LDA

CS 530
Chapman

Spring 2020

1

TABLE OF CONTENTS

2

LINEAR DISCRIMINANT ANALYSIS (LDA)

3

Intro to Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) models each category (class) as a *multivariate Gaussian distribution* with equal variances across classes and uses Bayes' Theorem to estimate probabilities.
Fischer's linear discriminant is a variety of LDA that uses projection to reduce dimensionality and aid in separating categories.

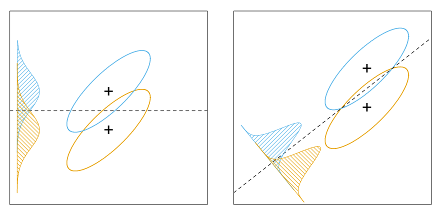


Image Source

4

Gaussian Distribution

The *Gaussian distribution* (a.k.a. normal distribution, bell curve) shows up frequently in various settings and datasets (due to the "central-limit theorem"):

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Here μ is the mean of the Gaussian and σ is its standard deviation (a measure of spread, the square root of the variance).

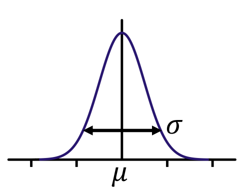


Image Source

5

Gaussian Distribution

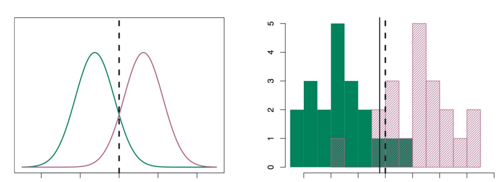


Image Source

For example, for the distribution on the right, LDA models the green and burgundy classes as 2 Gaussians and puts a decision boundary or classification boundary (dashed line) based on which category is more probable. In this case, the boundary computed by LDA (dashed line) is close to the optimal one (solid line), on the right.

6

Bayes' Theorem

Bayes' Theorem:

$$P(A|B) = \frac{P(A \wedge B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

7

Bayes' Theorem

Bayes' Theorem:

Likelihood Prior

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Posterior Marginal

8

Bayes' Theorem

Bayes' Theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Example: How accurate is the expression "where there's smoke there's fire"?

- We want to compute $P(\boldsymbol{Fire}|\boldsymbol{Smoke})$
- $P(F|S) = \frac{P(S|F)P(F)}{P(S)}$
- If
  - $P(S|F) = 0.9$ (likelihood)
  - $P(F) = 0.01$ (prior)
  - $P(S) = 0.1$ (marginal)
- then $P(F|S) = \frac{0.9 \cdot 0.01}{0.1} = 0.09$
- Under these probabilities, maybe check for fire when you see smoke

9

Bayes' Theorem and LDA

Bayes' Theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayes' Theorem for LDA:

$$\Pr(Y = k|X = x) = \frac{\Pr(X = x|Y = k)\Pr(Y = k)}{\sum_{i=1}^{K} \Pr(X = x|Y = i)\Pr(Y = i)} = \frac{f_k(x)\pi_k}{\sum_{i=1}^{K} f_i(x)\pi_i},$$

where

$$f_k(x) = \Pr(X = x|Y = k)$$

and

$$\pi_k = \Pr(Y = k)$$

10

Bayes' Theorem and LDA

In words:

The probability that label Y designates some specific class k given that data X is some specific value x is the same as:

The the probability of X being x given Y is the specific class k

times

The prior probability that any observation is from class k

divided by

a normalization factor (that is not dependent on, or the same for, every k).

11

Bayes' Theorem and LDA

For LDA we assume that each class k has a Gaussian distribution,

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x - \mu_k)^2\right).$$

Bayes' Theorem then tells us that the probability for each class k at a given point x is

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_k)^2\right)}{\sum_{i=1}^{K} \pi_i \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_i)^2\right)}$$

where we assume all classes have the same variance, σ^2 (i.e., ${\sigma_k}^2 = \sigma^2 \;\forall k$). In other words, $X \sim N(\mu_k, \sigma^2)$ for every k.

12

Bayes' Theorem and LDA

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_l)^2\right)}$$

We want to assign observation x to the class k for which $p_k(x)$ is largest.

