linear discriminant analysis (LDA)

Author and editors

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland (ORCID 0000-0001-7538-0990)
Editors: Konstantina Olioumtsevits (ORCID 0000-0003-2697-8659), Ekkehard Schnoor (ORCID 0000-0001-8252-1499), Aalto University, Espoo, Finland
*Corresponding author: alex.jung@aalto.fi

Abstract

Linear discriminant analysis (LDA) learns a linear feature transformation that replaces the feature vector of a data point by one feature, chosen so that the two label values of a binary classification problem are told apart as well as possible. The criterion is the Fisher criterion, the ratio of the between-class to the within-class scatter of the new feature. The direction attaining its maximum is obtained by solving one linear system, so LDA is a closed form rather than an iterative method. The two scatter matrices add up to the sample covariance matrix, the matrix whose quadratic form principal component analysis (PCA) maximizes, so PCA keeps the direction of largest total sample variance while LDA weighs the two parts against each other. A direction can carry nearly all of the sample variance and nearly none of the between-class part. With more than two label values, LDA delivers at most one direction fewer than there are values. The name also denotes the classifier that reaches the same direction from a multivariate normal distribution per label value with one shared covariance matrix.

Definition

A weather station at Krems an der Donau (Austria) records the morning minimum and the maximum daytime temperature of each day, and a forecaster wants one number per day that tells the days with precipitation from the dry ones. The two temperatures move together: a cold day is cold in the morning and in the afternoon, a warm day is warm in both. That shared swing is the direction along which the days vary most, and it is the direction principal component analysis (PCA) keeps. It is also close to useless for the question asked, because it says how warm a day was and not whether it rained. LDA produces one number per day as well, and it chooses the direction from the labels.

LDA is a feature learning method for binary classification. Consider a training set $\mathcal{D}^{(\text{train})} = \big\{(\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(m)}, y^{(m)})\big\}$ of sample size m, with feature vectors $\mathbf{x}^{(r)} \in \mathbb{R}^d$ and labels $y^{(r)} \in \{-1, +1\}$. LDA learns a linear feature transformation

$$\phi^{(\mathbf{w})} : \mathbb{R}^d \to \mathbb{R} : \mathbf{x} \mapsto \mathbf{w}^\top \mathbf{x},$$

which replaces the d raw features of a data point by the single new feature $\phi^{(\mathbf{w})}(\mathbf{x})$. The labels enter through two matrices. Let $\widehat{\boldsymbol{\mu}}^{(c)}$ be the sample mean

of the feature vectors carrying the label value c, and p_c the fraction of the training set that carries it. The sample mean of all feature vectors is then $\widehat{\boldsymbol{\mu}} = \sum_{c\in\{-1,+1\}} p_c \widehat{\boldsymbol{\mu}}^{(c)}$. The two matrices are (Hastie et al., 2009, Sect. 4.3.3):

$$\mathbf{S} := \frac{1}{m} \sum_{c\in\{-1,+1\}} \sum_{r\,:\, y^{(r)}=c} \big(\mathbf{x}^{(r)} - \widehat{\boldsymbol{\mu}}^{(c)}\big)\big(\mathbf{x}^{(r)} - \widehat{\boldsymbol{\mu}}^{(c)}\big)^\top, \tag{1}$$

$$\mathbf{B} := \sum_{c\in\{-1,+1\}} p_c \big(\widehat{\boldsymbol{\mu}}^{(c)} - \widehat{\boldsymbol{\mu}}\big)\big(\widehat{\boldsymbol{\mu}}^{(c)} - \widehat{\boldsymbol{\mu}}\big)^\top. \tag{2}$$

The within-class scatter matrix $\mathbf{S}$ averages the spread of the feature vectors around the sample mean of their own label value, and the between-class scatter matrix $\mathbf{B}$ measures how far the two sample means lie apart. Along a direction $\mathbf{w}$, the new feature has within-class sample variance $\mathbf{w}^\top\mathbf{S}\mathbf{w}$ and between-class sample variance $\mathbf{w}^\top\mathbf{B}\mathbf{w}$.

