Machine Learning I

MICRO-455

Classification with GMM and KNN

The Teaching Team

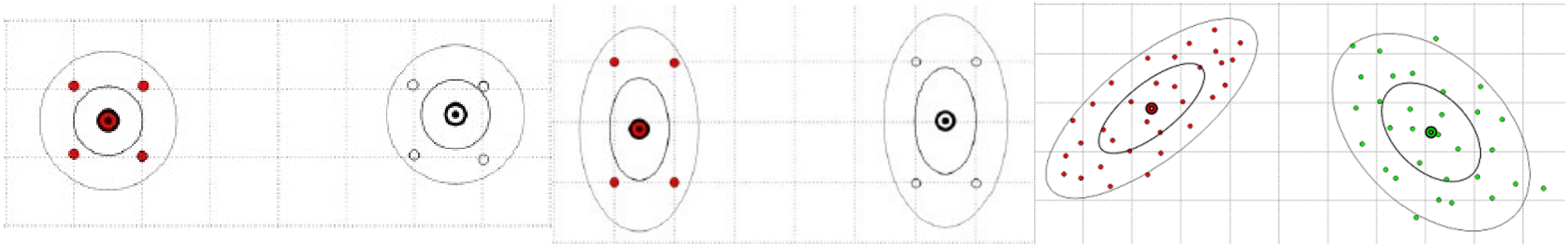
EPFL

Fall 2025

LASA

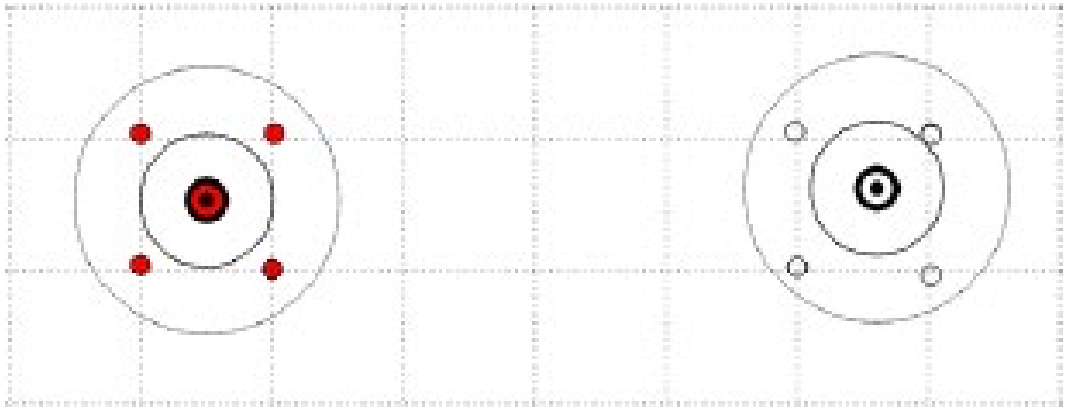
Q1.A

- Drawing the classification boundary.
  - Two classes, each fitted with a single Gauss function.
  - Two classes have the *same* number of datapoints.
  - The Gauss functions are *not* normalized.



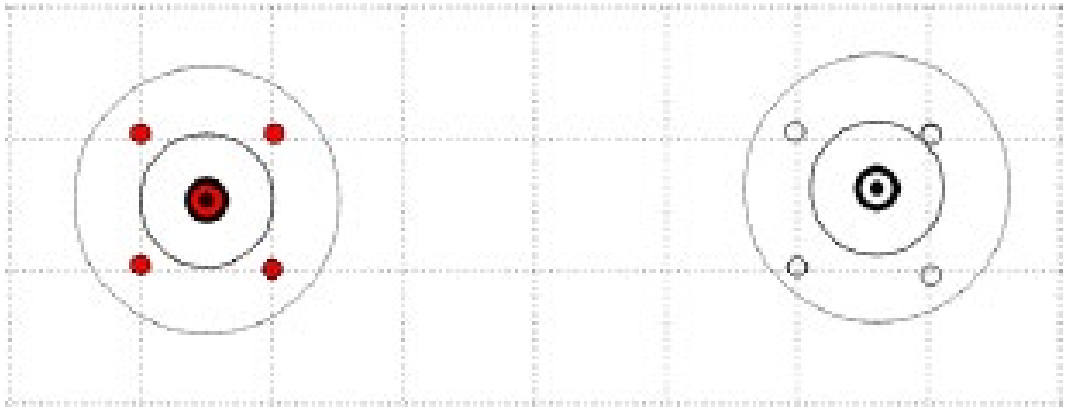
Q1.A > a

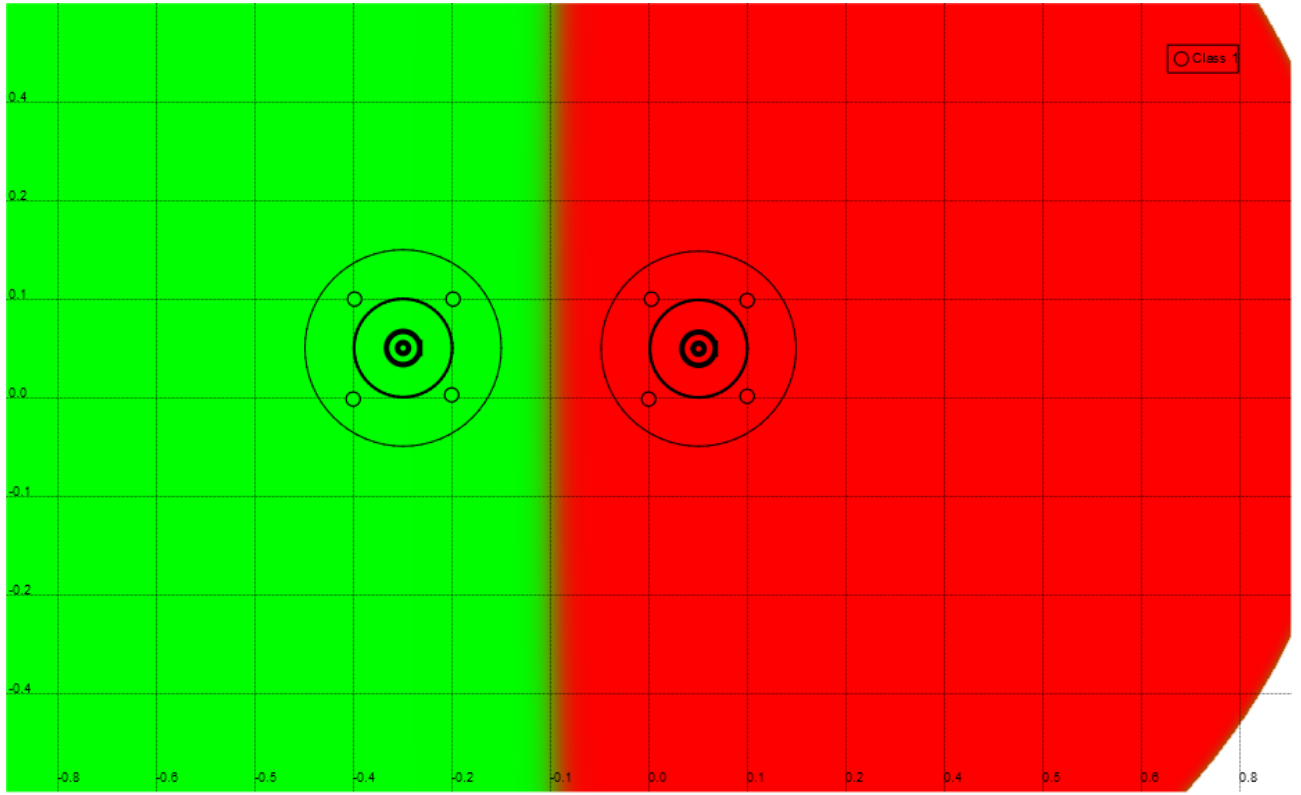
- Identical spherical covariance matrices.



Q1.A > a

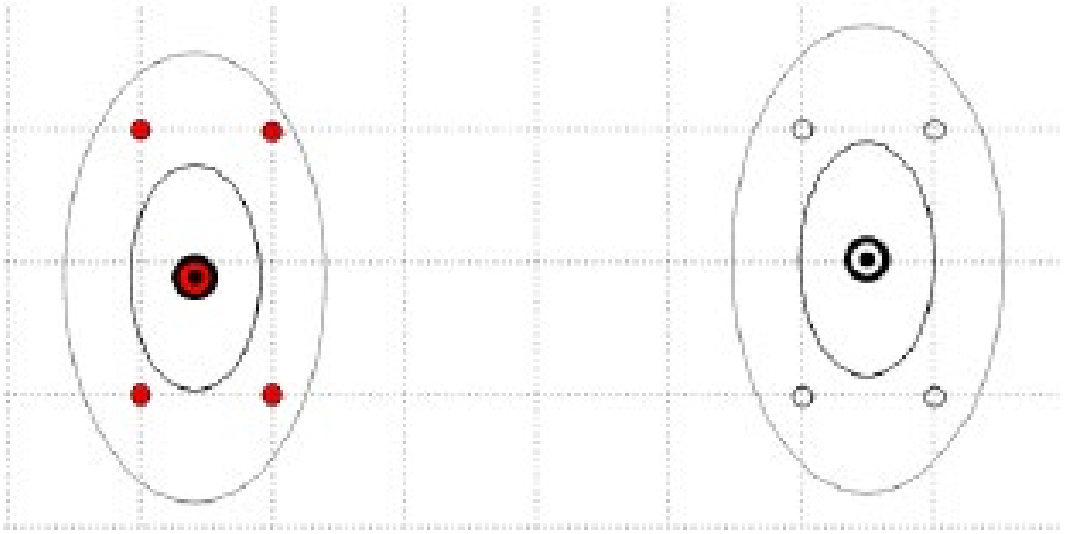
- Identical spherical covariance matrices.
- Linear boundary is equidistant from the two centroids.





