Statistical Pattern Recognition
Lecture4
Bayesian Learning

Dr Zohreh Azimifar

School of Electrical and Computer Engineering

Shiraz University

Fall2014

Table of contents

Introduction

- Classification based on the theory Bayesian Learning

$$P(y = 0|\mathbf{X}) \gtrless_{y=1}^{y=0} P(y = 1|\mathbf{X})$$

- Classification involves determining $P(y|\mathbf{X})$, from different perspectives.

Generative Learning and Discriminative Learning

1. Discriminative Learning:
   - Direct learning of $P(y|\mathbf{X})$.
   - Modelling of decision boundary, to which side a new sample is assigned.
   - Logistic and softmax regression are called discriminative learners.
2. Generative Learning
   - Explicit modelling of each class separately.
   - Compare new sample with each class probability, based on Bayesian rule:

$$P(y|\mathbf{X}) = \frac{\overbrace{P(\mathbf{X}|y)}^{\text{Likelihood}}\overbrace{P(y)}^{\text{Prior}}}{\underbrace{P(\mathbf{X})}_{\text{normalizing factor}}}$$

Gaussian Linear Discriminant Analysis

- Model $P(y)$ and $P(\mathbf{X}|y)$ for each class y:

$$P(y) = \phi_1^{1\{y=1\}} \phi_2^{1\{y=2\}} \cdots \phi_c^{1\{y=c\}}$$

$$P(\mathbf{X}|y=i) = \frac{1}{(\sqrt{2\pi})^n |\Sigma|^{\frac{1}{2}}} \; exp(\frac{-1}{2}(\mathbf{X} - \boldsymbol{\mu}_i)^T \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu}_i))$$

- Parameter set: $\boldsymbol{\theta} = \{\phi_1, \phi_2, ..., \phi_c, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, ..., \boldsymbol{\mu}_c, \Sigma\}$

Gaussian Linear Discriminant Analysis

- Estimate parameters: $\boldsymbol{\theta} = \{\phi_1, \phi_2, ..., \phi_c, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, ..., \boldsymbol{\mu}_c, \Sigma\}$

$$l(\boldsymbol{\theta}) = log \prod_{j=1}^{m} P(\mathbf{X}^{(j)}, y^{(j)}) = log \prod_{j=1}^{m} P(\mathbf{X}^{(j)} | P(y^{(j)}))P(y^{(j)})$$

Gaussian Linear Discriminant Analysis

- Estimate parameters: $\boldsymbol{\theta} = \{\phi_1, \phi_2, ..., \phi_c, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, ..., \boldsymbol{\mu}_c, \Sigma\}$

$$l(\boldsymbol{\theta}) = log \prod_{j=1}^{m} P(\mathbf{X}^{(j)}, y^{(j)}) = log \prod_{j=1}^{m} P(\mathbf{X}^{(j)} | P(y^{(j)})) P(y^{(j)})$$

- Take partial derivative in terms of each individual parameter:

$$\phi_i^{MLE} = \frac{\sum_{j=1}^{m} 1\{y^{(j)} = i\}}{m}$$

Gaussian Linear Discriminant Analysis

- Estimate parameters: $\boldsymbol{\theta} = \{\phi_1, \phi_2, ..., \phi_c, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, ..., \boldsymbol{\mu}_c, \Sigma\}$

$$l(\boldsymbol{\theta}) = log \prod_{j=1}^{m} P(\mathbf{X}^{(j)}, y^{(j)}) = log \prod_{j=1}^{m} P(\mathbf{X}^{(j)} | P(y^{(j)}))P(y^{(j)})$$

- Take partial derivative in terms of each individual parameter:

$$\phi_i^{MLE} = \frac{\sum_{j=1}^{m} 1\{y^{(j)} = i\}}{m}$$

$$\boldsymbol{\mu}_i^{MLE} = \frac{\sum_{j=1}^{m} 1\{y^{(j)} = i\}\mathbf{X}^{(j)}}{\sum_{j=1}^{m} 1\{y^{(j)} = i\}}$$

Gaussian Linear Discriminant Analysis

- Estimate parameters: $\boldsymbol{\theta} = \{\phi_1, \phi_2, ..., \phi_c, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, ..., \boldsymbol{\mu}_c, \Sigma\}$

$$l(\boldsymbol{\theta}) = log \prod_{j=1}^{m} P(\mathbf{X}^{(j)}, y^{(j)}) = log \prod_{j=1}^{m} P(\mathbf{X}^{(j)} | P(y^{(j)})) P(y^{(j)})$$

