Joke, Www wwwjs1, Jfmantis, EmausBot, Dewritech, Radshashi, Manyu aditya, Marion.cuny, Vldscore, WikiMSL, Helpful Pixie Bot, BG19bot, Solomon7968, CitationCleanerBot, Tbrknt, Khazar2, Illia Connell, Jerry Hintze, I am One of Many, Lcparra, Ashleyleia, ArtfulVampire, SJ Defender, Monkbot, BazzaHarp, Екатерина Конь, Degill, Olosko, HelpUsStopSpam, Mathematician1983 and Anonymous: 113

12.2 Images

- **File:R._A._Fischer.jpg** *Source:* https://upload.wikimedia.org/wikipedia/commons/4/46/R._A._Fischer.jpg *License:* Public domain *Contributors:* Transferred from en.wikipedia to Commons. *Original artist:* The original uploader was Bletchley at English Wikipedia

12.3 Content license

- Creative Commons Attribution-Share Alike 3.0