The criterion LDA maximizes is the ratio of the two, the Fisher criterion (Bishop, 2006, Sect. 4.1.4, Eq. (4.26); Fisher, 1936; Hastie et al., 2009, Sect. 4.3.3, Eq. (4.15)),

$$\max_{\mathbf{w}\in\mathbb{R}^d\,:\,\|\mathbf{w}\|=1} \frac{\mathbf{w}^\top\mathbf{B}\mathbf{w}}{\mathbf{w}^\top\mathbf{S}\mathbf{w}}. \tag{3}$$

A direction scores high when the two sample means are far apart on it and the days of one label value stay close together on it. The ratio of the two sample variances in the Fisher criterion (3) does not change when $\mathbf{w}$ is rescaled, so the constraint costs nothing and what is learned is a direction. For an invertible $\mathbf{S}$, the maximizing direction is that of

$$\widehat{\mathbf{w}} = \mathbf{S}^{-1}\big(\widehat{\boldsymbol{\mu}}^{(+1)} - \widehat{\boldsymbol{\mu}}^{(-1)}\big) \tag{4}$$

(Bishop, 2006, Sect. 4.1.4, Eq. (4.30)). The LDA direction (4) is computed in closed form, by solving one linear system: LDA is not an iterative method, so no operator has its output as a fixed point.

The name also denotes a classifier, not only the feature transformation. Modeling each label value by a multivariate normal distribution with its own mean and one shared covariance matrix, and estimating these model parameters by maximum likelihood, gives a linear classifier whose decision boundary is a hyperplane with the normal vector (4) (Hastie et al., 2009, Sect. 4.3). The direction is the same, and the probabilistic model supplies the threshold that the Fisher criterion leaves open.

The difference from PCA is one matrix. The within-class scatter matrix (1) and the between-class scatter matrix (2) add up to the sample covariance matrix (Hastie et al., 2009, Sect. 4.3.3),

$$\mathbf{S} + \mathbf{B} = \widehat{\mathbf{C}}. \tag{5}$$

The labels have disappeared from the sum (5). PCA maximizes $\mathbf{w}^\top\widehat{\mathbf{C}}\mathbf{w}$ over the directions of unit norm, which is the total sample variance of the new feature

and says nothing about which part of the sum (5) it comes from. The Fisher criterion (3) weighs the two parts against each other instead. A direction can therefore carry nearly all of the total sample variance and nearly none of the between-class part, which is what the opening example describes.

That example is measurable. The training set holds the 1461 days of 2020 to 2023 at Krems, each with the morning minimum and the maximum daytime temperature as its two features and the label $+1$ when precipitation was recorded. The eigenvector of $\widehat{\mathbf{C}}$ belonging to its largest eigenvalue carries 93 percent of the total sample variance and reaches a Fisher criterion of 0.00034, while the LDA direction (4) reaches 0.103, larger by a factor of 303. The two directions are 89 degrees apart, so what the labels depend on lies almost exactly along the direction PCA ranks last (Fig. 1). The LDA direction weighs the two features with opposite signs, $(0.817, -0.577)$, so its new feature is large on a day whose temperature changed little between morning and afternoon, which is what cloud cover does.

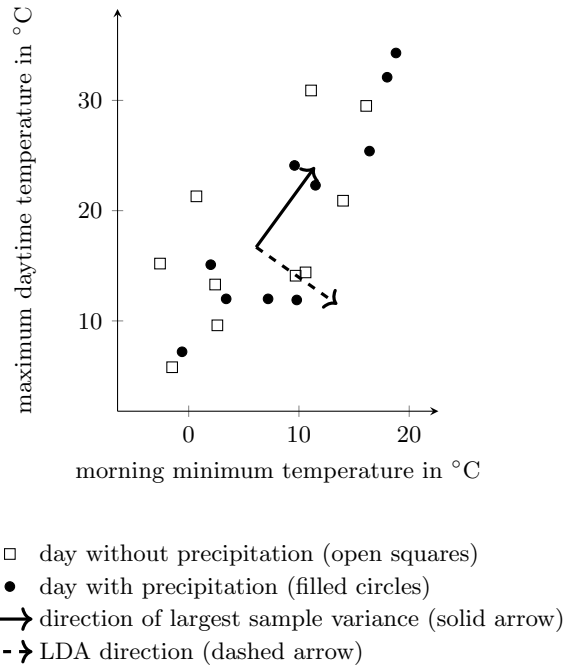


Figure 1: Twenty of the 366 days of 2024 at Krems an der Donau (Austria), ten with and ten without recorded precipitation, drawn by taking every 14th of the 144 wet days and every 22nd of the 222 dry days. Each day is a data point with its morning minimum and its maximum daytime temperature as the two features, drawn as a filled circle when precipitation was recorded and as an open square otherwise. The two arrows start at the sample mean of the training set and are drawn equally long, since only their directions are learned. The solid arrow is the direction of largest sample variance and the dashed arrow the direction along which the two kinds of day are told apart best. Data generated by `pythondemos/lda.py`