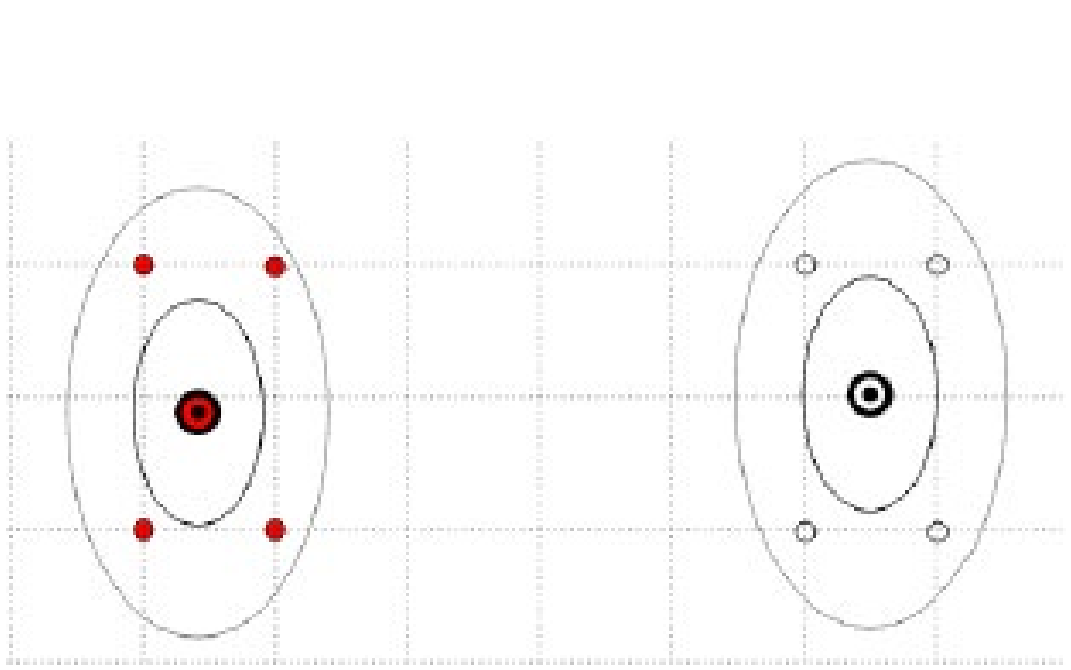
Q1.A > b

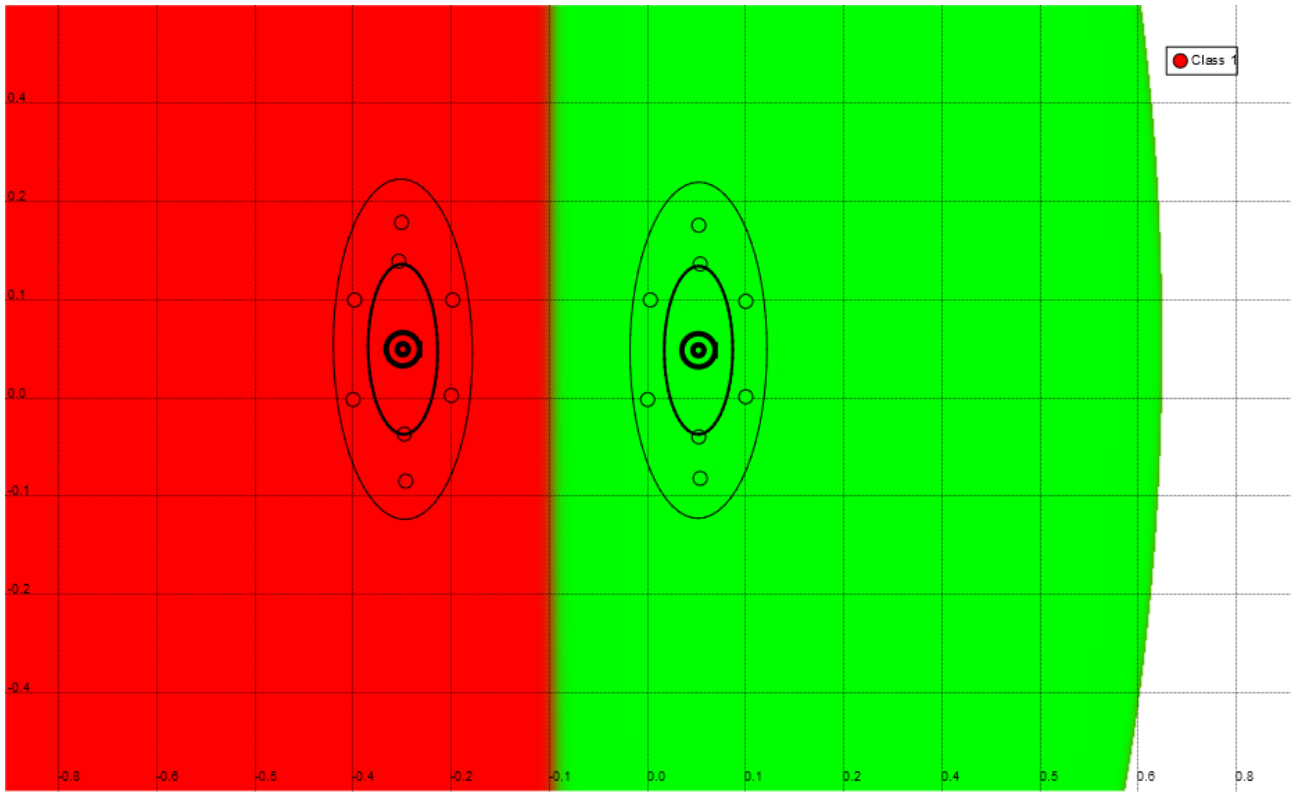
- Identical diagonal covariance matrices.



Q1.A > b

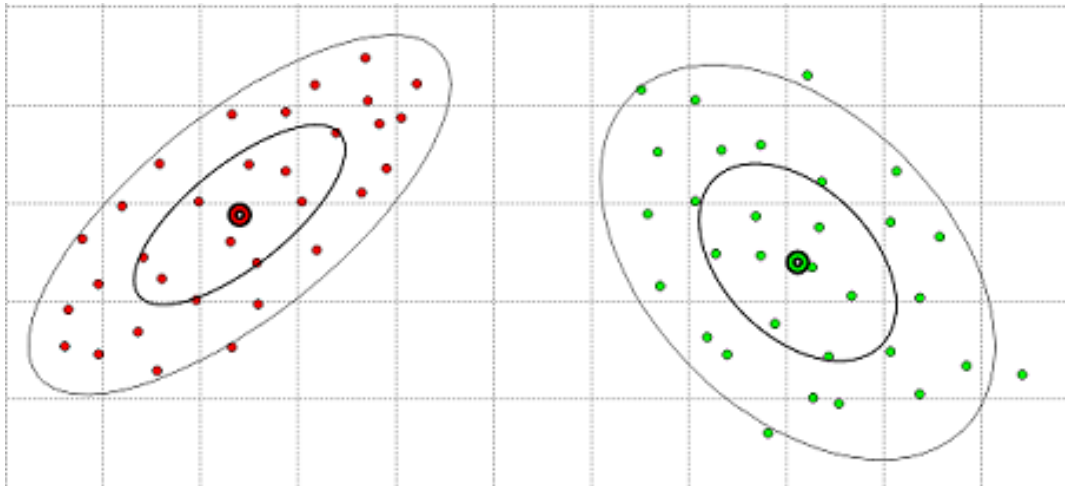
- Identical diagonal covariance matrices.
- Linear boundary is equidistant from the two centroids.





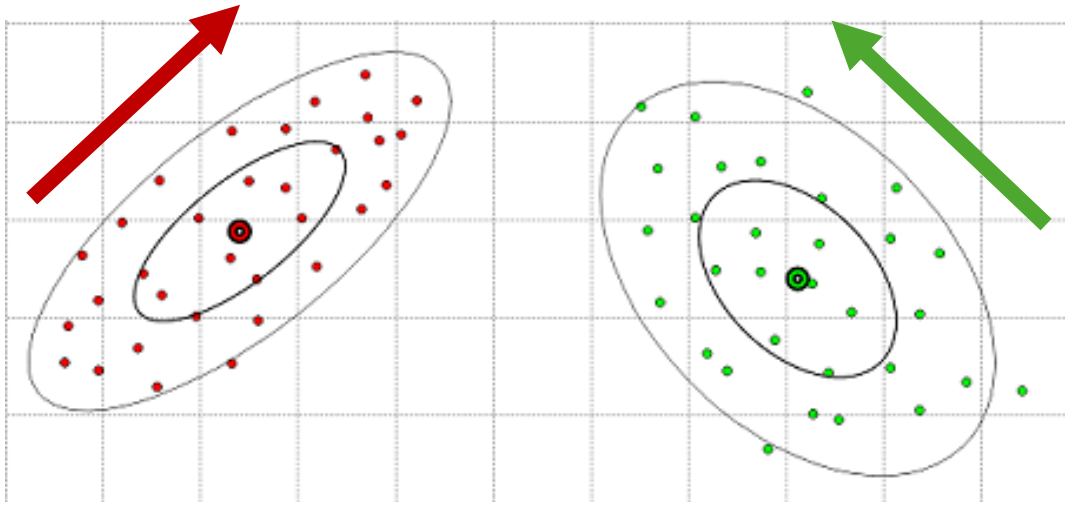
Q1.A > c

- Full covariance matrices.



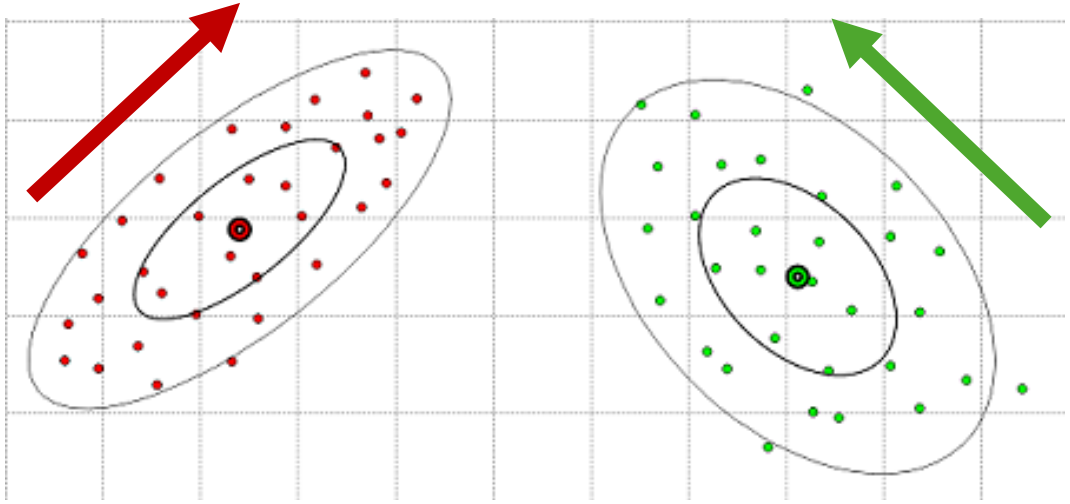
Q1.A > c

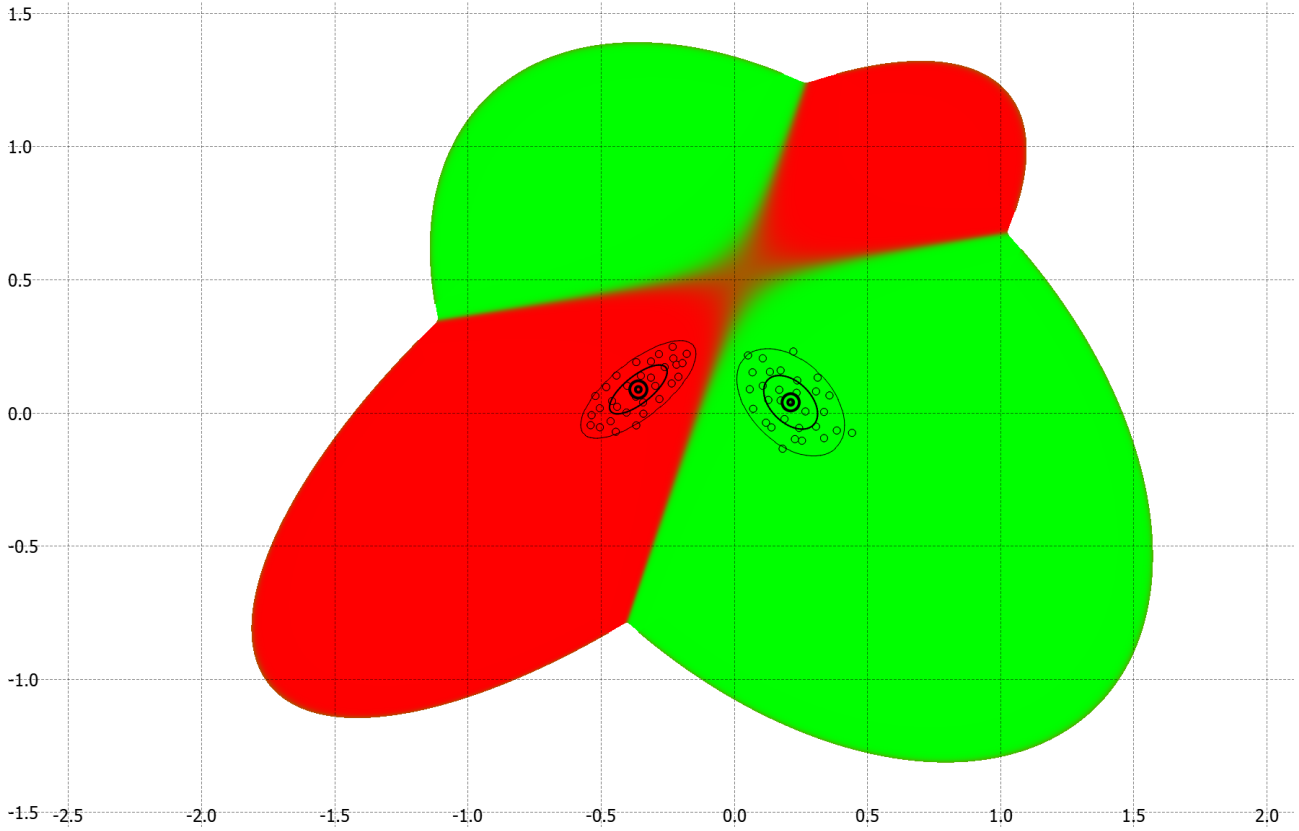
- Full covariance matrices.
- Notice the direction of spread (variance) for each Gaussian.



