Unmasking the Classifiers: Naive Bayes and LDA

Team Noble: Amanda Adams, Grace Tugado, Elizabeth Esquivel, Jack Rodrigue, Wai Yuen (Wylliam) Zheng

01.

Naive Bayes

Algorithm, Types of NB, and
Example Application

Bayes- Conditional Probability

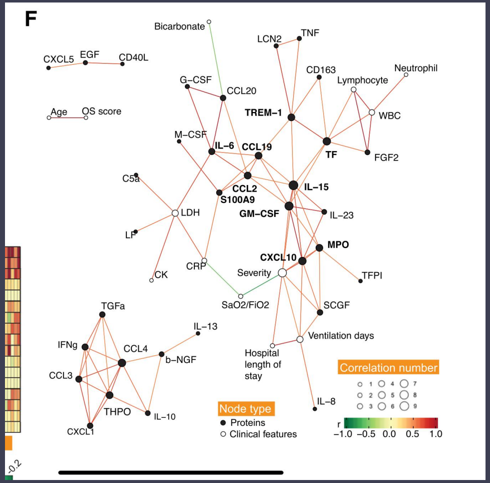
$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$$

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

Bayesian Algorithm

- **Start with a prior (initial assumption)** *## $P(A)$ where A is a target variable*
- **Integrate data / evidence** *## $P(B)$ where B is a input variable*
- **Update posterior probability (new beliefs)** *## $P(A|B)$*

  ## "training" [one-shot only, no iterative training]
- **Ready for prediction**



Put uncertainty into account

- Probability spectrum computed from the Bayesian algorithm summarize confident interval of each prediction

Naive Bayes

- Methods of supervised learning algorithms, based on Bayes Theorem, with a “Naive” Assumption
- Assumes conditional independence
- Why is it Naive?
- Different types of classifiers, but they all use Bayes Theorem
  - Bernoulli
  - Gaussian
  - Multinomial
- What is the difference between these?
  - Input variable types

Naive Bayes Classifiers

Bernoulli

- Focuses solely on (1) presence/true or (0) absence/false
- Ex: a spam filter

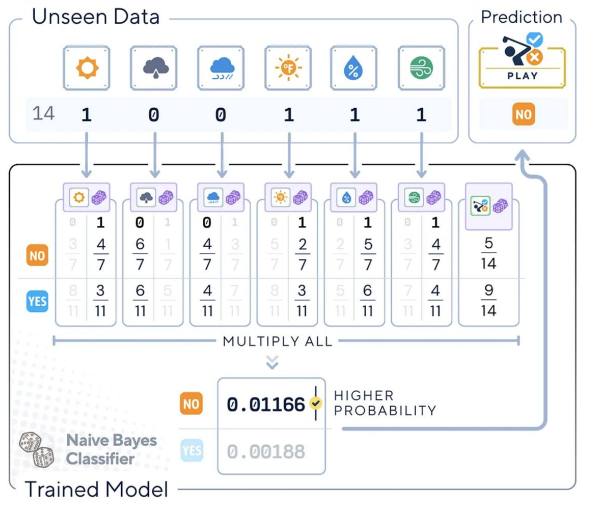
Multinomial

- This classifier counts the number of occurrences instead of the word itself
- Ex: a better spam classifier