Taking the log and rearranging, we can show that this is equivalent to assigning x to class k for which $\delta_k(x)$ is largest, where

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{{\mu_k}^2}{2\sigma^2} + \log(\pi_k).$$

13

Bayes' Theorem and LDA

The best estimates of these quantities from the data are:

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

$$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$$

$$\hat{\pi}_k = n_k / n$$

Hence, the class means estimates, $\hat{\mu}_k$, are sample means for each class. The overall probability estimate for each class, $\hat{\pi}_k$, is the number of elements of each class divided by the total number of observations.

14

Bayes' Theorem and LDA

This means that the LDA classifier would assign x to the class for which

$$\hat{\delta}_k(x) = x \cdot \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{{\hat{\mu}_k}^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k)$$

is largest.

Note that $\hat{\delta}_k$ is a linear function of x, hence the name *linear* discriminant analysis, or LDA.

15

Bayes' Theorem and LDA

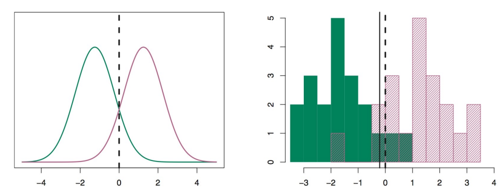


Image Source

For $K = 2$ (2 classes) and 1 dimension, with class means μ_1 and μ_2 and identical variance σ^2, and assuming $\pi_1 = \pi_2 = \frac{1}{2}$, it can be shown that the decision boundary (dashed line) is at the point where

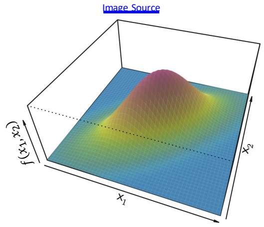
$$x = \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)} = \frac{\mu_1 + \mu_2}{2}.$$

16

Multivariate Gaussian and LDA

We have so far discussed the case of one-dimensional data (a single feature). For multiple dimensions, LDA assumes a multivariate Gaussian distribution. So, $X \sim N(\vec{\mu}_k, \overleftrightarrow{\Sigma})$ for some k. The math becomes more complicated. But the idea remains the same.

Image Source



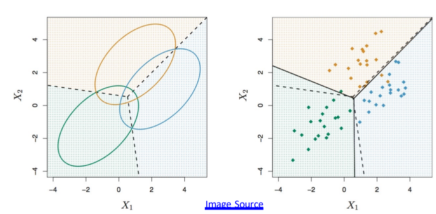
For a covariance matrix, $\overleftrightarrow{\Sigma} \in \mathbb{R}^{p \times p}$, and means vector, $\vec{\mu} \in \mathbb{R}^p$, the multivariate Gaussian function is:

$$f(\vec{x}) = \frac{1}{2\pi^{p/2} |\overleftrightarrow{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\vec{x} - \vec{\mu})^T \overleftrightarrow{\Sigma}^{-1} (\vec{x} - \vec{\mu})\right)$$

And when, for data $\vec{x}$, this function is greatest for class k, LDA predicts the label k for that $\vec{x}$.

17

Multivariate Gaussian and LDA



With 2 features and 3 classes, LDA models the distributions for each class as separate multivariate Gaussian functions, with the 95% probabilities indicated by the ovals on the left. The dashed lines are the optimal probabilities (which we know in this simulated example because we know the generative model). The LDA model determines decision boundaries, indicated by the solid lines on the right.

18

Multivariate Gaussian and LDA

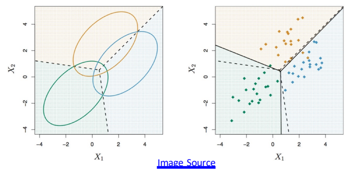


Image Source

Now the LDA classifier would assign $\vec{x}$ to the class for which

$$\hat{\delta}_k(\vec{x}) = \vec{x}^T \cdot \hat{\Sigma}^{-1} \cdot \hat{\vec{\mu}}_k - \frac{1}{2}\hat{\vec{\mu}}_k^{\,T} \cdot \hat{\Sigma}^{-1} \cdot \hat{\vec{\mu}}_k + \log(\hat{\pi}_k)$$

is largest. Here $\hat{\pi}_k$, $\hat{\vec{\mu}}_k$, and $\hat{\Sigma}$ are the estimators for the prior of class k, the means of Gaussian k, and the joint covariance matrix, respectively.