Q1.A > c

- Full covariance matrices.
- Notice the direction of spread (variance) for each Gaussian.





Q1.B

- Studying the effect of the number of datapoints per class on the classification boundary.
  - Two classes, each fitted with a single Gauss function.
  - Two classes have identical covariance matrices.

$$-\frac{1}{2}(x-\mu^1)^\top\Sigma^{-1}(x-\mu^1)+\ln p(y=1) = -\frac{1}{2}(x-\mu^2)^\top\Sigma^{-1}(x-\mu^2)+\ln p(y=2)$$

Q1.B

$$-\frac{1}{2}(x-\mu^1)^\top\Sigma^{-1}(x-\mu^1)+\ln p(y=1) = -\frac{1}{2}(x-\mu^2)^\top\Sigma^{-1}(x-\mu^2)+\ln p(y=2)$$

Q1.B

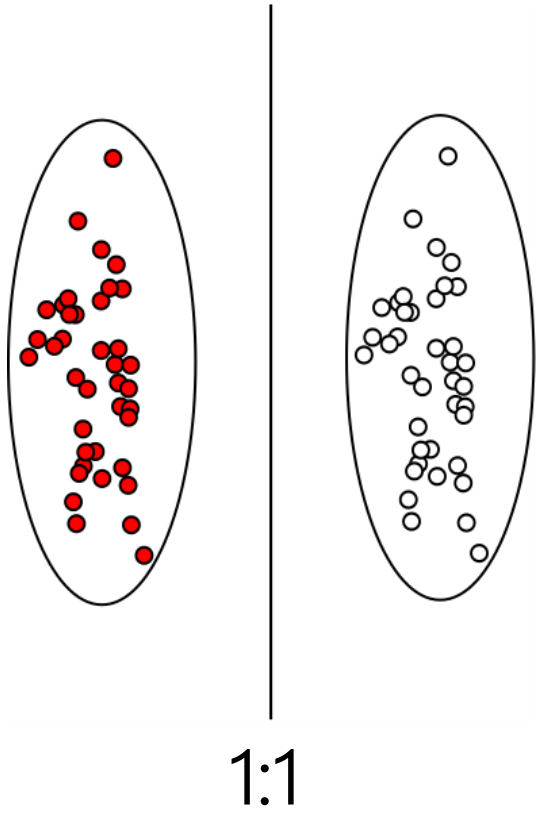
$$-\frac{1}{2}(x-\mu^1)^\top\Sigma^{-1}(x-\mu^1)+\ln p(y=1) = -\frac{1}{2}(x-\mu^2)^\top\Sigma^{-1}(x-\mu^2)+\ln p(y=2)$$

Q1.B

$$-\frac{1}{2}(x-\mu^1)^\top\Sigma^{-1}(x-\mu^1)+\ln p(y=1)=-\frac{1}{2}(x-\mu^2)^\top\Sigma^{-1}(x-\mu^2)+\ln p(y=2)$$

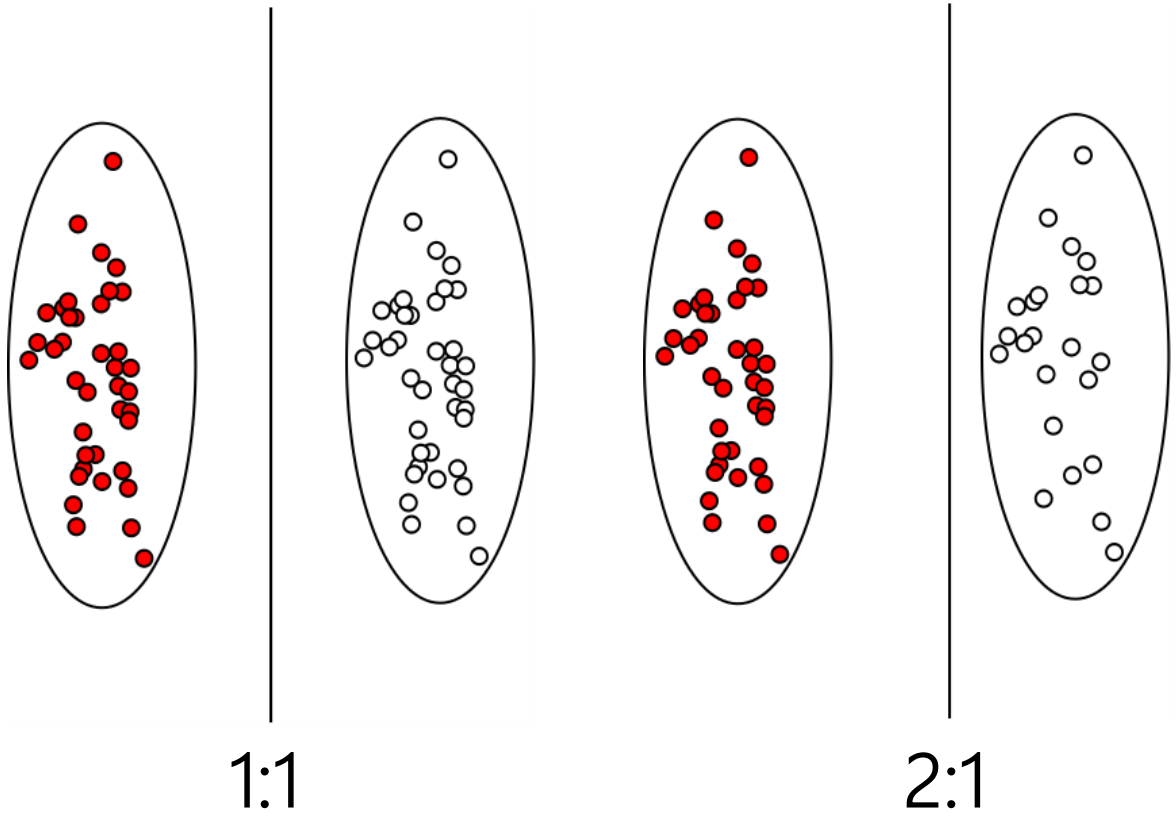
Q1.B

$$-\frac{1}{2}(x-\mu^1)^\top\Sigma^{-1}(x-\mu^1)+\ln p(y=1)=-\frac{1}{2}(x-\mu^2)^\top\Sigma^{-1}(x-\mu^2)+\ln p(y=2)$$



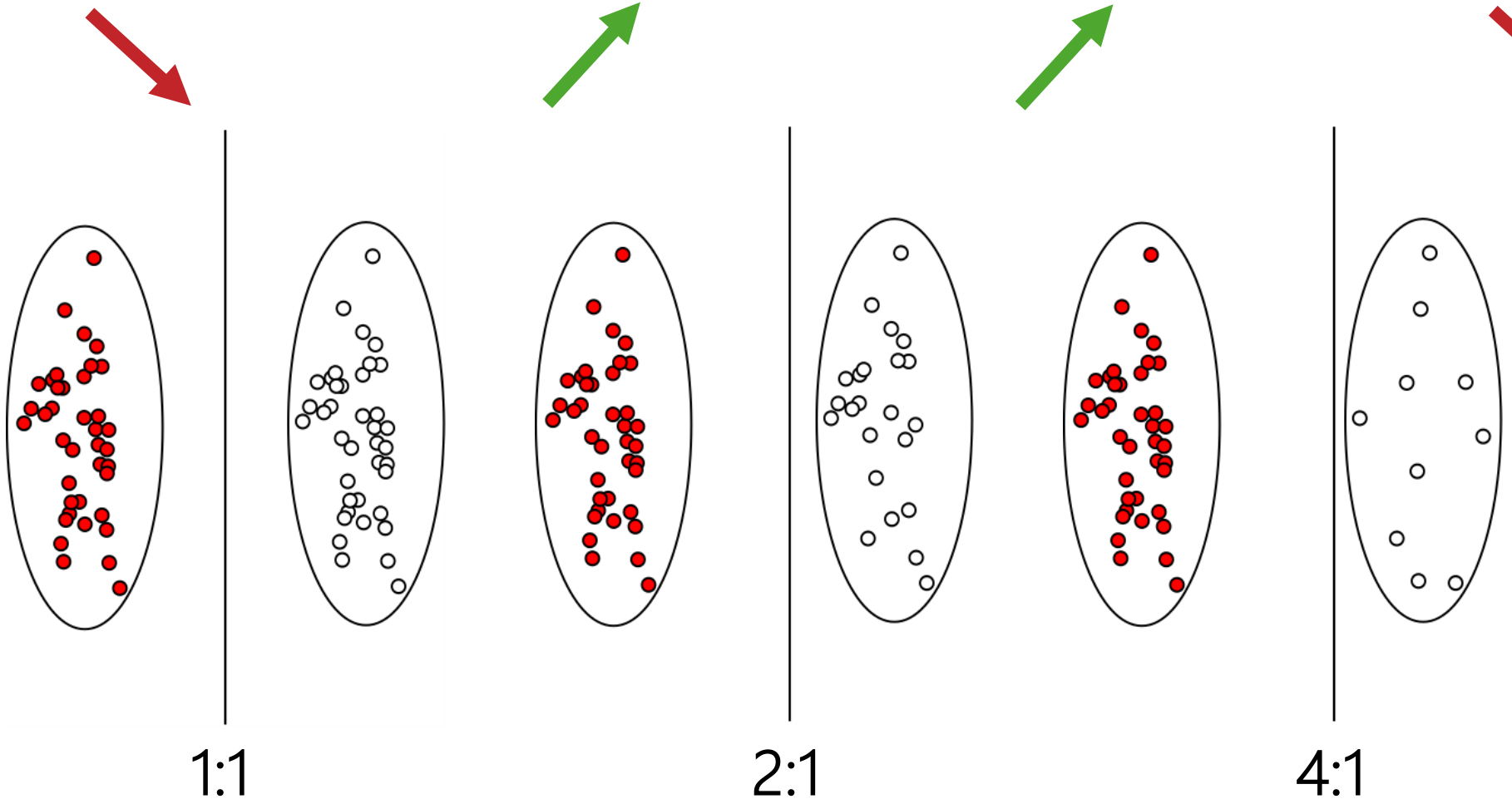
Q1.B

$$-\frac{1}{2}(x-\mu^1)^\top\Sigma^{-1}(x-\mu^1)+\ln p(y=1)=-\frac{1}{2}(x-\mu^2)^\top\Sigma^{-1}(x-\mu^2)+\ln p(y=2)$$