- Take partial derivative in terms of each individual parameter:

$$\phi_i^{MLE} = \frac{\sum_{j=1}^{m} 1\{y^{(j)} = i\}}{m}$$

$$\boldsymbol{\mu}_i^{MLE} = \frac{\sum_{j=1}^{m} 1\{y^{(j)} = i\} \mathbf{X}^{(j)}}{\sum_{j=1}^{m} 1\{y^{(j)} = i\}}$$

$$\Sigma^{MLE} = \frac{1}{m} \sum_{j=1}^{m} (\mathbf{X}^{(j)} - \boldsymbol{\mu}_{y^{(j)}})(\mathbf{X}^{(j)} - \boldsymbol{\mu}_{y^{(j)}})^T$$

Gaussian Linear Discriminant Analysis

- Determine class label of a new sample $\mathbf{X}^{new}$:

$$\begin{aligned} y^{new} &= argmax_y \; P(y|\mathbf{X}) \\ &= argmax_y \; \frac{P(\mathbf{X}|y)P(y)}{P(X)} \\ &= argmax_y \; P(\mathbf{X}|y)P(y) \end{aligned}$$

- Note that $P(\mathbf{X}|y)$ is a class dependent density.

Boundary Decision for GLDA

- Decision boundary is a line, a plane, or a hyper-plane. Why?

$$P(y = i|\mathbf{X}) = P(y = j|\mathbf{X})$$

$$\frac{P(\mathbf{X}|y = i)P(y = i)}{P(\mathbf{X})} = \frac{P(\mathbf{X}|y = j)P(y = j)}{P(\mathbf{X})}$$

Boundary Decision for GLDA

- Decision boundary is a line, a plane, or a hyper-plane. Why?

$$P(y = i|\mathbf{X}) = P(y = j|\mathbf{X})$$

$$\frac{P(\mathbf{X}|y = i)P(y = i)}{P(\mathbf{X})} = \frac{P(\mathbf{X}|y = j)P(y = j)}{P(\mathbf{X})}$$

$$P(\mathbf{X}|y = i)P(y = i) = P(\mathbf{X}|y = j)P(y = j)$$

Boundary Decision for GLDA

- Decision boundary is a line, a plane, or a hyper-plane. Why?

$$P(y=i|\mathbf{X}) = P(y=j|\mathbf{X})$$
$$\frac{P(\mathbf{X}|y=i)P(y=i)}{P(\mathbf{X})} = \frac{P(\mathbf{X}|y=j)P(y=j)}{P(\mathbf{X})}$$

$$P(\mathbf{X}|y=i)P(y=i) = P(\mathbf{X}|y=j)P(y=j)$$

$$\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}\ exp(\frac{-1}{2}(\mathbf{X}-\boldsymbol{\mu}_i)^T\Sigma^{-1}(\mathbf{X}-\boldsymbol{\mu}_i))P(y=i)$$
$$= \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}\ exp(\frac{-1}{2}(\mathbf{X}-\boldsymbol{\mu}_j)^T\Sigma^{-1}(\mathbf{X}-\boldsymbol{\mu}_j))P(y=j)$$

Boundary Decision for GLDA

- Decision boundary cont'ed:

$$exp(\frac{-1}{2}(\mathbf{X}-\boldsymbol{\mu}_i)^T\Sigma^{-1}(\mathbf{X}-\boldsymbol{\mu}_i))P(y=i) = exp(\frac{-1}{2}(\mathbf{X}-\boldsymbol{\mu}_j)^T\Sigma^{-1}(\mathbf{X}-\boldsymbol{\mu}_j))P(y=j)$$

Boundary Decision for GLDA

- Decision boundary cont'ed:

$$exp(\frac{-1}{2}(\mathbf{X}-\boldsymbol{\mu}_i)^T\Sigma^{-1}(\mathbf{X}-\boldsymbol{\mu}_i))P(y=i) = exp(\frac{-1}{2}(\mathbf{X}-\boldsymbol{\mu}_j)^T\Sigma^{-1}(\mathbf{X}-\boldsymbol{\mu}_j))P(y=j)$$

$$\frac{-1}{2}(\mathbf{X}-\boldsymbol{\mu}_i)^T\Sigma^{-1}(\mathbf{X}-\boldsymbol{\mu}_i) + log\ P(y=i) = \frac{-1}{2}(\mathbf{X}-\boldsymbol{\mu}_j)^T\Sigma^{-1}(\mathbf{X}-\boldsymbol{\mu}_j) + log\ P(y=j)$$