19

Multivariate Gaussian and LDA

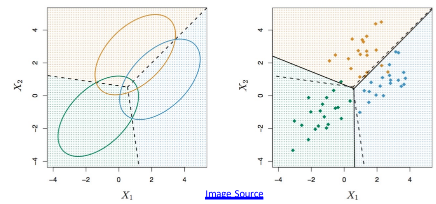


Image Source

For the multivariate case too Σ *is assumed to be the same across all the classes*. Hence, the ellipses (95% confidence intervals of the Gaussians) are the same sizes above, just with different means $\vec{\mu}_1, \vec{\mu}_2, \vec{\mu}_3$. And the separator between the classes is always a hyperplane in $\mathbb{R}^p$ (or line in $\mathbb{R}^2$ here).

20

LDA and QDA

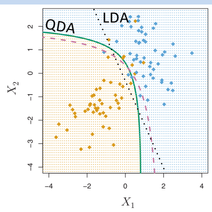


Image Source

What about cases where the assumption that $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_K$ is not merited? There is an extension of LDA that does not assume equal covariance matrices. The solution is then quadratic in $\vec{x}$. Hence, the algorithm is termed Quadratic Discriminant Analysis, or QDA.

21

Fisher's Linear Discriminant

Fisher's Linear Discriminant is a variant of LDA (actually, the original article on an LDA-like methods that Fisher published in 1936) that uses projections, in a similar manner to PCA (but supervised), to aid in classification. The data is projected to an optimal hyperplane that aims to simulatneously

- maximize between-class variance
- minimize within-class variance.

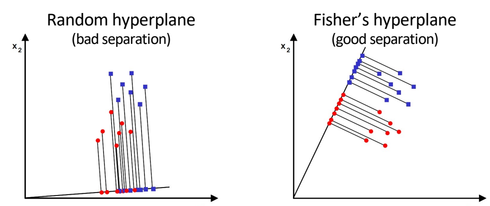


Image Source

22

Fisher's Linear Discriminant

We want to project our samples, $\vec{x}$, onto a line (or, generally, hyperplane) with normal vector $\vec{w}$ (assume $\vec{x} \in \mathbb{R}^2$ for visualization):

$$y = \vec{w}^T\vec{x}$$

Of all possible lines—i.e., all possible $\vec{w}$'s, we select the one that maximizes separability of the y's.

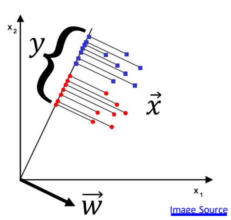


Image Source

23

Fisher's Linear Discriminant

To find a good projection, we need a measure of separation

The mean μ_i for each class i is

$$\vec{\mu}_i = \frac{1}{N_i}\sum_{\vec{x} \in \text{class } i} \vec{x}$$

The mean $\tilde{\mu}_i$ for each class i as projected onto the line (generally a hyperplane) is

$$\tilde{\mu}_i = \frac{1}{N_i}\sum_{y \in \text{class } i} y = \frac{1}{N_i}\sum_{x \in \text{class } i} \vec{w}^T\vec{x} = \vec{w}^T\vec{\mu}_i$$

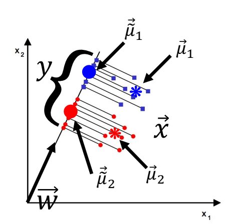


Image Source

24

Fisher's Linear Discriminant

Define the scatter within each class as

$$\tilde{s}_i^2 = \sum_{y \in \text{class } i} (y - \tilde{\mu}_i)^2 .$$

Then the overall *within-class scatter* is $(\tilde{s}_1^2 + \tilde{s}_2^2)$. Fisher's linear discriminant maximize the following function:

$$J(\vec{w}) = \frac{|\tilde{\mu}_1 - \tilde{\mu}_2|^2}{s_1^2 + s_2^2}.$$

In other words, the method tries to minimize the distance between samples within the same class and, at the same time, maximize the distance between the projected means.