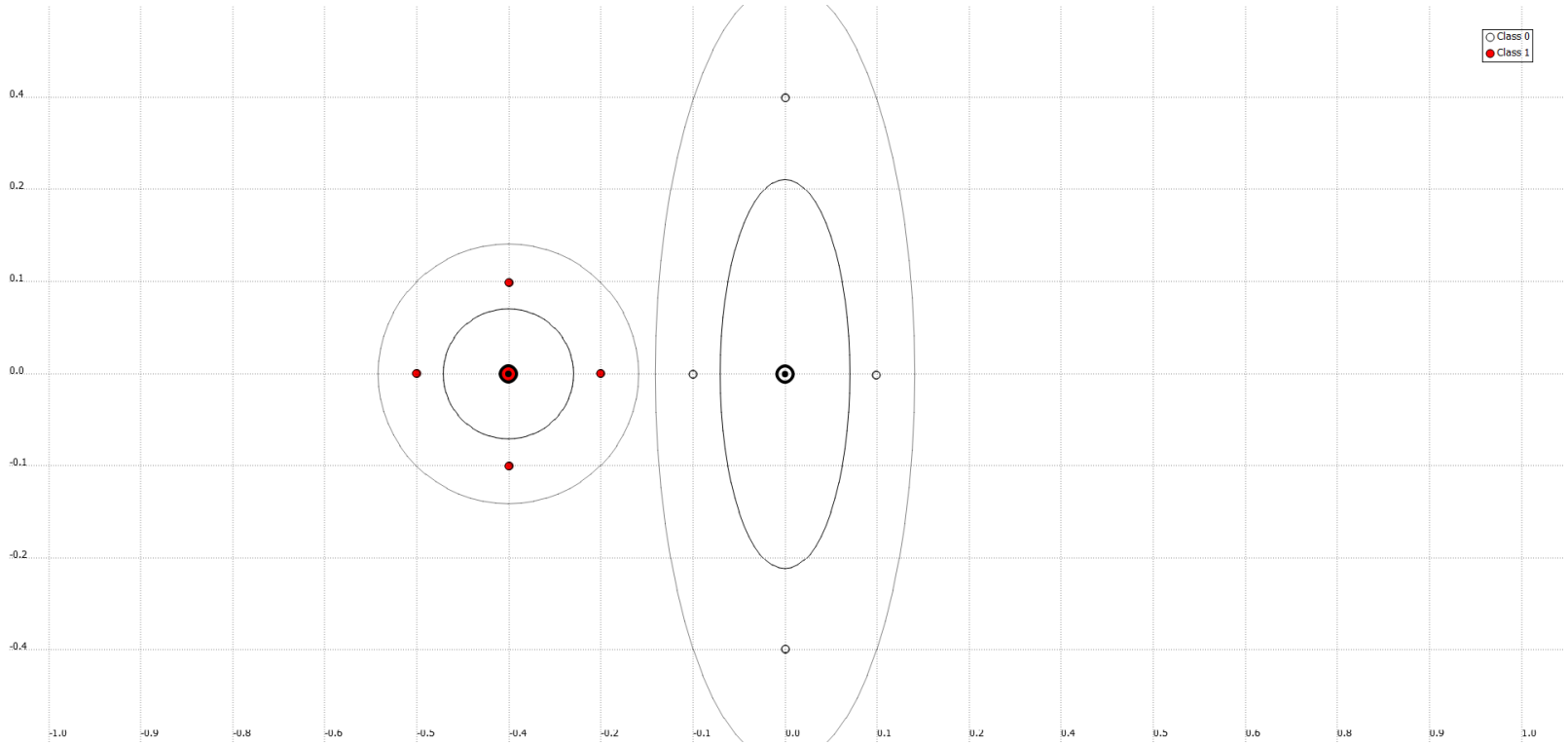
Q1.B

$$-\frac{1}{2}(x-\mu^1)^\top\Sigma^{-1}(x-\mu^1)+\ln p(y=1)=-\frac{1}{2}(x-\mu^2)^\top\Sigma^{-1}(x-\mu^2)+\ln p(y=2)$$



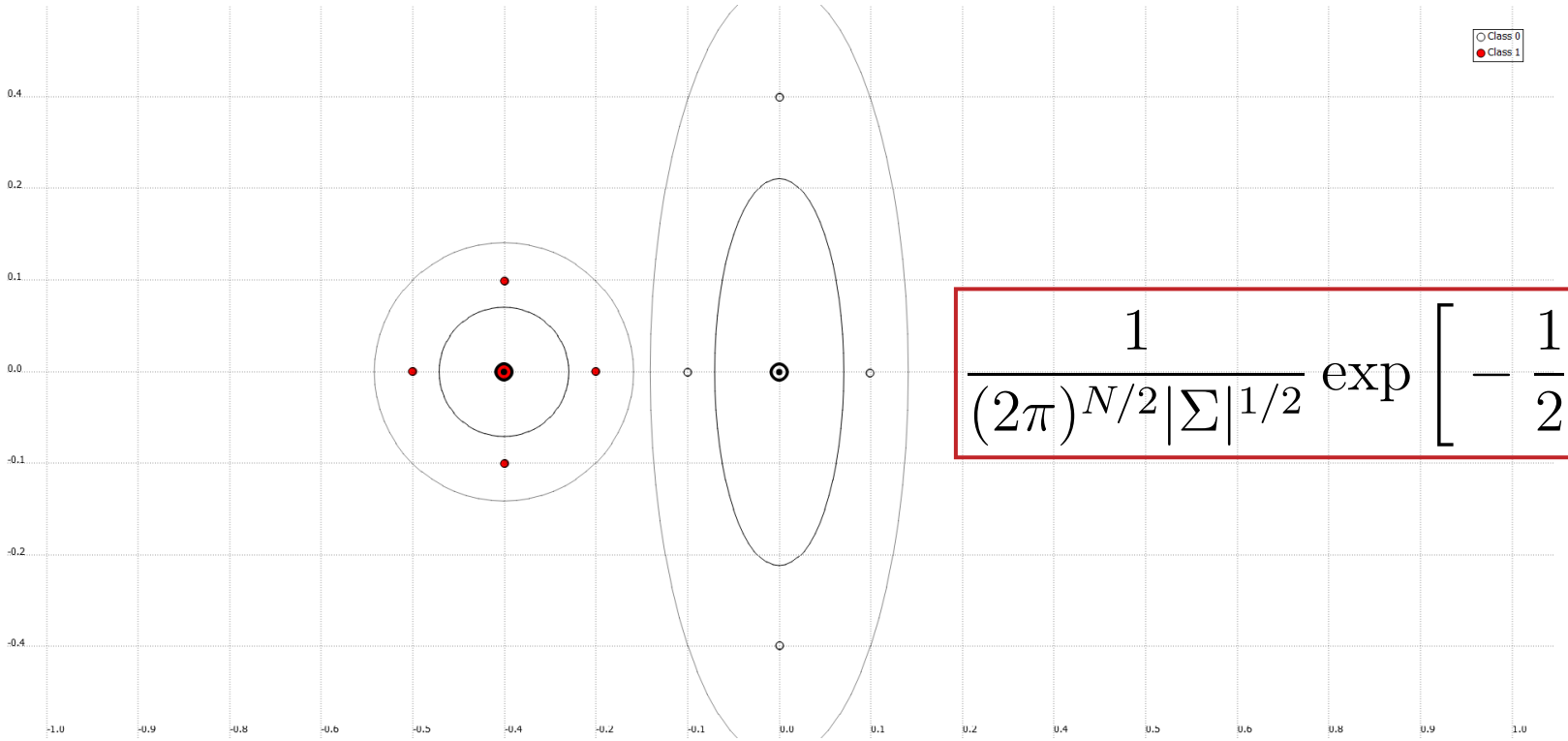
Q1.C

- Drawing the classification boundary when the Gauss functions are *normalized* and the covariance matrices are different.
  - Both Gaussian distributions have the same variance along the X-axis.



Q1.C

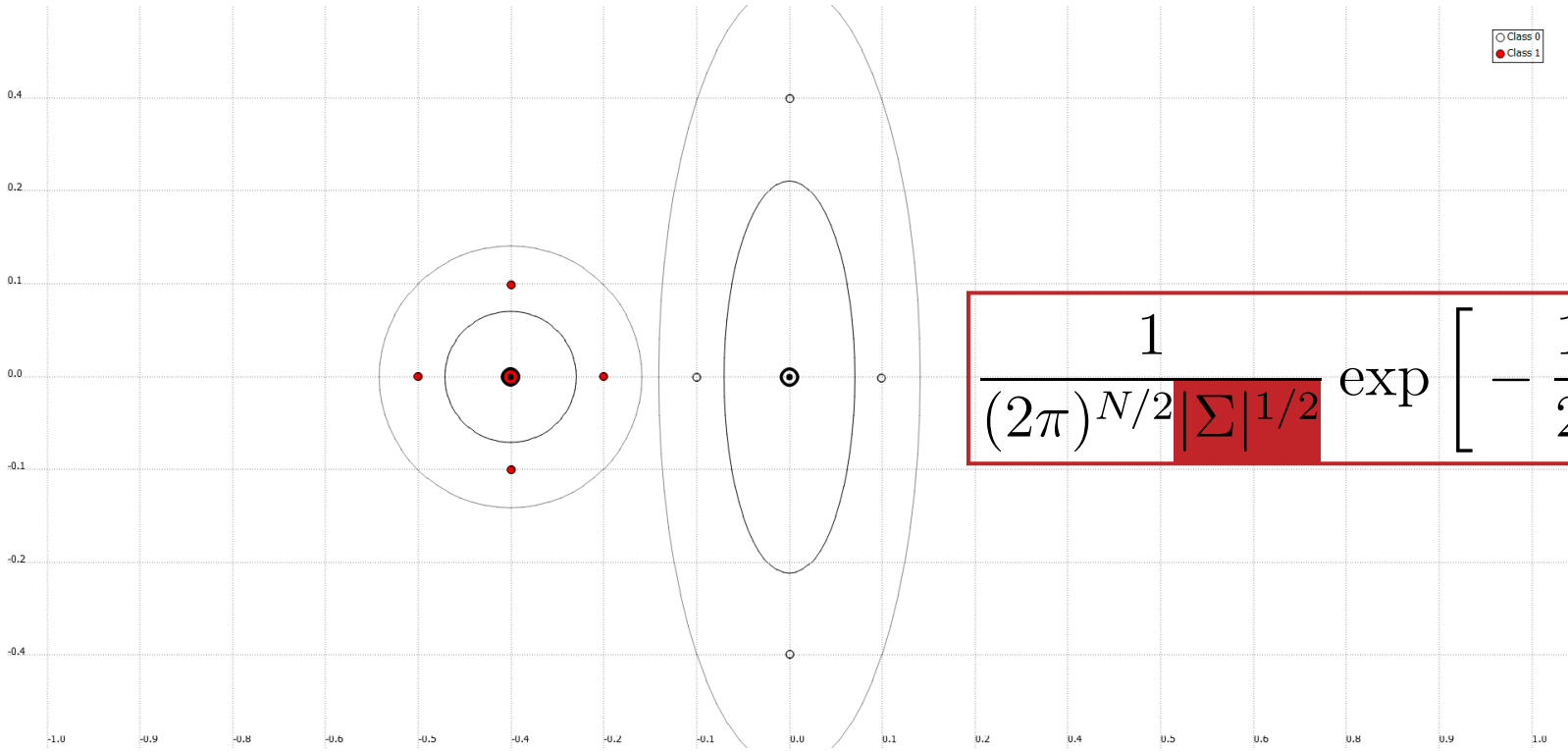
- Drawing the classification boundary when the Gauss functions are *normalized* and the covariance matrices are different.
  - Both Gaussian distributions have the same variance along the X-axis.



$$\frac{1}{(2\pi)^{N/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^{\top}\Sigma^{-1}(x-\mu)\right]$$

Q1.C

- Drawing the classification boundary when the Gauss functions are *normalized* and the covariance matrices are different.
  - Both Gaussian distributions have the same variance along the X-axis.



$$\frac{1}{(2\pi)^{N/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^{\top}\Sigma^{-1}(x-\mu)\right]$$

Q1.C

- Isolines no longer represent the same probability.