The criterion (3) has one maximum over the directions of the plane and one minimum, and Fig. 2 shows where the two directions sit on it. Thresholding the single LDA feature at the value learned on the 1461 training days classifies 63.7 percent of the 366 days of 2024, kept aside as the test set, correctly. The single PCA feature reaches 60.9 percent and always predicting the majority label value 60.7 percent. Two temperatures of one day say little about rain, and the little they say lies in the direction PCA discards.

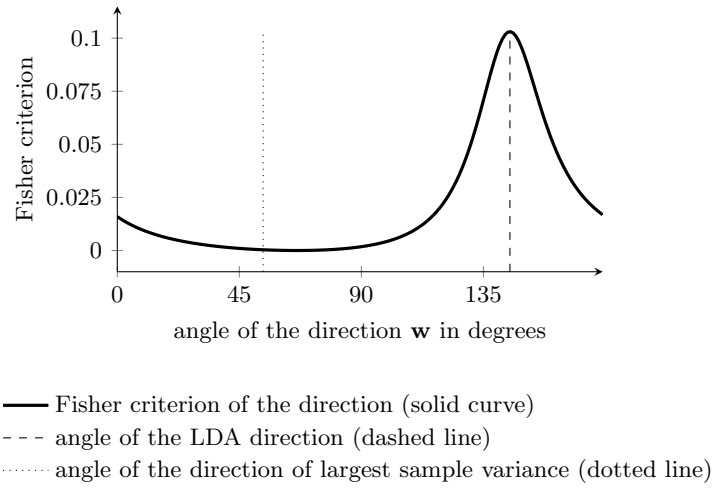


Figure 2: The Fisher criterion (3) of every direction of the plane, for the two temperature features of the 1461 days of 2020 to 2023 at Krems. A direction and its opposite give the same value, so the angle runs over half a turn. The dashed line marks the angle of the LDA direction, where the solid curve attains its maximum, and the dotted line the angle of the direction of largest sample variance, which sits near the minimum. Data generated by `pythondemos/lda.py`

Three limits come with the closed form (4). It needs an invertible $\mathbf{S}$, which fails when the training set holds fewer data points than features. Shrinking $\mathbf{S}$ towards a multiple of the identity matrix, which is regularization by a penalty, is the usual repair (Hastie et al., 2009, Sect. 4.3.1, Eq. (4.14)). With more than two label values, the between-class scatter matrix is built from all the sample means and has rank at most one below their number, so LDA delivers at most that many directions, however large d is (Hastie et al., 2009, Sect. 4.3.3). And the single new feature is only as good as a linear combination can be: two labels separated by a curved decision boundary overlap on every direction (Hastie et al., 2009, Sect. 12.4).

Where a linear combination is not enough, the answer is to run LDA on an enlarged feature vector that also holds the squares and the products of the raw features. A hyperplane in the feature space of the enlarged feature vector is a curved decision boundary in the feature space of the raw features (Hastie et al., 2009, Sect. 12.5), and a kernel method is the general form of the step. What LDA buys for the cost of solving one linear system is the factor of 303 in the Fisher criterion (3). Of the two single features at Krems, the one chosen from the labels gains 3.0 percentage points on predicting the majority label value and the PCA feature gains 0.3.

Cross References

PCA, feature learning, feature transformation, dimensionality reduction, binary classification, linear classifier, sample covariance matrix, decision boundary, kernel method, logistic regression

References

1. Bishop (2006). Pattern Recognition and Machine Learning. Springer Science+Business Media, New York, NY, USA.

2. Fisher (1936). The Use of Multiple Measurements in Taxonomic Problems. Ann. Eugenics, 7(2), pp. 179–188.

3. Hastie et al. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed., Springer Science+Business Media, New York, NY, USA.