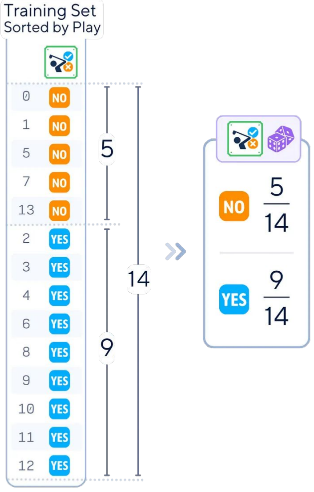
Gaussian

- The classifier assumes that the input follows a Gaussian Distribution
- Ex: IQ

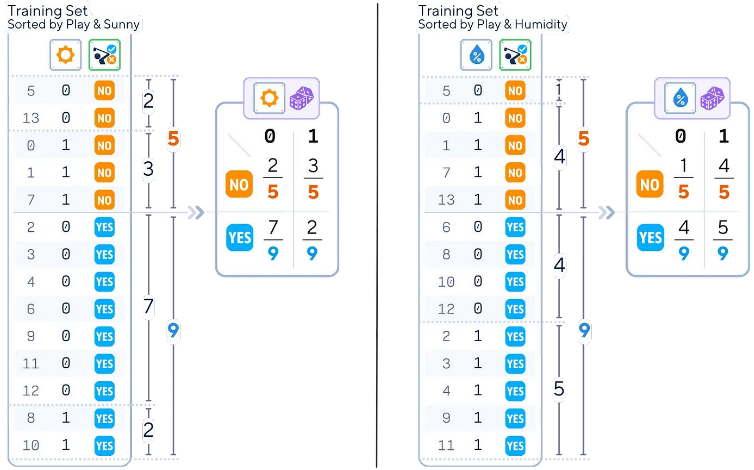
How does it work using Bayes?



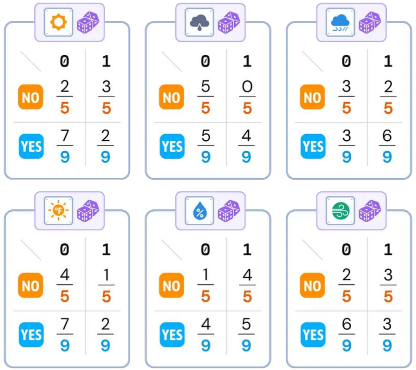
For our golf dataset, a Bernoulli NB classifier look at the probability of each feature happening for each class (YES & NO) then make decision based on which class has higher chance.



Target variable probability



Conditional variable probability

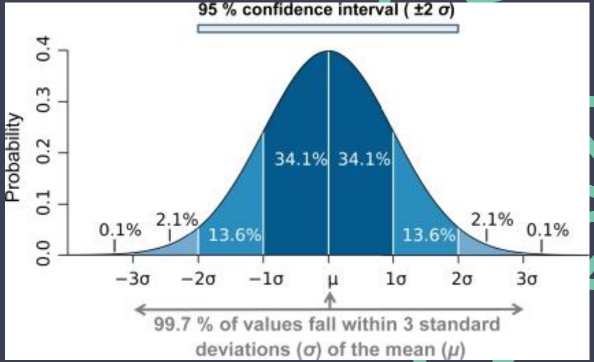


02.

Linear Discriminant Analysis

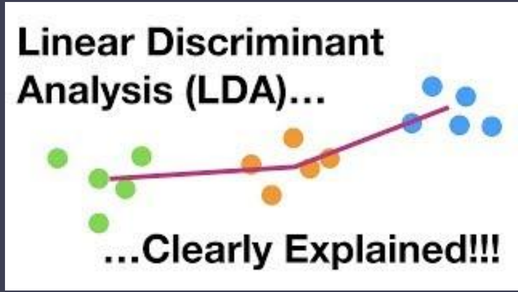
Overview, Strengths & Weaknesses of LDA, and its Applications

Objective: categorize observations into predefined classes based on input features.



Used for large datasets with many different features.

- Assumes normal (Gaussian) distribution within each class
- Aims to find a decision boundary that best separates these classes based on maximizing the ratio of between-class variance to within-class variance.
- Used to classify observations by finding a linear combination of features that best separates two or more classes.

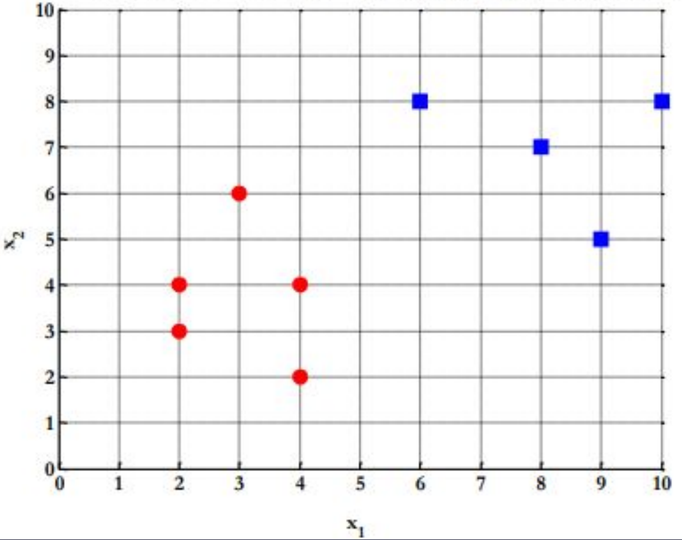


Main Steps and Equations for LDA (will add pics later)

