High Dimensional Data Classification

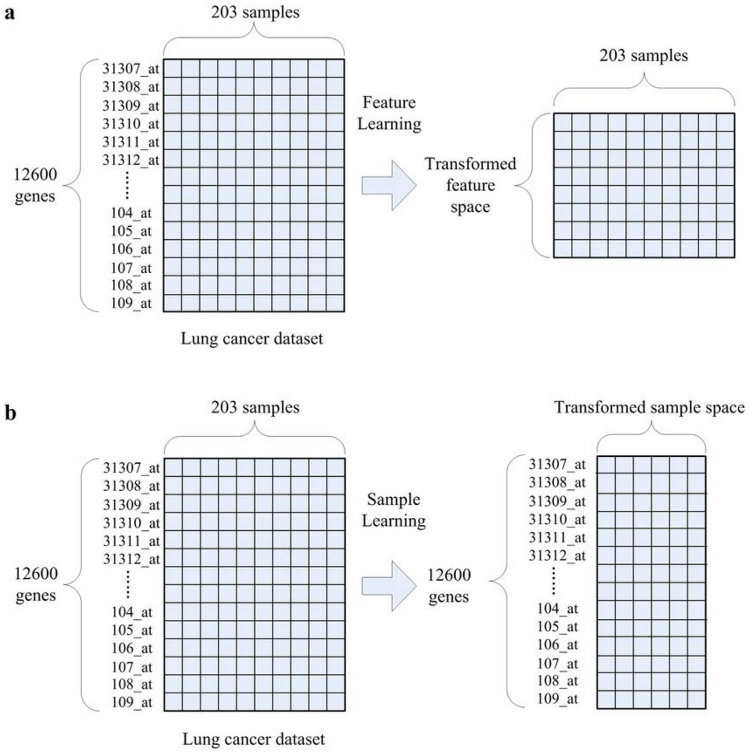
Mentee: Elvin Liu
Mentor: Zhaoxing Wu

Overview

1. **Introduction**
2. **Background** (Fisher’s Linear Discriminant Analysis)
3. **Solution** (Penalized Discriminant Analysis Projection Pursuit)
4. **Analysis**

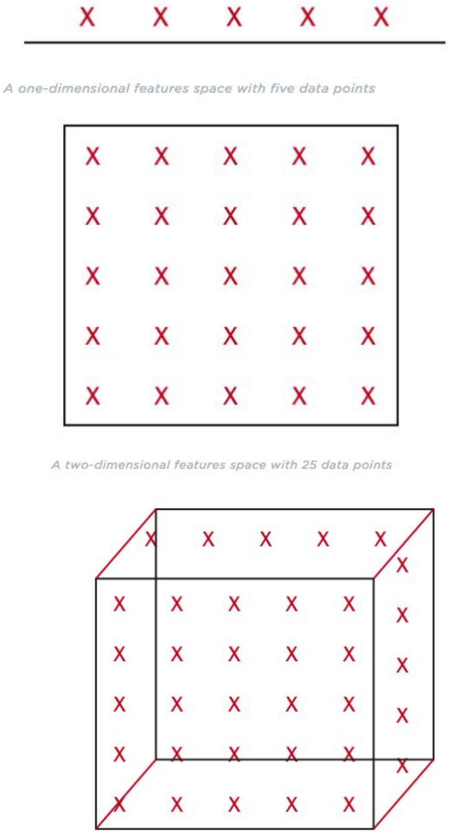
Premise and Motivation

- Classification is predicting labels from a dataset’s features
- Prediction algorithms rely on sufficiently *large* **sample sizes** (n) to train such **features** (p)



Curse of Dimensionality

Accuracies of classification algorithms tend to dip in high dimensions due to the **curse of dimensionality**