Q1.C

- Isolines no longer represent the same probability.

$$\frac{p(x|y=1)}{p(x|y=2)} \times \frac{p(y=1)}{p(y=2)} = 1$$

Q1.C

- Isolines no longer represent the same probability.

$$\frac{p(x|y=1)}{p(x|y=2)} \times \frac{p(y=1)}{p(y=2)} = 1$$

Q1.C

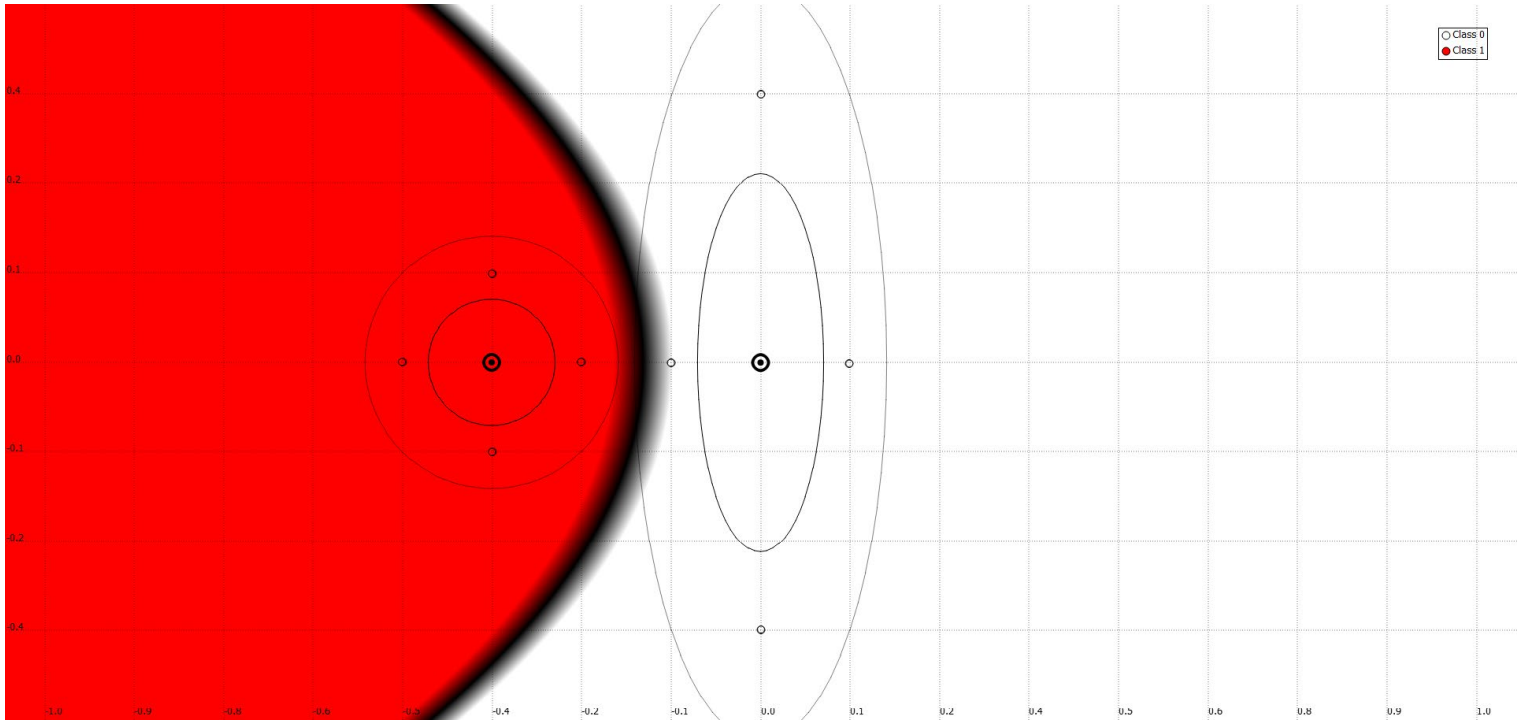
- Isolines no longer represent the same probability.

$$\frac{p(x|y=1)}{p(x|y=2)} \times \frac{p(y=1)}{p(y=2)} = 1$$

Q1.C

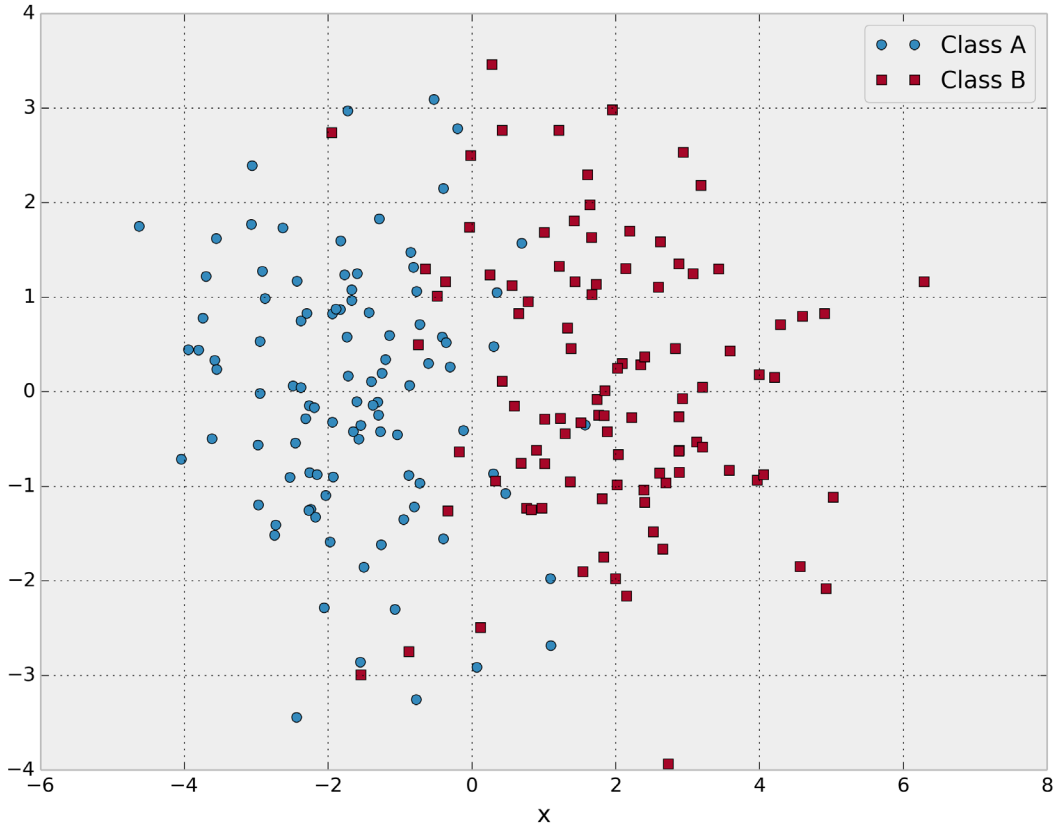
- Isolines no longer represent the same probability.

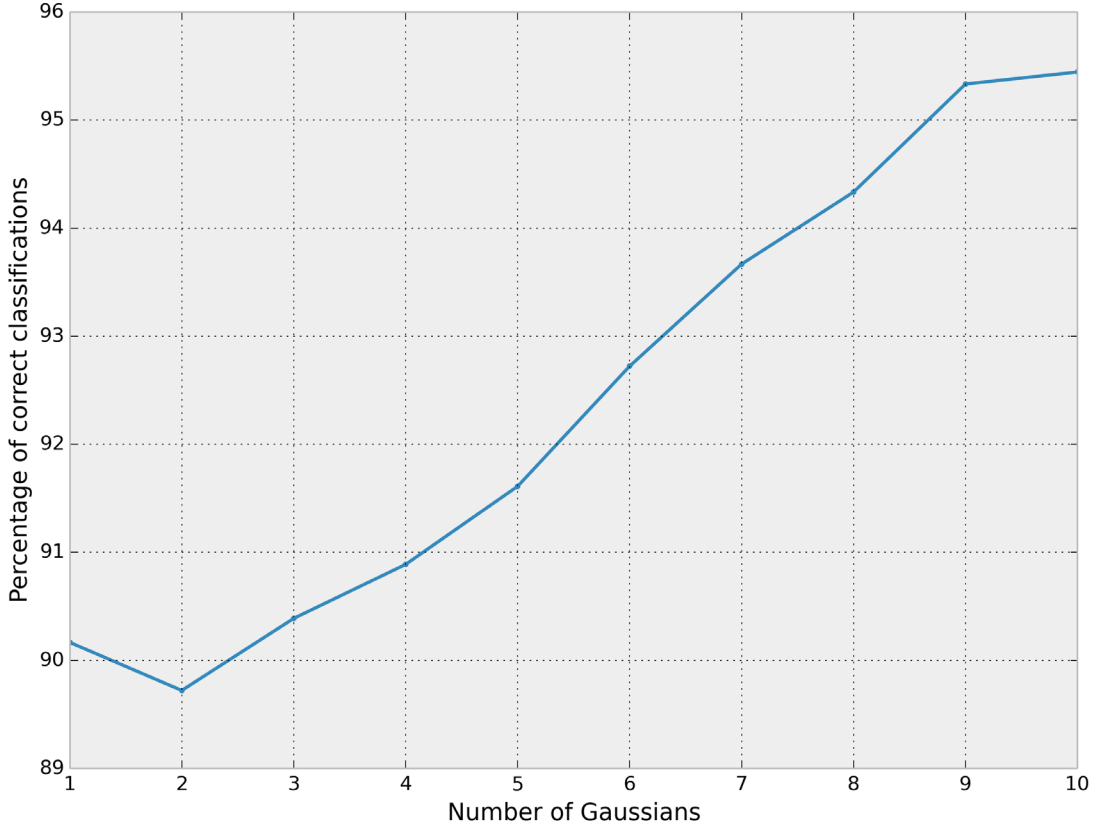
$$\frac{p(x|y=1)}{p(x|y=2)} \times \frac{p(y=1)}{p(y=2)} = 1$$



Q2.A

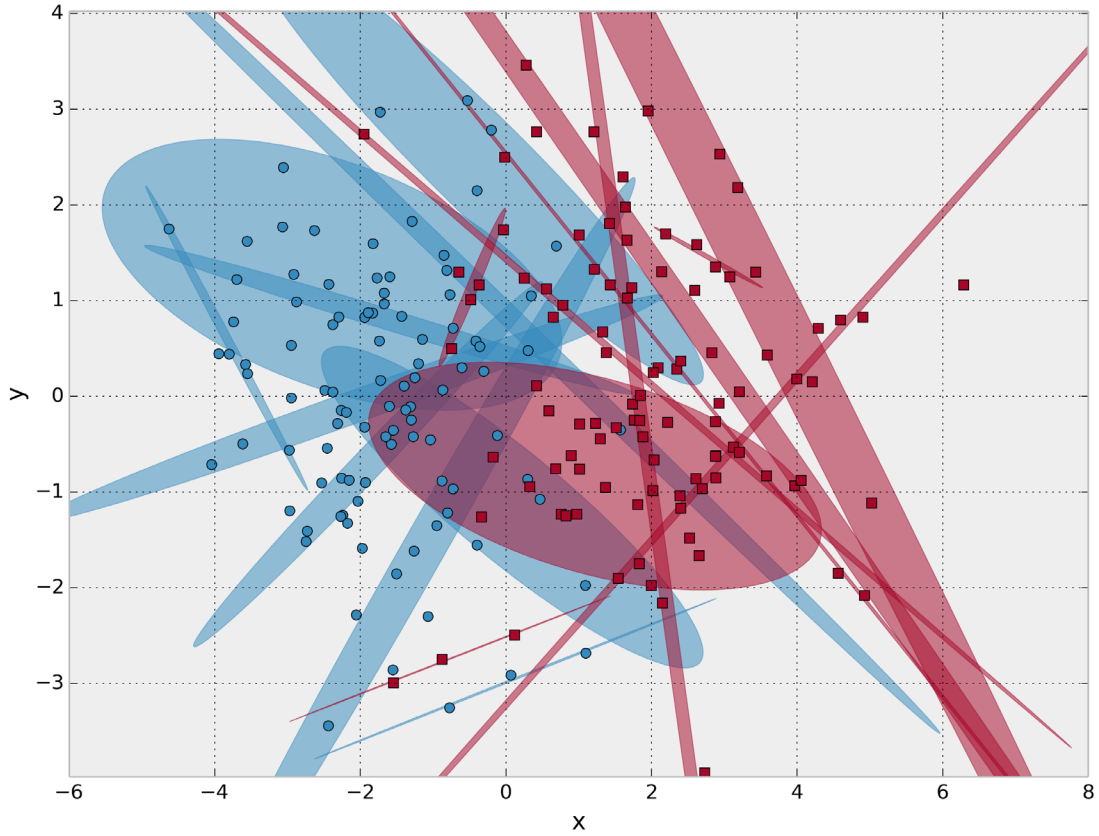
- Choosing the number of Gauss functions in GMM based on the performance of the classifier.