1. Compute the mean vectors for the different classes
2. Compute the scatter matrices (in between class & within class)
   a. Sw (within class scatter matrix); Si = scatter matrix for a specific class
   b. Mi = mean vector for that class
   c. Sb = between-class scatter matrix
3. Compute the eigenvectors and eigenvalues for the scatter matrices
4. Sort the eigenvectors by decreasing eigenvalues & choose k eigenvectors with the largest eigenvalues (k = number of dimensions)
5. Use eigenvector matrix to transform the samples onto the new subspace
   a. Y = X * W
   b. W = d x k-dim eigenvector matrix

Integrated LDA Example

- Compute the Linear Discriminant projection for the following two-dimensional dataset.
  - Samples for class ω_1 : $\mathbf{X}_1=(x_1,x_2)=\{(4,2),(2,4),(2,3),(3,6),(4,4)\}$
  - Sample for class ω_2 : $\mathbf{X}_2=(x_1,x_2)=\{(9,10),(6,8),(9,5),(8,7),(10,8)\}$




```
% samples for class 1
X1 = [4,2;
      2,4;
      2,3;
      3,6;
      4,4];

% samples for class 2
X2 = [9,10;
      6,8;
      9,5;
      8,7;
      10,8];
```

Compute the...

1. Mean vectors for the different classes
1. Covariance matrices
2. Within-class scatter matrix Sw (sum of #2)
3. Between-class scatter matrix
4. LDA projection = soln of generalized eigenvalue problem

Detailed example: https://www.sci.utah.edu/~shireen/pdfs/tutorials/Elhabian_LDA09.pdf

Example: Results from 3-Class LDA

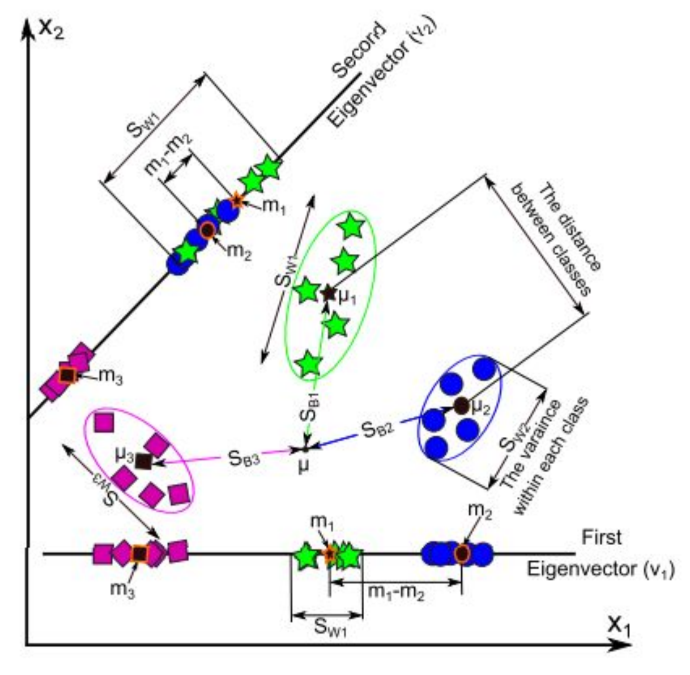


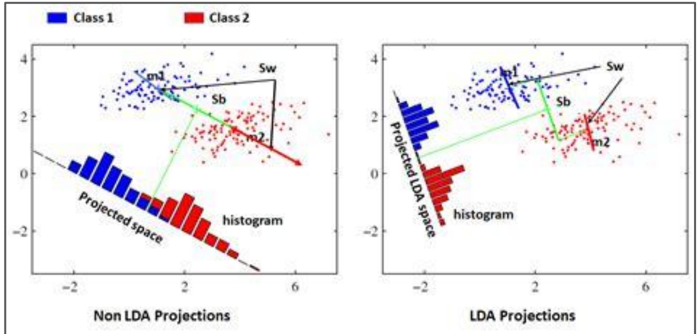
Fig. 3. A visualized comparison between the two lower-dimensional sub-spaces which are calculated using three different classes.