Boundary Decision for GLDA

- Decision boundary cont'ed:

$$exp(\frac{-1}{2}(\mathbf{X}-\boldsymbol{\mu}_i)^T\Sigma^{-1}(\mathbf{X}-\boldsymbol{\mu}_i))P(y=i) = exp(\frac{-1}{2}(\mathbf{X}-\boldsymbol{\mu}_j)^T\Sigma^{-1}(\mathbf{X}-\boldsymbol{\mu}_j))P(y=j)$$

$$\frac{-1}{2}(\mathbf{X}-\boldsymbol{\mu}_i)^T\Sigma^{-1}(\mathbf{X}-\boldsymbol{\mu}_i) + log\ P(y=i) = \frac{-1}{2}(\mathbf{X}-\boldsymbol{\mu}_j)^T\Sigma^{-1}(\mathbf{X}-\boldsymbol{\mu}_j) + log\ P(y=j)$$

$$\Rightarrow \quad log\ \frac{P(y=i)}{P(y=j)} - \frac{1}{2}[\mathbf{X}^T\Sigma^{-1}\mathbf{X} - 2\mathbf{X}^T\Sigma^{-1}\boldsymbol{\mu}_i + \boldsymbol{\mu}_i^T\Sigma^{-1}\boldsymbol{\mu}_i] + \frac{1}{2}[\mathbf{X}^T\Sigma^{-1}\mathbf{X} - 2\mathbf{X}^T\Sigma^{-1}\boldsymbol{\mu}_j + \boldsymbol{\mu}_j^T\Sigma^{-1}\boldsymbol{\mu}_j] = 0$$

Boundary Decision for GLDA

- Decision boundary cont'ed:

$$exp(\frac{-1}{2}(\mathbf{X}-\boldsymbol{\mu}_i)^T\Sigma^{-1}(\mathbf{X}-\boldsymbol{\mu}_i))P(y=i) = exp(\frac{-1}{2}(\mathbf{X}-\boldsymbol{\mu}_j)^T\Sigma^{-1}(\mathbf{X}-\boldsymbol{\mu}_j))P(y=j)$$

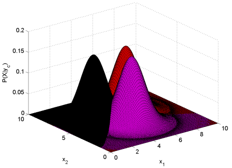
$$\frac{-1}{2}(\mathbf{X}-\boldsymbol{\mu}_i)^T\Sigma^{-1}(\mathbf{X}-\boldsymbol{\mu}_i) + log\ P(y=i) = \frac{-1}{2}(\mathbf{X}-\boldsymbol{\mu}_j)^T\Sigma^{-1}(\mathbf{X}-\boldsymbol{\mu}_j) + log\ P(y=j)$$

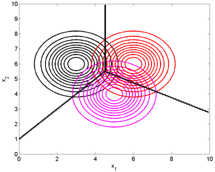
$$\Rightarrow \quad log\ \frac{P(y=i)}{P(y=j)} - \frac{1}{2}[\mathbf{X}^T\Sigma^{-1}\mathbf{X} - 2\mathbf{X}^T\Sigma^{-1}\boldsymbol{\mu}_i + \boldsymbol{\mu}_i^T\Sigma^{-1}\boldsymbol{\mu}_i] + \frac{1}{2}[\mathbf{X}^T\Sigma^{-1}\mathbf{X} - 2\mathbf{X}^T\Sigma^{-1}\boldsymbol{\mu}_j + \boldsymbol{\mu}_j^T\Sigma^{-1}\boldsymbol{\mu}_j] = 0$$

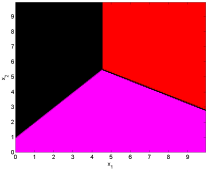
$$\Rightarrow \quad \underbrace{\mathbf{X}^T\Sigma^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)}_{a\mathbf{X}} + \underbrace{\frac{1}{2}\boldsymbol{\mu}_i^T\Sigma^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\boldsymbol{\mu}_j^T\Sigma^{-1}\boldsymbol{\mu}_j + log\ \frac{P(y=i)}{P(y=j)}}_{b} = 0$$

Analysis of GLDA when $\Sigma = \sigma^2 I$

- Classes are of identical distribution, but different means.
- Cross-section of classes distribution is spherical.
- Decision boundary is linear.
- Called classifier with nearest Euclidean distance to the class mean, when?