A one-dimensional features space with five data points

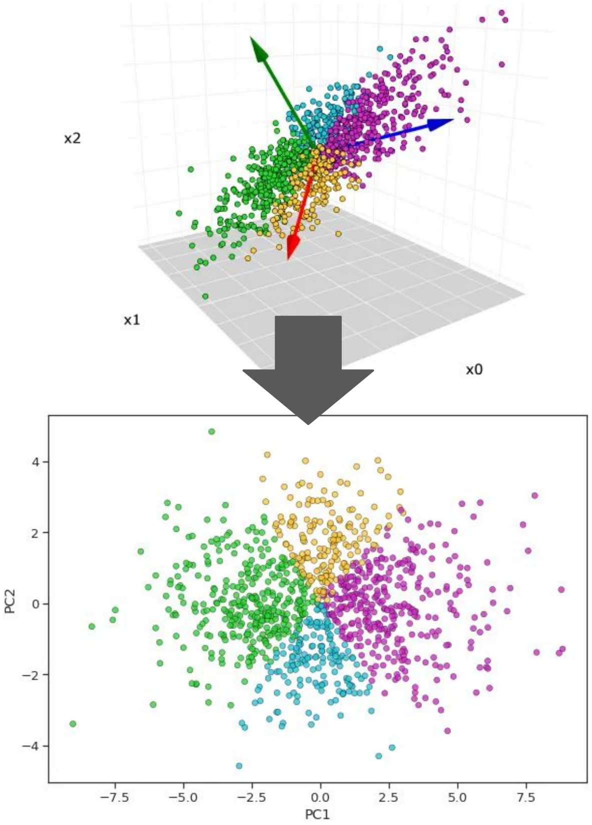
A two-dimensional features space with 25 data points

A three-dimensional features space with 125 data points

V. Pappu, & P.M. Pardalos. High Dimensional Data Classification. Clusters, Orders and Trees: Methods and Applications.

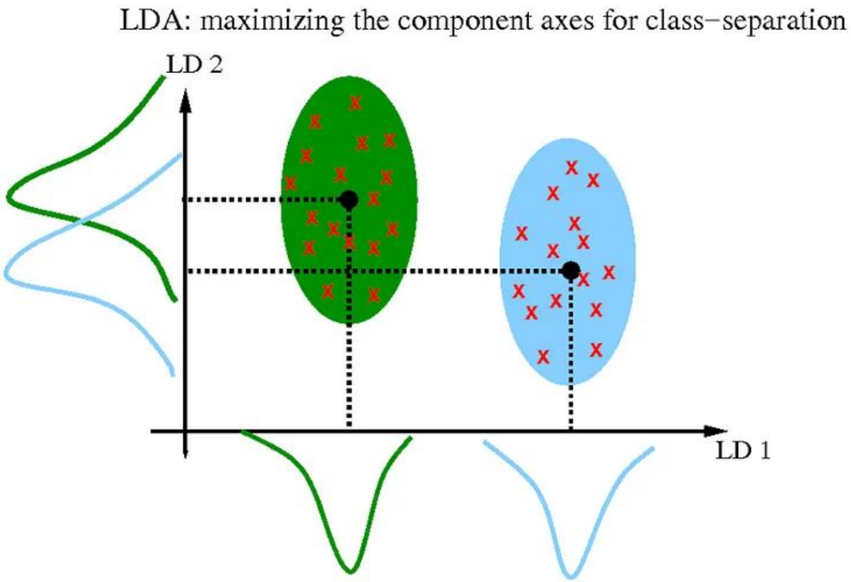
Projection Pursuit

- Seek interesting *low-dimensional* projections



Lee, E.-K., & Cook, D. (2009). A projection pursuit index for large P Small N Data. Statistics and Computing, 20(3), 381–392.

Fisher’s Linear Discriminant Analysis (LDA)



- Supervised dimensionality reduction
- Reduce dimensions but preserve features relevant for class discrimination

Lee, E.-K., & Cook, D. (2009). A projection pursuit index for large P Small N Data. Statistics and Computing, 20(3), 381–392.

Between-Class and Within-Class Scatter Matrices

1. The **between-class** scatter matrix $\mathbf{S_B}$ is computed by the following equation,

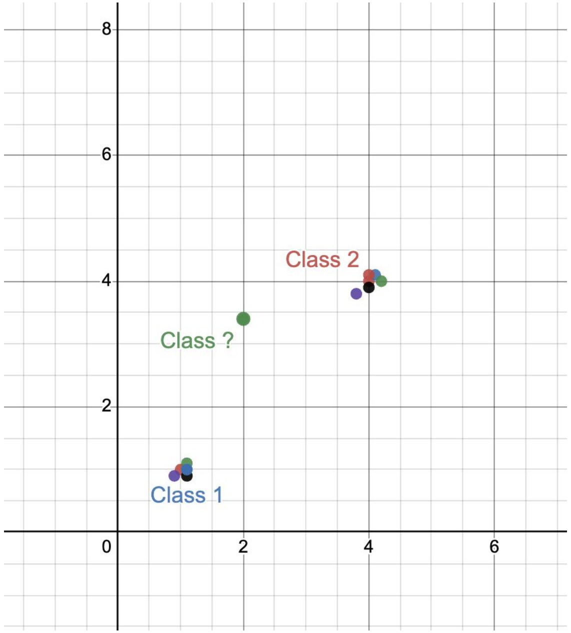
$$S_B = \sum_{i=1}^{c} N_i(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T$$