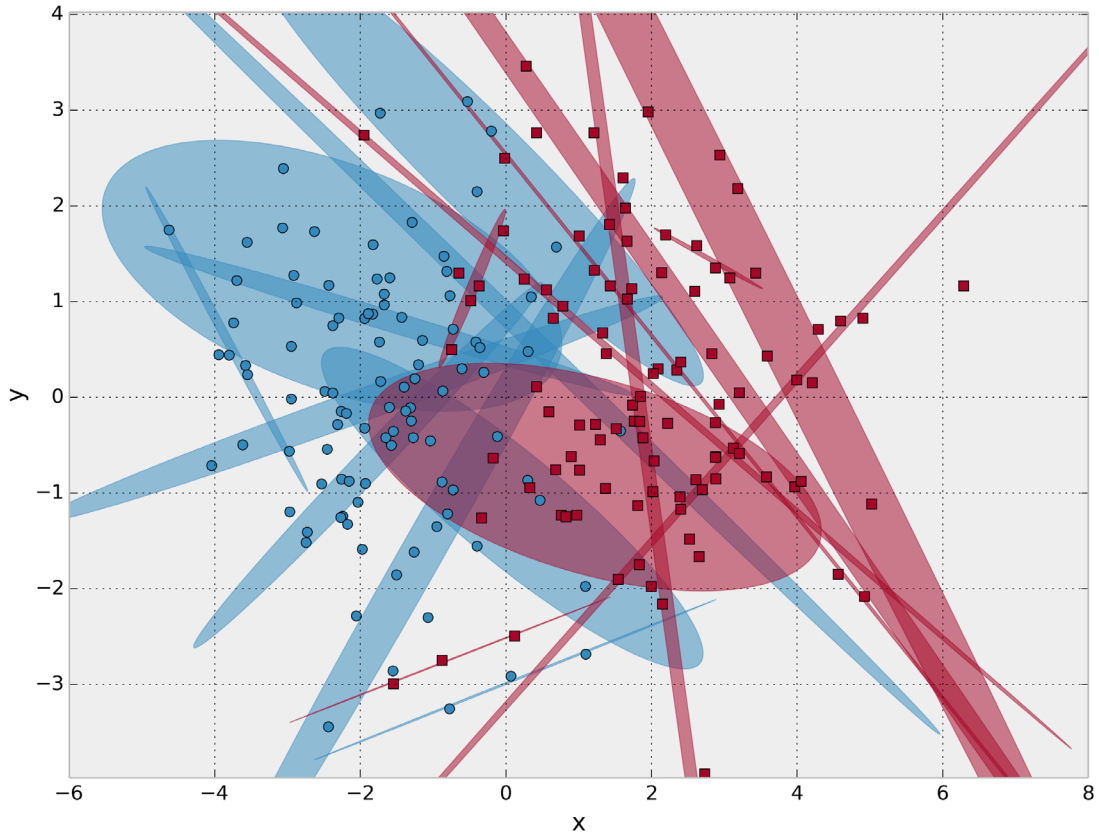
Q2.A

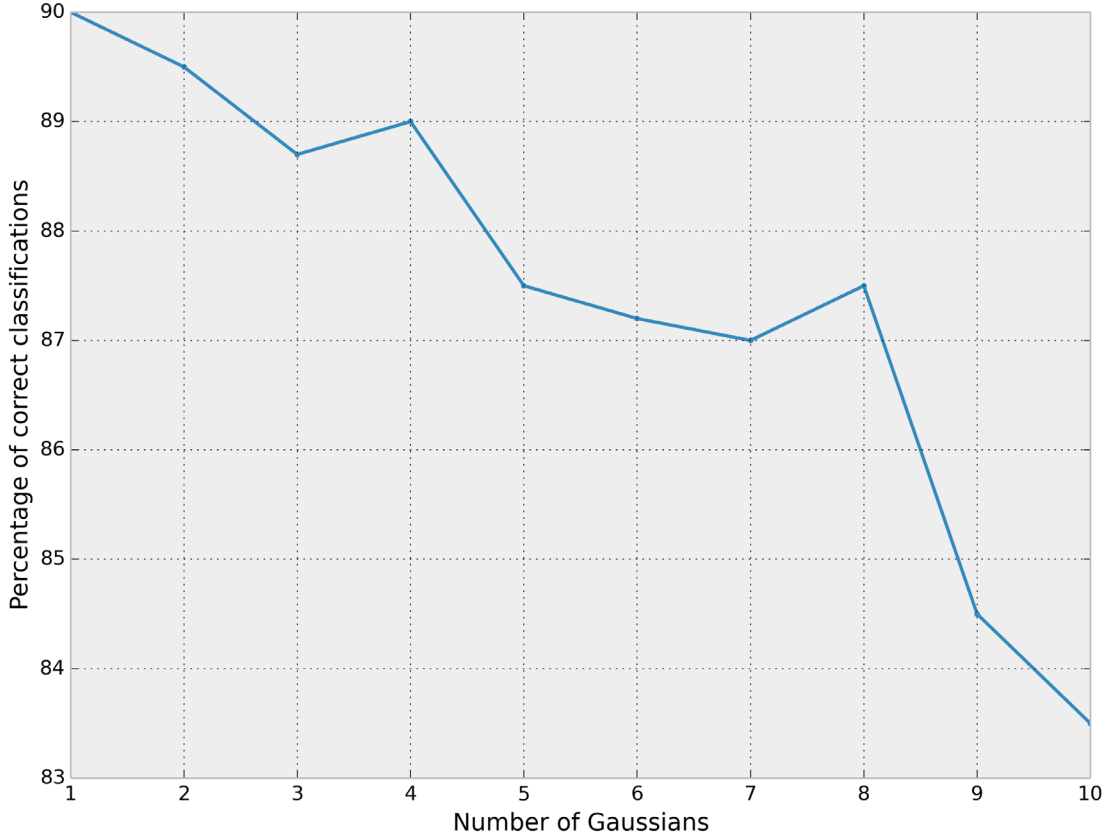
- Risk of overfitting.



Q2.A

- Risk of overfitting.
- Proper usage of K-fold cross-validation.



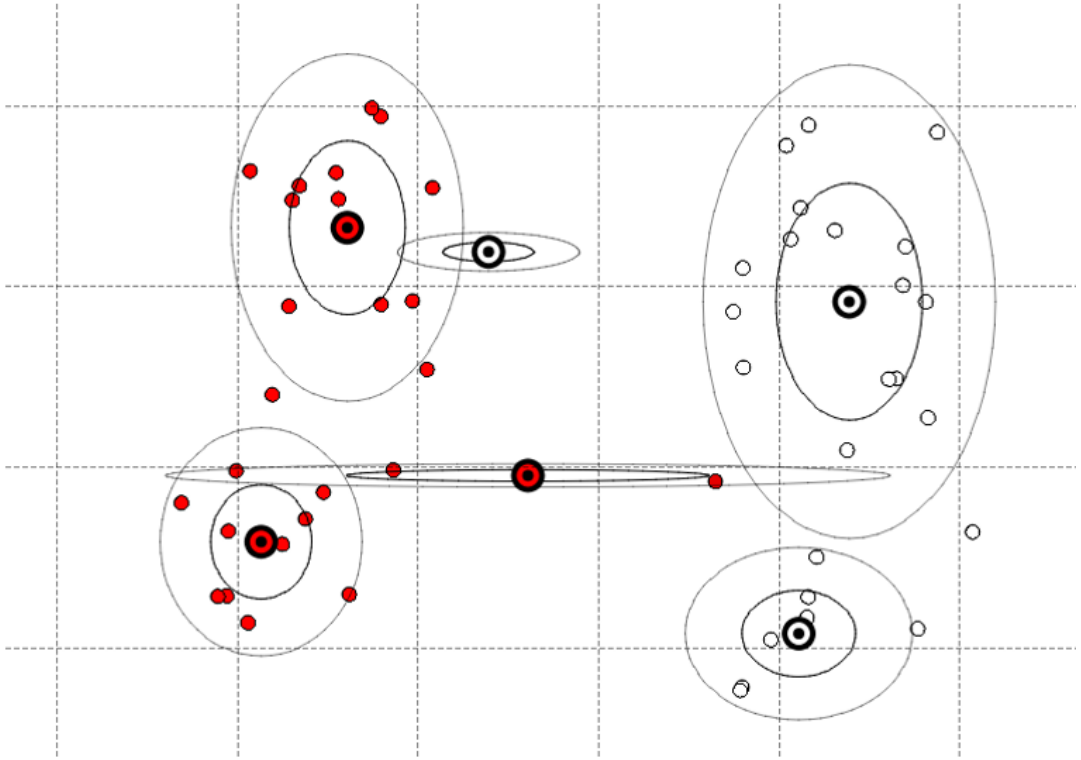


Q2.B

- An example dataset and choice of GMM that leads to overfitting.