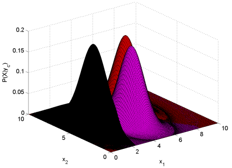


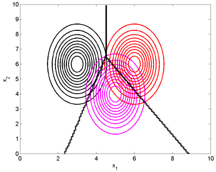


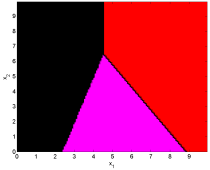
$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Analysis of GLDA when Σ is not identity

- Classes are of identical distribution, but different means.
- Cross-section of classes distribution is ellipsoidal.
- Decision boundary is linear.



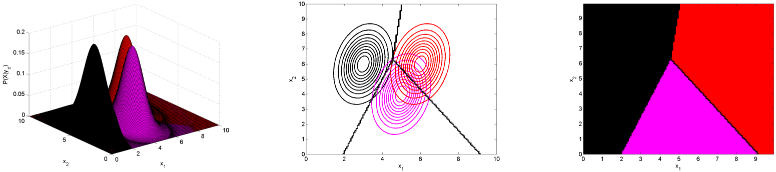




$$\Sigma = \begin{bmatrix} 0.5 & 0 \\ 0 & 1.5 \end{bmatrix}$$

Analysis of GLDA with arbitrary Σ

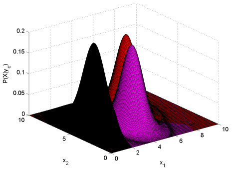
- Classes are of identical distribution, but different means.
- Cross-section of classes distribution is ellipsoidal. Linear decision boundary.
- Classes are aligned with direction of covariance eigenvectors.

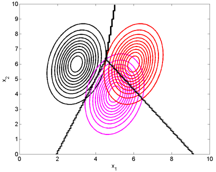


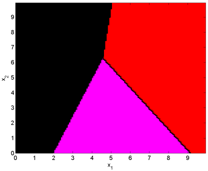
$$\Sigma = \begin{bmatrix} 0.5 & 0.2 \\ 0.2 & 1.5 \end{bmatrix}$$

Analysis of GLDA with arbitrary Σ

- Classes are of identical distribution, but different means.
- Cross-section of classes distribution is ellipsoidal. Linear decision boundary.
- Classes are aligned with direction of covariance eigenvectors.







$$\Sigma = \begin{bmatrix} 0.5 & 0.2 \\ 0.2 & 1.5 \end{bmatrix}$$

- Called classifier with nearest Mahalanobis distance to the class mean, when? $Dist(\mathbf{X}, \boldsymbol{\mu}_\mathbf{i}) = (\mathbf{X} - \boldsymbol{\mu}_\mathbf{i})^T \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu}_\mathbf{i})$

Quadratic Discriminant Analysis

- Another generative learning model; a Bayesian classifier
- Here, classes are of multinomial distribution, and likelihoods are multivariate Gaussian with separate covariance Σ_i.
- Decision boundary becomes non-linear, Why?

$$P(\mathbf{X}|y=i) = \frac{1}{(\sqrt{2\pi})^n|\Sigma_i|^{\frac{1}{2}}}\ exp(\frac{-1}{2}(\mathbf{X}-\boldsymbol{\mu}_i)^T\Sigma_i^{-1}(\mathbf{X}-\boldsymbol{\mu}_i))$$

Quadratic Discriminant Analysis

Decision boundary is parabolic:

$$P(y = i|\mathbf{X}) = P(y = j|\mathbf{X})$$
$$\frac{P(\mathbf{X}|y = i)P(y = i)}{P(\mathbf{X})} = \frac{P(\mathbf{X}|y = j)P(y = j)}{P(\mathbf{X})}$$
$$P(\mathbf{X}|y = i)P(y = i) = P(\mathbf{X}|y = j)P(y = j)$$