where $\mathbf{m}$ is the overall mean, and $\mathbf{m_i}$ and $\mathbf{N_i}$ are the sample mean and sizes of the respective classes.

Lee, E.-K., & Cook, D. (2009). A projection pursuit index for large P Small N Data. Statistics and Computing, 20(3), 381–392.

Fisher’s LDA

Intuitively, makes sense to *maximize* **between-class** scatter matrix and *minimize* **within-class** scatter matrix.



Lee, E.-K., & Cook, D. (2009). A projection pursuit index for large P Small N Data. Statistics and Computing, 20(3), 381–392.

Fisher’s LDA

$$\mathbf{Qv} = \boldsymbol{\lambda}\mathbf{v}$$

where…

$$\mathbf{Q} = \mathrm{S_w}^{-1}\mathrm{S_B}$$

$\mathbf{v}$ = eigenvector

$\boldsymbol{\lambda}$ = eigenvalue

Eigenvectors of largest eigenvalues *maximize* **between-class/within-class variance** most

k most important eigenvectors of $\mathbf{S_w}^{-1}\mathbf{S_B}$ are **linear discriminant function**

Projected data trims least important features

$$I_{\mathrm{LDA}}(\mathbf{A}) = \begin{cases} 1 - \dfrac{|\mathbf{A}^T \mathbf{S_W} \mathbf{A}|}{|\mathbf{A}^T (\mathbf{S_W} + \mathbf{S_B}) \mathbf{A}|}, & \text{for } |\mathbf{A}^T(\mathbf{S_W} + \mathbf{S_B})\, \mathbf{A}| \neq 0, \\ 0, & \text{for } |\mathbf{A}^T(\mathbf{S_W} + \mathbf{S_B})\, \mathbf{A}| = 0, \end{cases}$$

Lee, E.-K., & Cook, D. (2009). A projection pursuit index for large P Small N Data. Statistics and Computing, 20(3), 381–392.

Penalized Discriminant Analysis

n < p leads to data piling issues in LDA

Lee, E.-K., & Cook, D. (2009). A projection pursuit index for large P Small N Data. Statistics and Computing, 20(3), 381–392.

PDA PP-Index

Use $\tilde{\boldsymbol{\Sigma}}(\lambda) = (1-\lambda)\hat{\boldsymbol{\Sigma}} + \lambda \cdot \mathbf{diag}(\bar{\hat{\boldsymbol{\Sigma}}})$ to estimate the variance-covariance matrix

Σ (hat) = maximum likelihood estimator (MLE) of **Σ**

Using standardized data vectors, the above equation turns to

$\tilde{\mathbf{R}}(\lambda) = (1-\lambda)\hat{\mathbf{R}} + \lambda\mathbf{I}$, where

R (hat) = MLE of correlation matrix

I = identity matrix

Lee, E.-K., & Cook, D. (2009). A projection pursuit index for large P Small N Data. Statistics and Computing, 20(3), 381–392.

PDA PP-Index

$$\boldsymbol{\Sigma}^{\mathbf{B}} = \sum_{i=1}^{g} n_i (\bar{\mathbf{X}}^*_{i.} - \bar{\mathbf{X}}^*_{..})(\bar{\mathbf{X}}^*_{i.} - \bar{\mathbf{X}}^*_{..})^T,$$

$$\boldsymbol{\Sigma}^{\mathbf{S}} = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\mathbf{X}^*_{ij} - \bar{\mathbf{X}}^*_{i.})(\mathbf{X}^*_{ij} - \bar{\mathbf{X}}^*_{i.})^T$$