Tharwat et al. 2017

Strengths vs Weaknesses of LDA

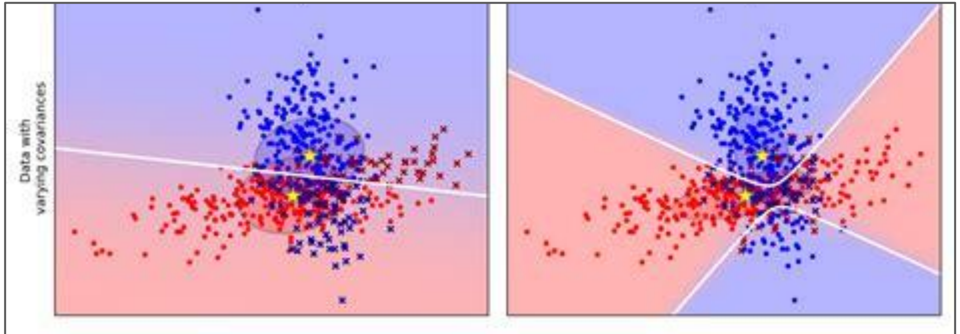
Strengths

- Efficient with multiple classification problems with well-separated classes
- Feature reduction → projecting data onto lower dimensional space
  - Improve computational efficiency and reduce noise
- Compatible for linearly separable data
  - Maximize ratio of between-class variance to within-class var



Weaknesses

- Fails to find the lower dimensional space if dimensions >> number of samples in data matrix
- The Small Sample Size (SSS)
  - Means of the 2 classes are equal → Sb and W = 0 (LDA space can't be found)
- Linearity problems
  - Classes are non-linearly separable → LDA can't discriminante between classes
  - Solution! Kernel functions



Practical Applications of LDA

- Biometrics with facial recognition
  - 160 images of 10 people → different lighting, expressions and eye-wear
  - Error rates → “leaving-one-out” strategy
  - Last image classification → nearest-neighbors classifier
- Fisherface (LDA) error rate gave lowest error rate

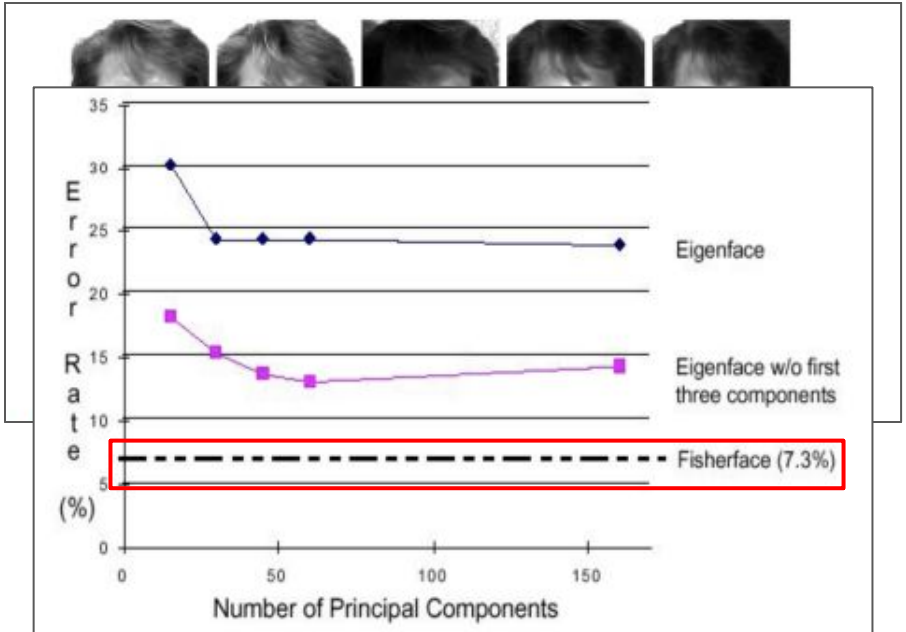


Figure 11: Eigenface vs. Fisherface. Source: Lecture 13, slide 61

Practical Applications of LDA

- Medical applications
  - DNA microarray gene expression datasets → determine expression levels of thousands of genes simultaneously
  - Cancer classification with gradient LDA (avoids SSS problem) → gene exp datasets: acute leukemia, small round blue-cell tumour and lung adenocarcinoma