Quadratic Discriminant Analysis

Decision boundary is parabolic:

$$P(y = i|\mathbf{X}) = P(y = j|\mathbf{X})$$
$$\frac{P(\mathbf{X}|y = i)P(y = i)}{P(\mathbf{X})} = \frac{P(\mathbf{X}|y = j)P(y = j)}{P(\mathbf{X})}$$
$$P(\mathbf{X}|y = i)P(y = i) = P(\mathbf{X}|y = j)P(y = j)$$

$$\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma_i|^{\frac{1}{2}}}\ exp(\frac{-1}{2}(\mathbf{X} - \boldsymbol{\mu}_i)^T\Sigma_i^{-1}(\mathbf{X} - \boldsymbol{\mu}_i))P(y = i)$$
$$= \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma_j|^{\frac{1}{2}}}\ exp(\frac{-1}{2}(\mathbf{X} - \boldsymbol{\mu}_j)^T\Sigma_j^{-1}(\mathbf{X} - \boldsymbol{\mu}_j))P(y = j)$$

Quadratic Discriminant Analysis

Decision boundary cont'ed:

$$\frac{1}{|\Sigma_i|^{\frac{1}{2}}}\ exp(\frac{-1}{2}(\mathbf{X}-\boldsymbol{\mu}_i)^T\Sigma_i^{-1}(\mathbf{X}-\boldsymbol{\mu}_i))P(y=i)$$
$$=\frac{1}{|\Sigma_j|^{\frac{1}{2}}}\ exp(\frac{-1}{2}(\mathbf{X}-\boldsymbol{\mu}_j)^T\Sigma_j^{-1}(\mathbf{X}-\boldsymbol{\mu}_j))P(y=j)$$

Quadratic Discriminant Analysis

Decision boundary cont'ed:

$$\frac{1}{|\Sigma_i|^{\frac{1}{2}}}\ exp(\frac{-1}{2}(\mathbf{X}-\boldsymbol{\mu}_i)^T\Sigma_i^{-1}(\mathbf{X}-\boldsymbol{\mu}_i))P(y=i)$$
$$=\frac{1}{|\Sigma_j|^{\frac{1}{2}}}\ exp(\frac{-1}{2}(\mathbf{X}-\boldsymbol{\mu}_j)^T\Sigma_j^{-1}(\mathbf{X}-\boldsymbol{\mu}_j))P(y=j)$$

$$\frac{-1}{2}log|\Sigma_i|\ -\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu}_i)^T\Sigma_i^{-1}(\mathbf{X}-\boldsymbol{\mu}_i)+logP(y=i)$$
$$=\frac{-1}{2}log|\Sigma_j|\ -\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu}_j)^T\Sigma_j^{-1}(\mathbf{X}-\boldsymbol{\mu}_j)+logP(y=j)$$

Quadratic Discriminant Analysis

Decision boundary cont'ed:

$$\frac{1}{|\Sigma_i|^{\frac{1}{2}}} \; exp(\frac{-1}{2}(\mathbf{X}-\boldsymbol{\mu}_i)^T\Sigma_i^{-1}(\mathbf{X}-\boldsymbol{\mu}_i))P(y=i)$$
$$= \frac{1}{|\Sigma_j|^{\frac{1}{2}}} \; exp(\frac{-1}{2}(\mathbf{X}-\boldsymbol{\mu}_j)^T\Sigma_j^{-1}(\mathbf{X}-\boldsymbol{\mu}_j))P(y=j)$$

$$\frac{-1}{2}log|\Sigma_i| \; - \frac{1}{2}(\mathbf{X}-\boldsymbol{\mu}_i)^T\Sigma_i^{-1}(\mathbf{X}-\boldsymbol{\mu}_i) + logP(y=i)$$
$$= \frac{-1}{2}log|\Sigma_j| \; - \frac{1}{2}(\mathbf{X}-\boldsymbol{\mu}_j)^T\Sigma_j^{-1}(\mathbf{X}-\boldsymbol{\mu}_j) + logP(y=j)$$

$$\Rightarrow \quad log\frac{P(y=i)}{P(y=j)} - \frac{1}{2}log\frac{|\Sigma_i|}{|\Sigma_j|} - \frac{1}{2}[\mathbf{X}^T\Sigma_i^{-1}\mathbf{X} + \boldsymbol{\mu}_i^T\Sigma_i^{-1}\boldsymbol{\mu}_i -$$
$$2\mathbf{X}^T\Sigma_i^{-1}\boldsymbol{\mu}_i - \mathbf{X}^T\Sigma_j^{-1}\mathbf{X} - \boldsymbol{\mu}_j^T\Sigma_j^{-1}\boldsymbol{\mu}_j + 2\mathbf{X}^T\Sigma_j^{-1}\boldsymbol{\mu}_j] = 0$$