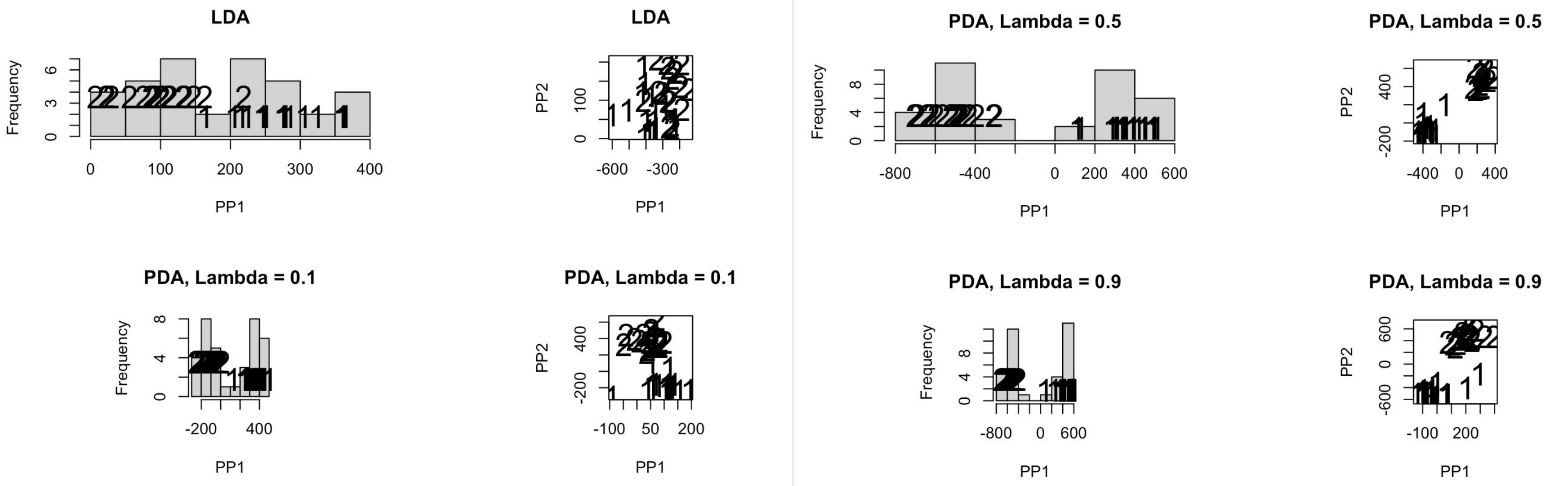
$\mathbf{X}^*_{\mathbf{i}}$ = *i*-th group mean of the standardized data and

$\mathbf{X}^{\mathbf{s}}_{**}$ = 0 = total mean of the standardized data

$$I_{\text{PDA}}(\mathbf{A}, \lambda) = 1 - \frac{|\mathbf{A}^T\{(1-\lambda)\ \boldsymbol{\Sigma}^{\text{S}} + n\lambda \mathbf{I}_p\}\mathbf{A}|}{|\mathbf{A}^T\{(1-\lambda)\ (\boldsymbol{\Sigma}^{\text{B}} + \boldsymbol{\Sigma}^{\text{S}}) + n\lambda \mathbf{I}_p\}\mathbf{A}|}$$

Lee, E.-K., & Cook, D. (2009). A projection pursuit index for large P Small N Data. Statistics and Computing, 20(3), 381–392.

Example Projections



Histogram on the left is 1-D projection
Scatter plot on the right is 2-D projection

Notterman, et al, Cancer Research, vol. 61: 2001

Sample Data Background

Sample:

- 18 malignant cancer tumors paired
- 18 normal tissue from the *same patient*

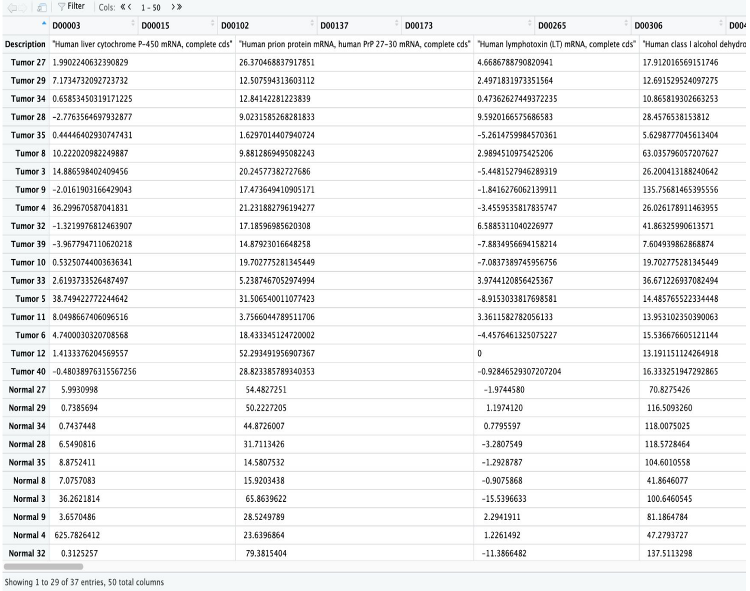
Features:

- 3200 full-length human cDNA
- 3400 expressed sequence tags

Classifier: Malignant Tumor | Normal Tissue

Features (≈6600)

Samples (36)



D00003 D00015 D00102 D00137 D00173 D00265 D00306 D004

Description	"Human liver cytochrome P-450 mRNA, complete cds"	"Human prion protein mRNA, human PrP 27-30 mRNA, complete cds"	"Human lymphotoxin (LT) mRNA, complete cds"	"Human class I alcohol dehydro
Tumor 27	1.9902240632390829	26.370468837917851	4.6686788790820941	17.912016569151746
Tumor 29	7.1734732092723732	12.507594313603112	2.4971831973351564	12.691529524097275
Tumor 34	0.65853450319171225	12.84142281223839	0.47362627449372235	10.865819302663253
Tumor 28	-2.7763564697932877	9.0231585268281833	9.5920166575686583	28.4576538153812
Tumor 35	0.44446402930747431	1.6297014407940724	-5.2614759984570361	5.6298777045613404
Tumor 8	10.222020982249887	9.8812869495082243	2.9894510975425206	63.035796057207627
Tumor 3	14.886598402409456	20.24577382727686	-5.4481527946289319	26.200413188240642
Tumor 9	-2.0161903166429043	17.473649410905171	-1.8416276062139911	135.75681465395556
Tumor 4	36.299670587041831	21.231882796194277	-3.4559535817835747	26.026178911463955
Tumor 32	-1.3219976812463907	17.18596985620308	6.5885311040226977	41.86325990613571
Tumor 39	-3.9677947110620218	14.87923016648258	-7.8834956694158214	7.604939862868874
Tumor 10	0.53250744003636341	19.702775281345449	-7.0837389745956756	19.702775281345449
Tumor 33	2.6193733526487497	5.2387467052974994	3.9744120856425367	36.671226937082494
Tumor 5	38.749422772244642	31.506540011077423	-8.9153033817698581	14.485765522334448
Tumor 11	8.0498667406096516	3.7566044789511706	3.3611582782056133	13.953102350390063
Tumor 6	4.7400030320708568	18.433345124720002	-4.4576461325075227	15.536676605121144
Tumor 12	1.4133376204569557	52.293491956907367	0	13.191151124264918
Tumor 40	-0.48038976315567256	28.823385789340353	-0.92846529307207204	16.333251947292865
Normal 27	5.9930998	54.4827251	-1.9744580	70.8275426
Normal 29	0.7385694	50.2227205	1.1974120	116.5093260
Normal 34	0.7437448	44.8726007	0.7795597	118.0075025
Normal 28	6.5490816	31.7113426	-3.2807549	118.5728464
Normal 35	8.8752411	14.5807532	-1.2928787	104.6010558
Normal 8	7.0757083	15.9203438	-0.9075868	41.8646077
Normal 3	36.2621814	65.8639622	-15.5396633	100.6460545
Normal 9	3.6570486	28.5249789	2.2941911	81.1864784
Normal 4	625.7826412	23.6396864	1.2261492	47.2793727
Normal 32	0.3125257	79.3815404	-11.3866482	137.5113298

Showing 1 to 29 of 37 entries, 50 total columns

Notterman, et al, Cancer Research, vol. 61: 2001 14

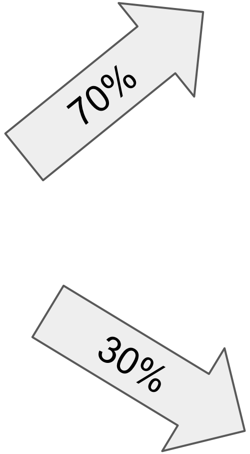
Training/Testing Split

Let’s split *training*/*testing* in a 70:30 ratio

Original dataset:

36 samples

18 normal tissue/**18** tumors



Training dataset:

25 samples

13 normal tissue/**12** tumors

Testing dataset:

11 samples

5 normal tissue/**6** tumors

Notterman, et al, Cancer Research, vol. 61: 2001

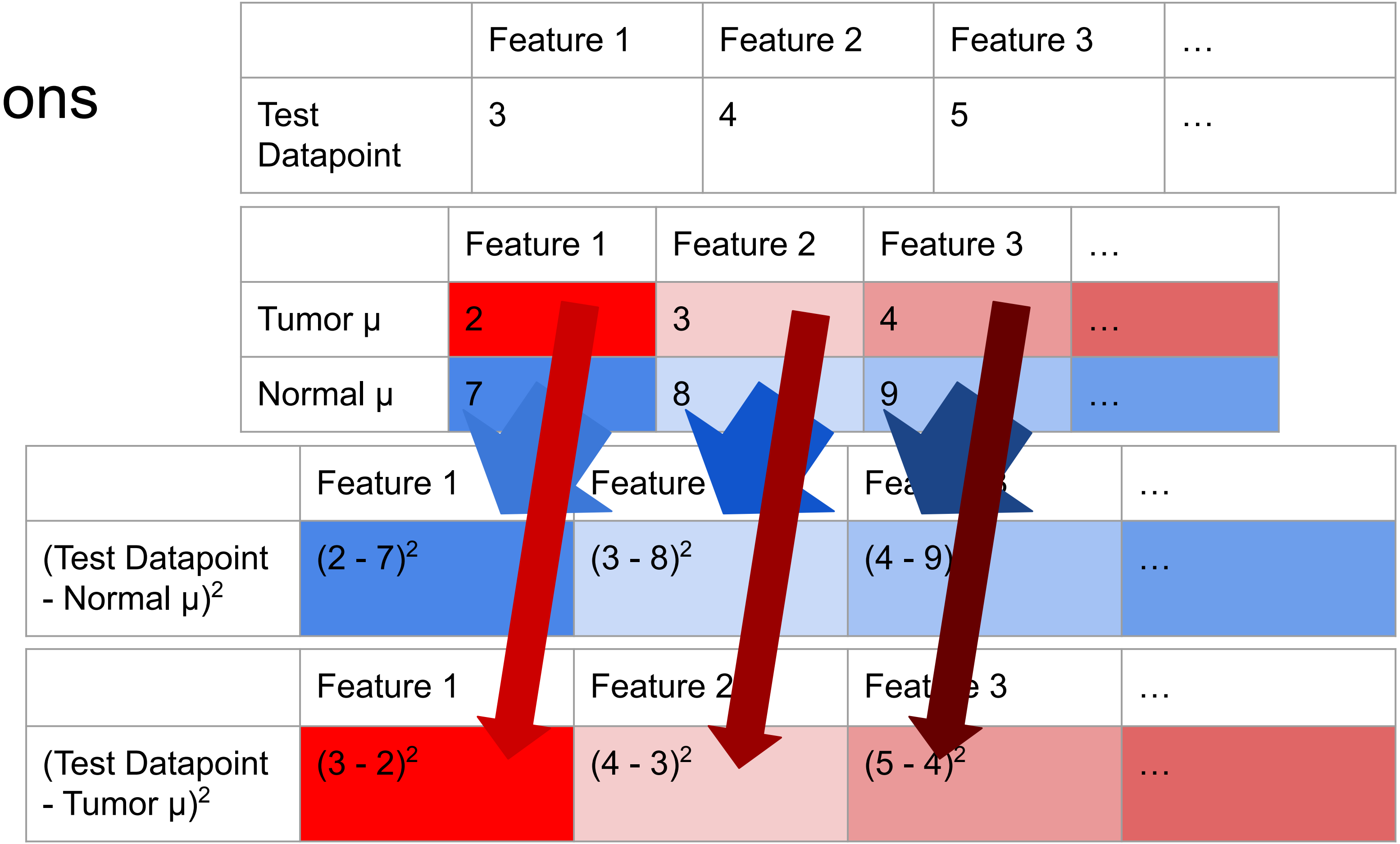
Predictions

	Feature 1	Feature 2	Feature 3	…
Tumor µ	2	3	4	…
Normal µ	7	8	9	…

	Feature 1	Feature 2	Feature 3	…
Tumor 1…	1	6	10	…
Tumor 2…	2	7	11	…
Tumor *x*…	…	…	…	…
Normal 1…	5	5	7	…
Normal 2…	6	6	8	…
Normal *y*…	…	…	…	…