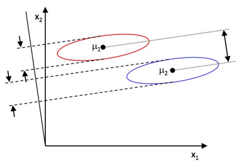


Image Source

25

Fisher's Linear Discriminant

Define S_B as the between-class scatter matrix:

$$\overleftrightarrow{S}_B = (\vec{\mu}_2 - \vec{\mu}_1)(\vec{\mu}_2 - \vec{\mu}_1)^T$$

and define S_w as the within-class scatter matrix:

$$\overleftrightarrow{S}_w = \sum_{i \in \text{class } 1} (\vec{x}_i - \vec{\mu}_1)(\vec{x}_i - \vec{\mu}_1)^T + \sum_{i \in \text{class } 2} (\vec{x}_i - \vec{\mu}_2)(\vec{x}_i - \vec{\mu}_2)^T$$

Then (with some math) we can rewrite the function J, which we want to maximize, as

$$J(\vec{w}) = \frac{\vec{w}^T \overleftrightarrow{S}_B \vec{w}}{\vec{w}^T \overleftrightarrow{S}_W \vec{w}}$$

This is maximized when

$$\vec{w}^* = \overleftrightarrow{S}_W^{-1}(\vec{\mu}_2 - \vec{\mu}_1).$$

26

Fisher's Linear Discriminant

This formulation generalizes well to multi-class classification problems.

Also, this original derivation does not assume Gaussian distributions. Nor does it assume $\overleftrightarrow{\Sigma}_1 = \overleftrightarrow{\Sigma}_2 = \cdots = \overleftrightarrow{\Sigma}_K$.

However, if all LDA assumptions are met, it can be shown to be equivalent to LDA.

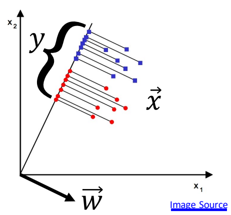


Image Source

27

LDA and PCA

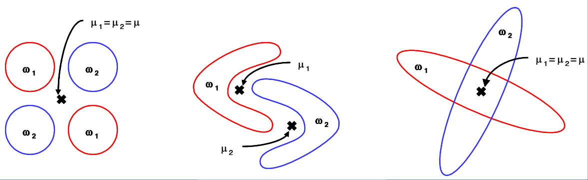
- Dimensionality reduction using LDA finds the subspace where the within-class variability (of the samples) is minimized while the between class variability (of the means) is maximized
- Dimensionality reduction using PCA find the subspace where the overall variability is maximized

28

Disadvantages of Linear Discriminant Analysis

LDA's dimensionality reduction may lead to more interpretable data and data that is easier to visualize. However,

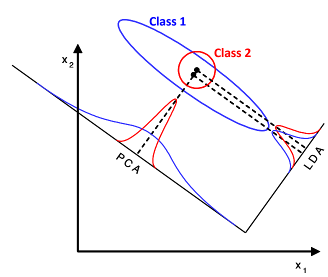
- LDA uses labels, so can lead to double dipping
  - Example (of what not to do): LDA on entire dataset (train & test) as classification preprocessing step
- For C classes, LDA produces at most $C - 1$ feature projections. If more features are needed for good classification, LDA cannot be used
- LDA assumes Gaussian (or at least Gaussian-like) likelihoods



29

Disadvantages of Linear Discriminant Analysis

- LDA assumes that all classes have the same variance. If this is not the case, and especially if the discriminatory information is in the variance and not the mean, LDA may fail.



30

LDA Code Example

```
>>> import numpy as np
>>> from sklearn.lda import LDA
>>> X = np.array([[-1, -1], [-2, -1], [-3, -2], [1, 1], [2, 1], [3, 2]])
>>> y = np.array([1, 1, 1, 2, 2, 2])
>>> clf = LDA()
>>> clf.fit(X, y)
LDA(n_components=None, priors=None, shrinkage=None, solver='svd',
  store_covariance=False, tol=0.0001)
>>> print(clf.predict([[-0.8, -1]]))
[1]
```

Source

31

CONCLUSIONS

- LDA is a supervised classification method that imposes equi-variance Gaussians on all classes and then derives optimal separation based on Bayes.
- Fisher Discriminant Analysis (LDA) is a supervised method that maximizes between class variability while minimizing within class variability. It does not require that each class be a Gaussian or equal variance but collapses to LDA if they are.

32

Take 🏠 message for rest of course: PCA & LDA

- Dimensionality reduction methods attempt to simplify the data by reducing its number of features
- Preprocessing: they can reduce classification error
  - PCA: class differentiability along maximal variance direction
  - LDA: when classes are (at least roughly) Gaussian
- PCA & LDA can improve data visualizability

33