Quadratic Discriminant Analysis

Decision boundary cont'ed:

$$log\frac{P(y=i)}{P(y=j)} - \frac{1}{2}log\frac{\left|\Sigma_i\right|}{\left|\Sigma_j\right|} - \frac{1}{2}[\mathbf{X}^T(\Sigma_i^{-1} - \Sigma_j^{-1})\mathbf{X} + \boldsymbol{\mu}_i^T\Sigma_i^{-1}\boldsymbol{\mu}_i$$
$$- \boldsymbol{\mu}_j^T\Sigma_j^{-1}\boldsymbol{\mu}_j - 2\mathbf{X}^T(\Sigma_i^{-1}\boldsymbol{\mu}_i - \Sigma_j^{-1}\boldsymbol{\mu}_j)] = 0$$

Quadratic Discriminant Analysis

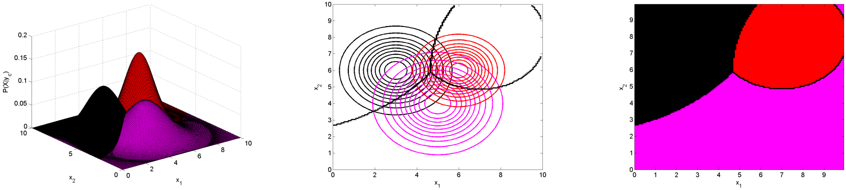
Decision boundary cont'ed:

$$log\frac{P(y=i)}{P(y=j)} - \frac{1}{2}log\frac{\left|\Sigma_i\right|}{\left|\Sigma_j\right|} - \frac{1}{2}[\mathbf{X}^T(\Sigma_i^{-1} - \Sigma_j^{-1})\mathbf{X} + \boldsymbol{\mu}_i^T\Sigma_i^{-1}\boldsymbol{\mu}_i$$
$$- \boldsymbol{\mu}_j^T\Sigma_j^{-1}\boldsymbol{\mu}_j - 2\mathbf{X}^T(\Sigma_i^{-1}\boldsymbol{\mu}_i - \Sigma_j^{-1}\boldsymbol{\mu}_j)] = 0$$

$$\Rightarrow \quad \mathbf{X}^T a\mathbf{X} + b^T\mathbf{X} + c = 0$$

Analysis of QDA when $\Sigma = \sigma_i^2 I$

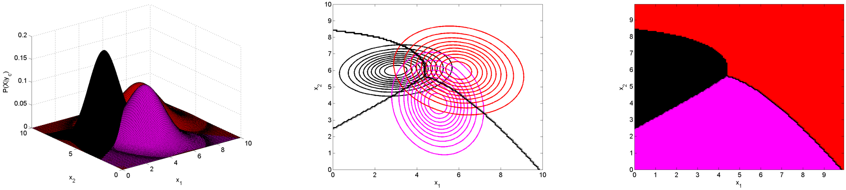
- Classes are of different distributions and different means.
- Cross-sections of classes distribution are spherical but of different sizes.



$$\Sigma_1 = \begin{bmatrix} 1.5 & 0 \\ 0 & 1.5 \end{bmatrix},\ \Sigma_2 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix},\ \Sigma_3 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Analysis of QDA with arbitrary $\Sigma_i \neq \Sigma_j$

- Classes are of different distributions and different means.
- Cross-sections of classes distribution are ellipsoidal and of different sizes.



$$\Sigma_1 = \begin{bmatrix} 1.5 & 0.1 \\ 0.1 & 0.5 \end{bmatrix}, \; \Sigma_2 = \begin{bmatrix} 1 & -0.2 \\ -0.2 & 2 \end{bmatrix}, \; \Sigma_3 = \begin{bmatrix} 2 & -0.25 \\ -0.25 & 1.5 \end{bmatrix}$$

GLAD and QDA, Another point of view

-
-
-

Naive Bayes

Naive Bayes: An Example

Summary