Table 3. Misclassification on lung adenocarcinoma dataset

Method	Number of misclassified sampl
HC-TSP [24]	9
k-NN [24]	8
HC-*k*-TSP [24]	7
Decision Trees [24]	7
Naïve Bayes [24]	6
1-vs-1-SVM [24]	4
GLDA	2

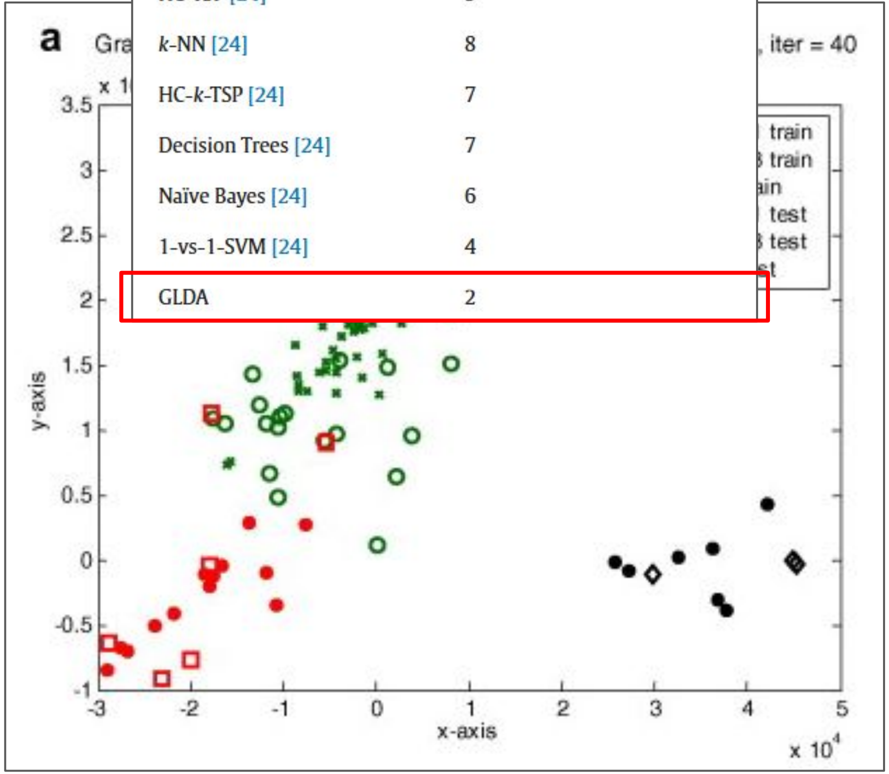


Figure A: Sharma et al. 2008 Cancer classification by gradient LDA technique using microarray gene expression data

03.

Comparison of NB and LDA

Comparing Naive Bayes and Linear Discriminant Analysis

- Probability: Posterior Vs Prior
- Feature Independence
- High Dimensionality
- Training Simplicity

Comparing Naive Bayes and Linear Discriminant Analysis-cont

Which should we use? Why?

Thanks!

Do you have any questions?

CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, and infographics & images by Freepik.

References

F. -J. Yang, "An Implementation of Naive Bayes Classifier," 2018 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 2018, pp. 301-306, doi: 10.1109/CSCI46756.2018.00065.

https://www.youtube.com/watch?app=desktop&v=azXCzI57Yfc (LDA video)

https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/gaussian-distribution (Normal distribution image)

https://www.ibm.com/topics/naive-bayes

https://medium.com/@mansi89mahi/simple-explanation-difference-between-naive-bayes-and-full-bayesian-network-model-505616545503#:~:text=In%20summary%2C%20the%20main%20difference,modeling%20of%20dependencies%20among%20variables.

https://medium.com/@kashishdafe0410/gaussian-naive-bayes-understanding-the-basics-and-applications-52098087b963

https://medium.com/@gridflowai/part-2-dive-into-bernoulli-naive-bayes-d0cbcbabb775

https://towardsdatascience.com/conditional-independence-the-backbone-of-bayesian-networks-85710f1b35b

https://www.researchgate.net/publication/316994943_Linear_discriminant_analysis_A_detailed_tutorial (LDA comprehensive review)

https://towardsdatascience.com/probabilistic-linear-discriminant-analysis-plda-explained-253b5effb96?gi=c65eb8d4c2c6 (pros of LDA image)