Notterman, et al, Cancer Research, vol. 61: 2001

Predictions



	Feature 1	Feature 2	Feature 3	…
Test Datapoint	3	4	5	…

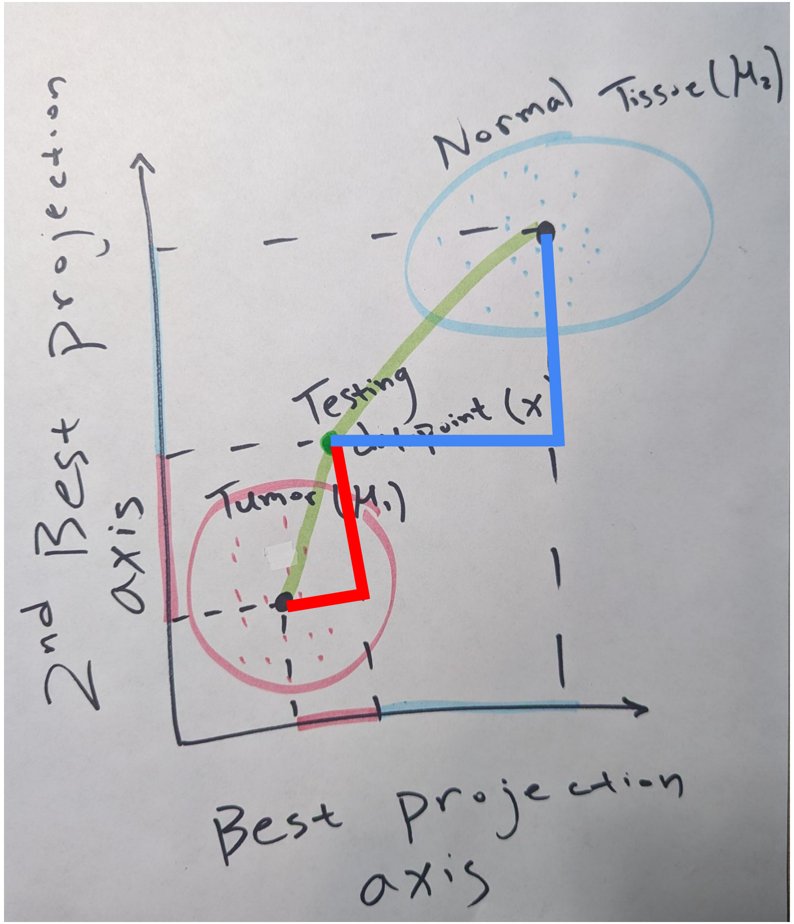
	Feature 1	Feature 2	Feature 3	…
Tumor μ	2	3	4	…
Normal μ	7	8	9	…

	Feature 1	Feature 2	Feature 3	…
(Test Datapoint - Normal μ)^2	$(2 - 7)^2$	$(3 - 8)^2$	$(4 - 9)^2$	…

	Feature 1	Feature 2	Feature 3	…
(Test Datapoint - Tumor μ)^2	$(3 - 2)^2$	$(4 - 3)^2$	$(5 - 4)^2$	…

Notterman, et al, Cancer Research, vol. 61: 2001

Predictions

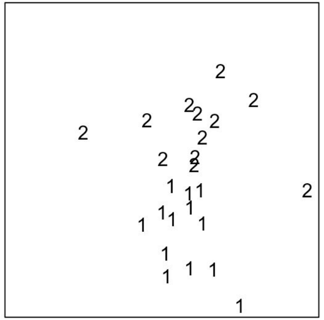


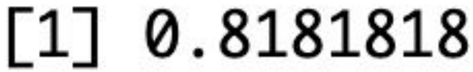
Testing datapoint is classified as **Tumor** since its distance to the **Tumor mean** is *smaller* than to **Normal Tissue's mean**