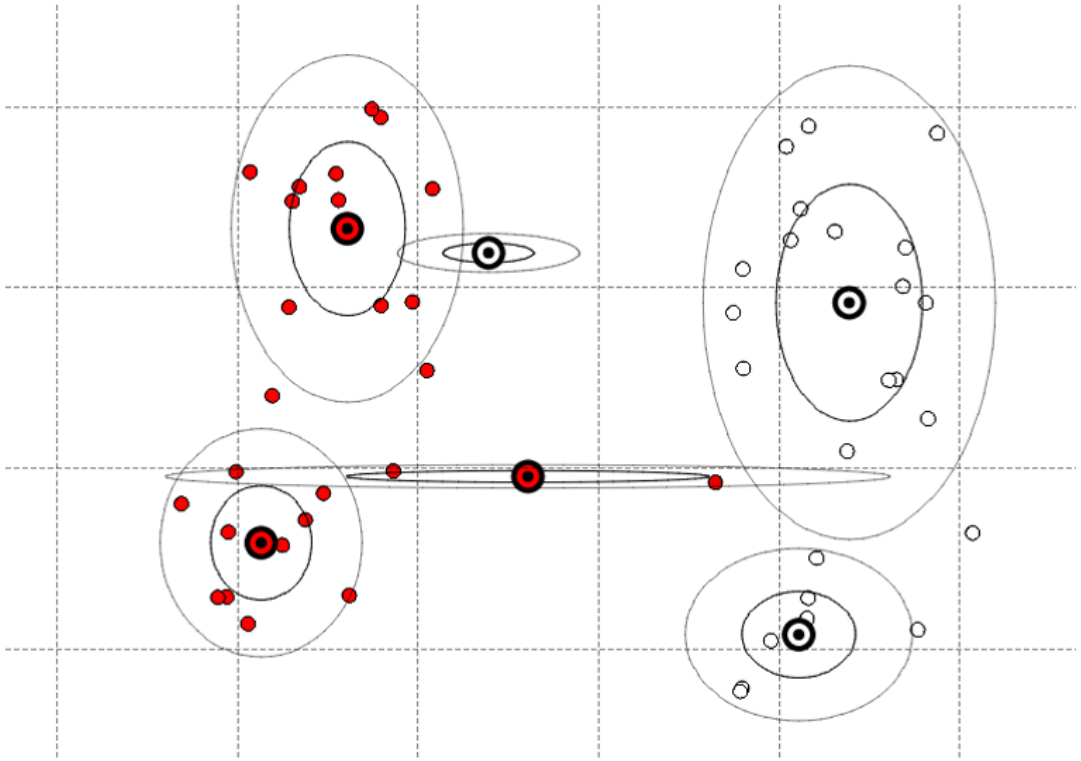
Q2.B

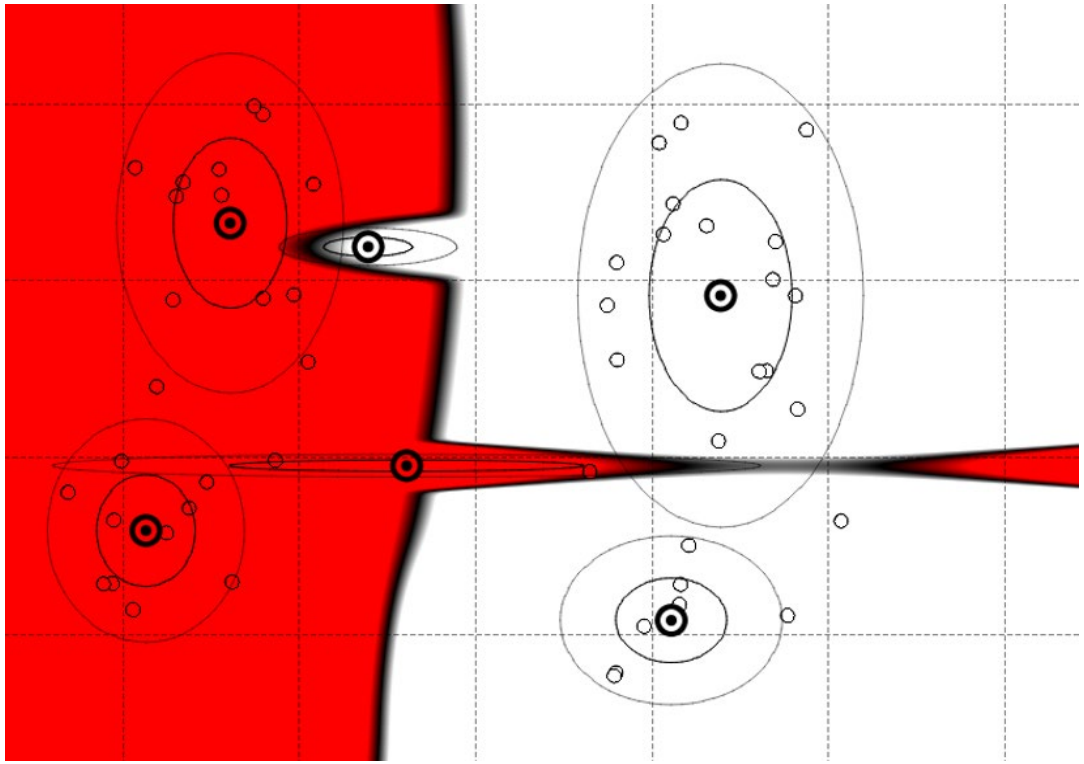
- An example dataset and choice of GMM that leads to overfitting.



Q2.B

- An example dataset and choice of GMM that leads to overfitting.





Q2.C

- What is the number of parameters to be estimated for GMM?
- What is the computational cost per iteration of the update step for GMM?

Q2.C > Number of parameters to estimate

- Number of Gaussians: K
- Data Dimension: N

Q2.C > Number of parameters to estimate

- Number of Gaussians: K
- Data Dimension: N
- Full covariance matrix:

$$K\left(\frac{N(N+1)}{2}+N\right)+K-1$$

Q2.C > Number of parameters to estimate

- Number of Gaussians: K
- Data Dimension: N
- Full covariance matrix:

$$K\left(\frac{N(N+1)}{2}+N\right)+K-1$$

covariance matrix

Q2.C > Number of parameters to estimate

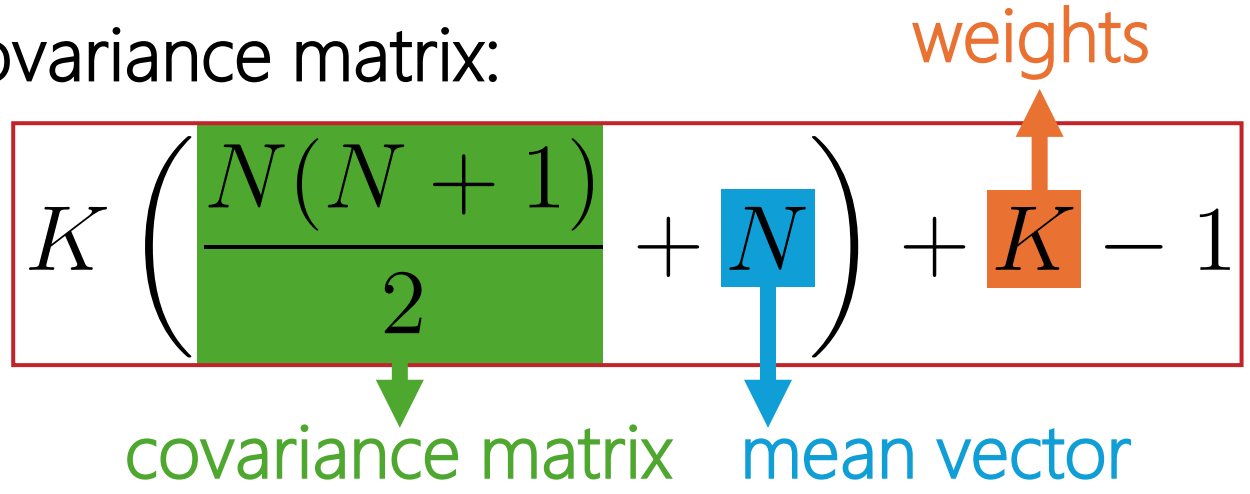
- Number of Gaussians: K
- Data Dimension: N
- Full covariance matrix:

$$K\left(\frac{N(N+1)}{2}+N\right)+K-1$$

covariance matrix mean vector

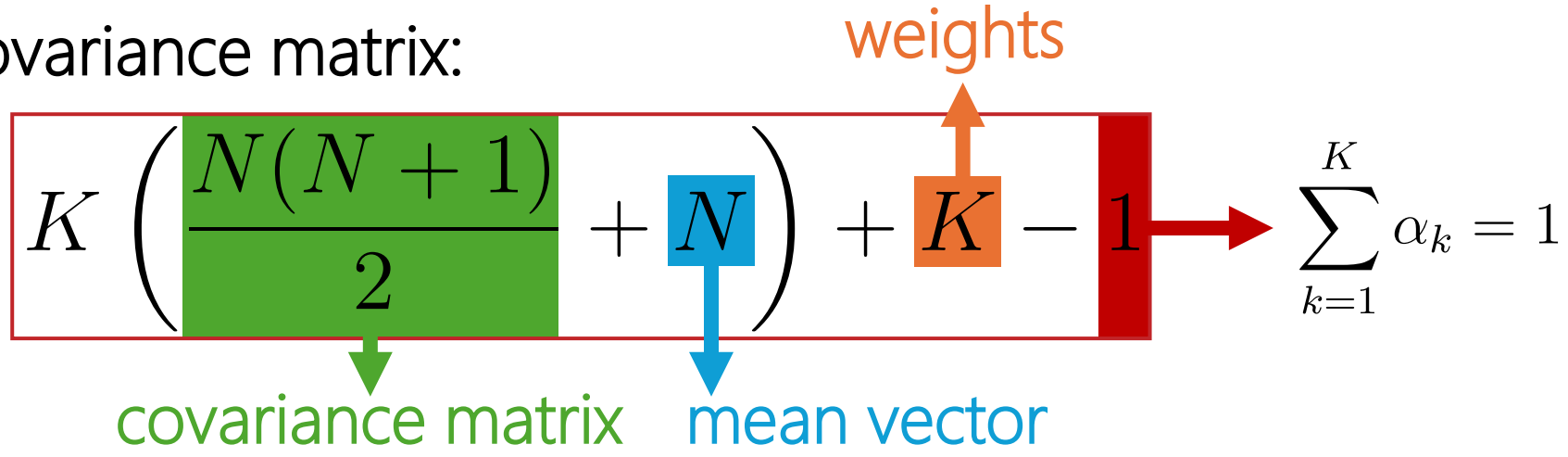
Q2.C > Number of parameters to estimate

- Number of Gaussians: K
- Data Dimension: N
- Full covariance matrix:



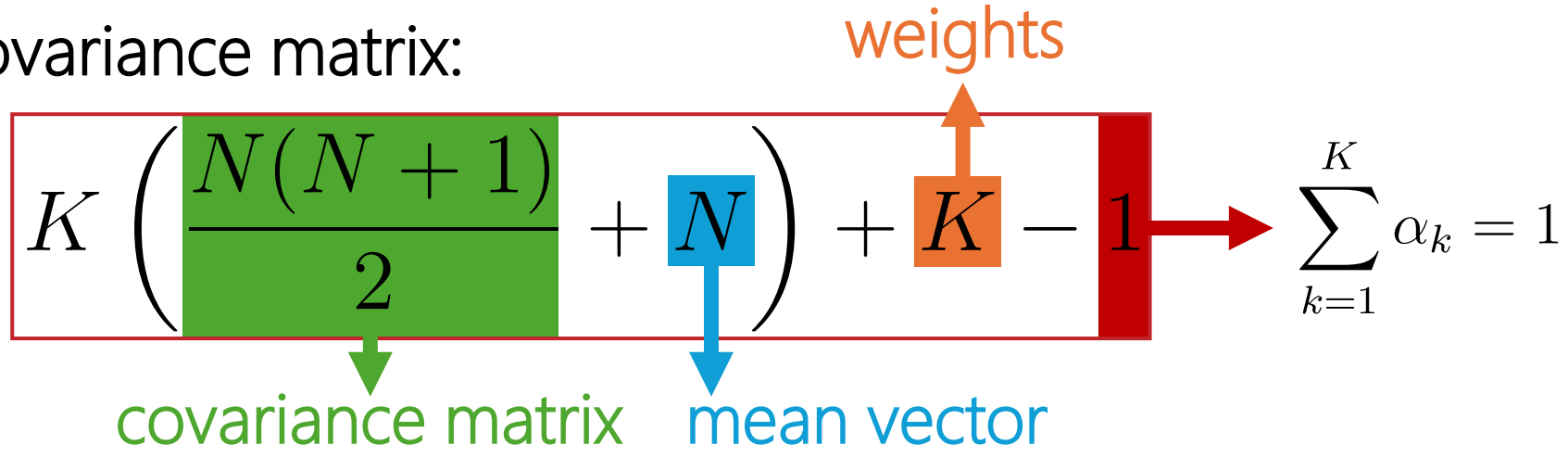
Q2.C > Number of parameters to estimate

- Number of Gaussians: K
- Data Dimension: N
- Full covariance matrix:



Q2.C > Number of parameters to estimate

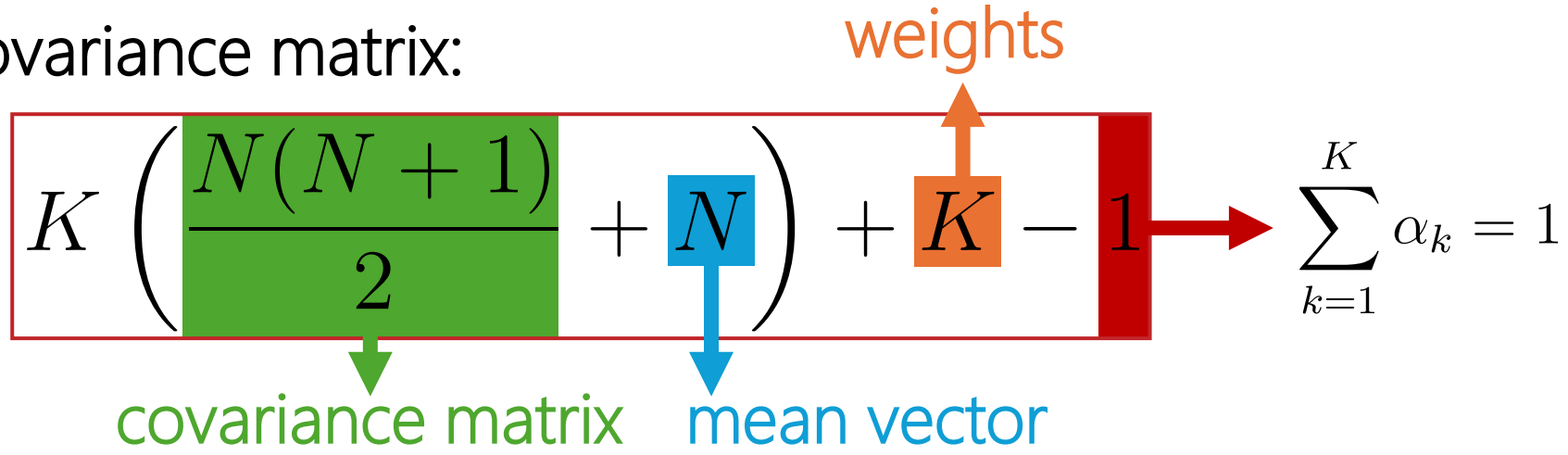
- Number of Gaussians: K
- Data Dimension: N
- Full covariance matrix:



- Diagonal covariance matrix: $K(N+N)+K-1$

Q2.C > Number of parameters to estimate

- Number of Gaussians: K
- Data Dimension: N
- Full covariance matrix:



- Diagonal covariance matrix: $K(N+N)+K-1$
- Spherical covariance matrix: $K(1+N)+K-1$

Q2.C > Computational cost of the update step

- Update of the covariance matrices.

Q2.C > Computational cost of the update step

- Update of the covariance matrices.

$$\Sigma_{\text{full}}^{k\ (t+1)} = \frac{\sum_j p(k|x^j, \Theta^{(t)})(x^j - \mu^{k(t+1)})(x^j - \mu^{k(t+1)})^\top}{\sum_j p(k|x^j, \Theta^{(t)})}$$

$$\Sigma_{\text{dia}}^{k\ (t+1)} = \text{diag}((\sigma_1^{k(t+1)})^2, \ldots, (\sigma_N^{k\ (t+1)})^2)$$

$$\Sigma_{\text{iso}}^{k\ (t+1)} = \text{diag}((\sigma^{k(t+1)})^2)$$

where

$$(\sigma_i^{k(t+1)})^2 = \frac{\sum_j p(k|x^j, \Theta^{(t)})(x_i^j - \mu_i^{k(t+1)})^2}{\sum_j p(k|x^j, \Theta^{(t)})} \quad \forall i = 1, \ldots, N$$

are the variances along each dimension i and

$$(\sigma^{k(t+1)})^2 = \frac{\sum_j p(k|x^j, \Theta^{(t)}) \|x^j - \mu^{k(t+1)}\|^2}{N \sum_j p(k|x^j, \Theta^{(t)})}$$

is the variance averaged over all samples.

Q2.C > Computational cost of the update step

- Update of the covariance matrices.

$$\Sigma_{\text{full}}^{k\ (t+1)} = \frac{\sum_j p(k|x^j, \Theta^{(t)})(x^j - \mu^{k(t+1)})(x^j - \mu^{k(t+1)})^\top}{\sum_j p(k|x^j, \Theta^{(t)})}$$

$$\Sigma_{\text{dia}}^{k\ (t+1)} = \text{diag}((\sigma_1^{k(t+1)})^2, \ldots, (\sigma_N^{k\ (t+1)})^2)$$

$$\Sigma_{\text{iso}}^{k\ (t+1)} = \text{diag}((\sigma^{k(t+1)})^2)$$

where

$$(\sigma_i^{k(t+1)})^2 = \frac{\sum_j p(k|x^j, \Theta^{(t)})(x_i^j - \mu_i^{k(t+1)})^2}{\sum_j p(k|x^j, \Theta^{(t)})} \quad \forall i = 1, \ldots, N$$

are the variances along each dimension i and

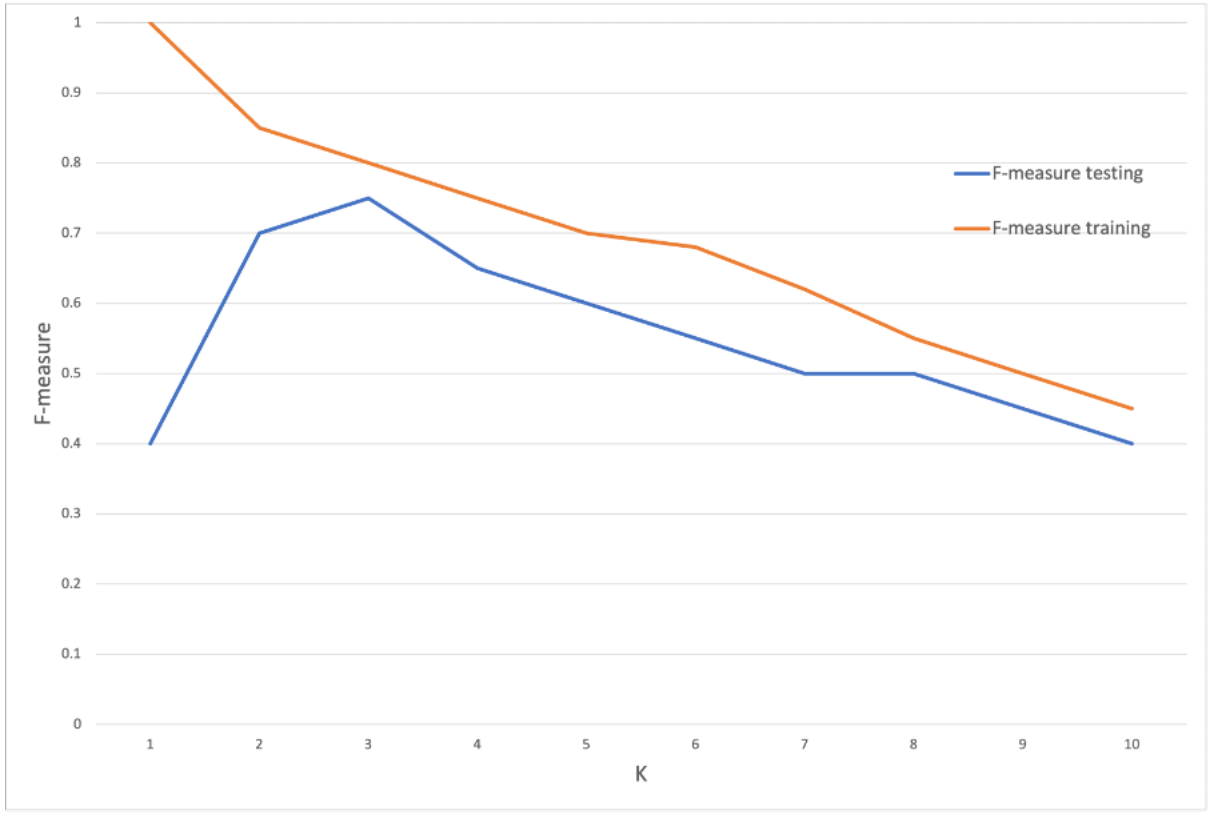
$$(\sigma^{k(t+1)})^2 = \frac{\sum_j p(k|x^j, \Theta^{(t)}) \|x^j - \mu^{k(t+1)}\|^2}{N \sum_j p(k|x^j, \Theta^{(t)})}$$

is the variance averaged over all samples.

Type	Complexity
Spherical	$\mathcal{O}(KM)$
Diagonal	$\mathcal{O}(KMN)$
Full	$\mathcal{O}(KMN^2)$

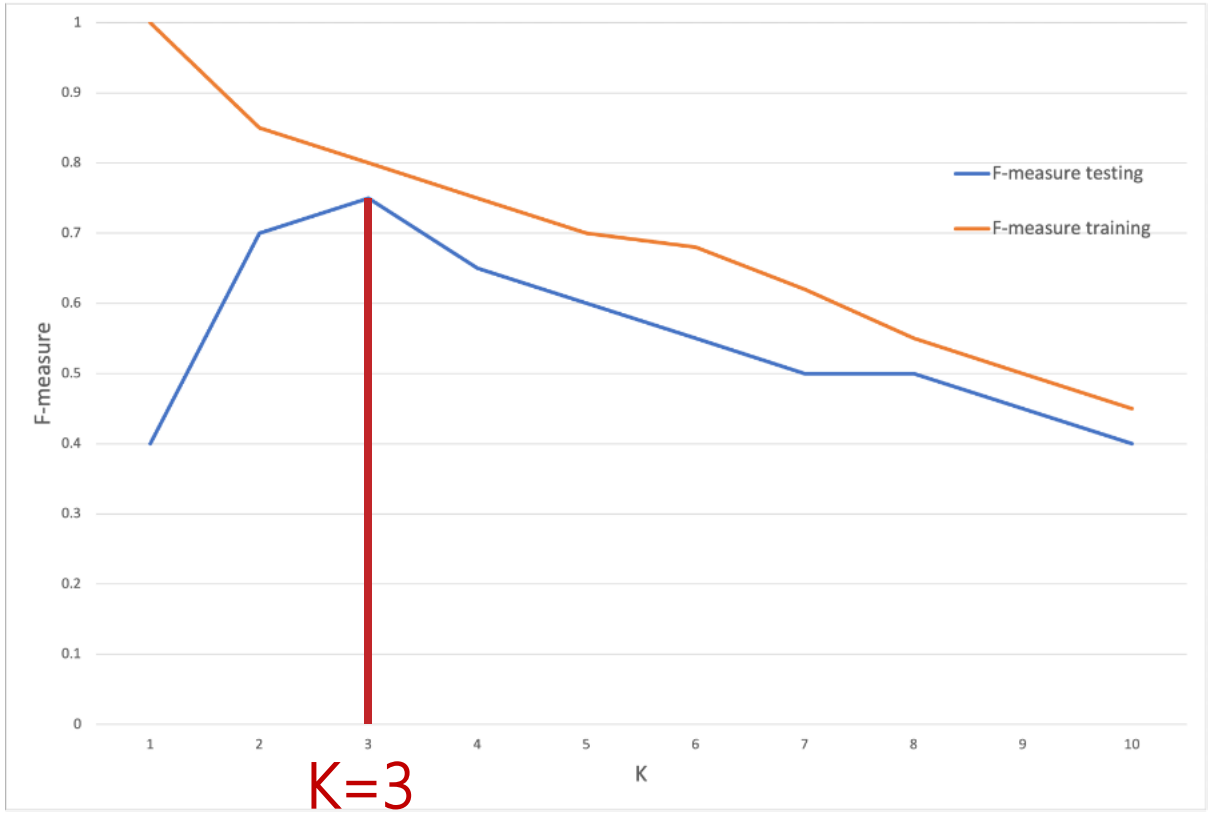
Q2.D

- Choosing hyperparameter K in KNN classification based on F-measure.



Q2.D

- Choosing hyperparameter K in KNN classification based on F-measure.



Q2.D

- Choosing hyperparameter K in KNN classification based on F-measure.