Notterman, et al, Cancer Research, vol. 61: 2001

Prediction Results

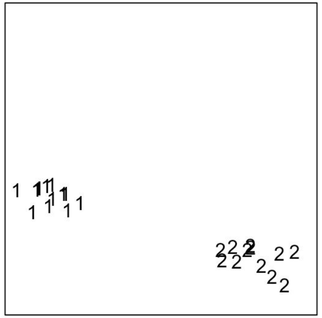
2-Dimensional LDA
projection of testing data

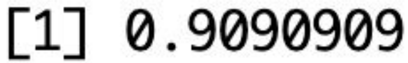




```
[1] 0.8181818
```

2-Dimensional PDA (**λ** = 0.5)
projection of testing data

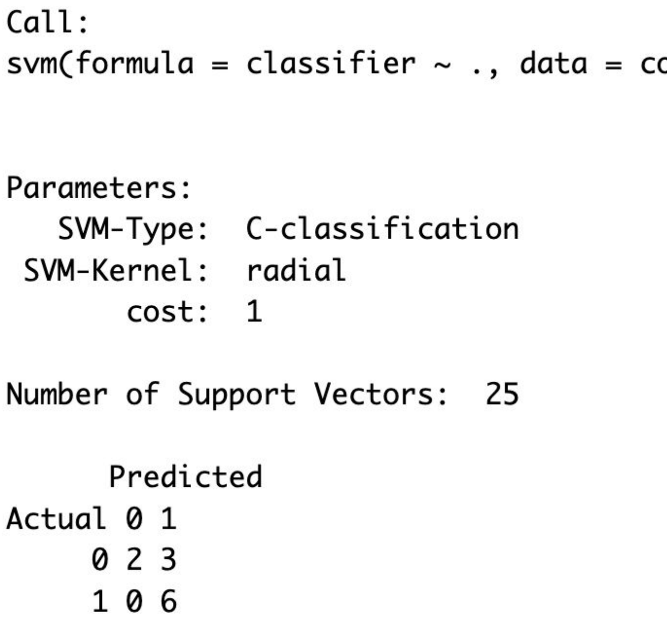




```
[1] 0.9090909
```

Notterman, et al, Cancer Research, vol. 61: 2001

Support Vector Machines



```
Call:
svm(formula = classifier ~ ., data = cc


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1

Number of Support Vectors:  25

      Predicted
Actual 0 1
     0 2 3
     1 0 6
```

e1071 package

2/5 normal tissues *correctly* classified

3/5 normal tissues *incorrectly* classified as tumors

6/6 tumors *correctly* classified

Accuracy: (2 + 6)/(5 + 6) = 8/11 ≈ 72.73%

Notterman, et al, Cancer Research, vol. 61: 2001

References (images)

1201904. “Dimensionality Reduction Techniques Skill Test for Data Scientists (Updated 2023).” Analytics Vidhya, 14 Mar. 2023, www.analyticsvidhya.com/blog/2017/03/questions-dimensionality-reduction-data-scientist/.

Liu, Jian, et al. “Cancer Characteristic Gene Selection via Sample Learning Based on Deep Sparse Filtering.” The Differences between Sample Learning and Feature Learning., May 2018, www.researchgate.net/figure/The-differences-between-sample-learning-and-feature-learning-a-A-feature-learning_fig1_325426703.

Raschka, Sebastian. “Linear Discriminant Analysis.” Sebastian Raschka, PhD, 3 Aug. 2014, sebastianraschka.com/Articles/2014_python_lda.html.

Shetty, Badreesh. “What Is the Curse of Dimensionality?” Built In, 19 Aug. 2022, builtin.com/data-science/curse-dimensionality.

Vungarala, Seshu Kumar. “PCA vs Lda‑No More Confusion!” Medium, Medium, 30 Apr. 2023, medium.com/@seshu8hachi/pca-vs-lda-no-more-confusion-fc21fb8d06e9#:~:text=PCA%20is%20an%20unsupervised%20method%20that%20aims%20to%20find%20the,the%20characteristics%20of%20the%20dataset.

References (papers)

Kabacoff, R. (2015). R in Action, 2nd edition, Chapter 17 (ISBN: 9781617291388). Manning Publications Co.

Lee, E.-K., & Cook, D. (2009). A projection pursuit index for large P Small N Data. Statistics and Computing, 20(3), 381–392.

Notterman, et al, Cancer Research vol. 61: 2001

V. Pappu, & P.M. Pardalos. High Dimensional Data Classification. Clusters, Orders and Trees: Methods and Applications.