CS 189 Introduction to Machine Learning
Spring 2023 Jonathan Shewchuk

DIS4

1 Bayesian Decision Theory: Case Study

We want to design an automated fishing system that captures fish, classifies them, and sends them off to two different companies, Salmonites, Inc., and Seabass, Inc. For some reason we only ever catch salmon ($Y = 1$) and seabass ($Y = 2$). Salmonites, Inc. wants salmon, and Seabass, Inc. wants seabass. Given only the weights of the fish we catch, we want to figure out what type of fish it is using machine learning!

Let us assume that the weight of both seabass and salmon are both normally distributed (univariate Gaussian), given by the p.d.f.

$$P(x|Y = i) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(x-\mu_i)^2/(2\sigma_i^2)}.$$

12 fish are randomly selected from our system and have the following weights.

Data for salmon: $\{3, 4, 5, 6, 7\}$.

Data for seabass: $\{5, 6, 7, 8, 9, 7 + \sqrt{2}, 7 - \sqrt{2}\}$.

When we classify seabass incorrectly, it gets sent to Salmonites, Inc. who won't pay us for the wrong fish and sells it themselves. When we classify salmon incorrectly, it gets sent to SeaBass, Inc., who is nice and returns our fish. This situation gives rise to this loss matrix.

Predicted:

Truth:		**salmon**	**seabass**
	salmon	0	1
	seabass	2	0

(a) First, compute the sample mean $\hat{\mu}_i$ and variance $\hat{\sigma}_i^2$ for the univariate Gaussian for both the seabass class and the salmon class. Also compute the empirical estimates of the priors $\hat{\pi}_i$.

$\hat{\mu}_1 =$
$\hat{\sigma}_1^2 =$
$\hat{\pi}_1 =$

$\hat{\mu}_2 =$
$\hat{\sigma}_2^2 =$
$\hat{\pi}_2 =$

(b) What is significant about $\hat{\sigma_1}$ and $\hat{\sigma_2}$?

(c) Next, find the decision rule when assuming a 0-1 loss function. Recall that a decision rule for the 0-1 loss function will minimize the probability of error.

(d) Now, find the decision rule using the loss matrix above. Recall that a decision rule, in general, minimizes the risk, or expected loss.

1. $X \in \mathbb{R}$ – weight of fish
   $Y \in \{1, 2\}$ – label of fish

$$P(X|Y=i) = \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right)$$

$N = 12$ – number of samples
Data w/ $Y=1$: $\{3, 4, 5, 6, 7\}$
Data w/ $Y=2$: $\{5, 6, 7, 8, 9, 7+\sqrt{2}, 7-\sqrt{2}\}$

Loss :

	$\hat{Y}=1$	$\hat{Y}=2$
$Y=1$	0	1
$Y=2$	2	0

a) Compute $\hat{\mu}_i$, $\hat{\sigma}_i^2$, π_i for $i = 1, 2$.

Let N_i = # of samples w/ label $Y = i$

$$\hat{\mu}_i = \frac{1}{N_i} \sum_{j : y_j = i} x_j$$

$$\hat{\sigma}_i^2 = \frac{1}{N_i} \sum_{j : y_j = i} (x_j - \hat{\mu}_i)^2$$

$$\hat{\pi}_i = N_i / N$$

$$\hat{\mu}_1 = \frac{1}{5}(3+4+5+6+7) = 25/5 = 5$$

$$\hat{\mu}_2 = \frac{1}{7}(5+6+7+8+9+7+\sqrt{2}+7-\sqrt{2}) = 49/7 = 7$$

$$\hat{\sigma}_1^2 = \frac{1}{5}\left[(3-5)^2 + (4-5)^2 + (5-5)^2 + (6-5)^2 + (7-5)^2\right]$$

$$= \frac{1}{5}(4+1+0+1+4) = 10/5 = 2$$

$$\hat{\sigma}_2^2 = \frac{1}{7}\left[(5-7)^2 + (6-7)^2 + (7-7)^2 + (8-7)^2 + (9-7)^2 + (7+\sqrt{2}-7)^2 + (7-\sqrt{2}-7)^2\right]$$

$$= \frac{1}{7}(4+1+0+1+4+2+2) = 14/7 = 2$$

$$\hat{\pi}_1 = 5/12$$

$$\hat{\pi}_2 = 7/12$$

b) Significance of $\hat{\sigma}_1$ & $\hat{\sigma}_2$ in context of Gaussian discriminant analysis?

$\hat{\sigma}_1^2 = \hat{\sigma}_2^2 = 2$

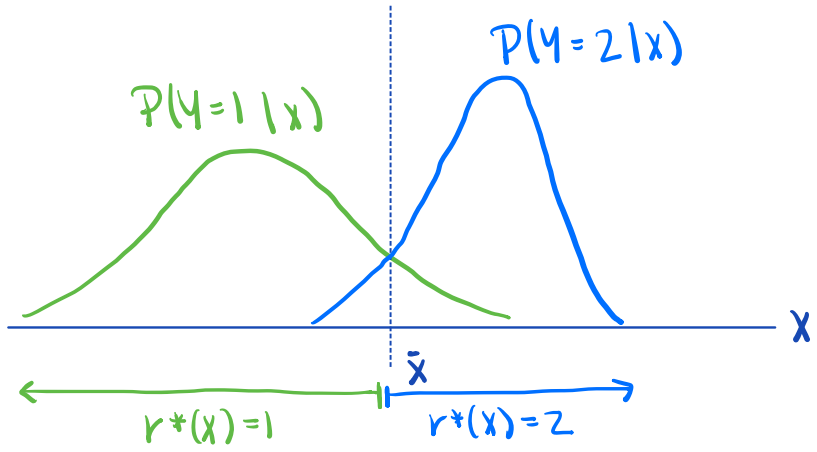
B/c the sample variances are equal for the two classes, QDA will produce the same results as LDA (i.e. the decision boundary will be linear).

C) Find QDA decision rule assuming 0-1 loss.

Bayes (optimal) decision rule:

$$r^*(x) = \underset{i \in \{1,2\}}{\operatorname{argmax}} P(Y=i \mid x)$$

e.g.



From Bayes rule,

$$P(Y=i \mid x) = \frac{P(x \mid Y=i)P(Y=i)}{P(x)}$$

$P(x)$ → doesn't depend on i

Therefore, the decision rule is also

$$r^*(x) = \underset{i \in \{1,2\}}{\operatorname{argmax}} P(x \mid Y=i)P(Y=i)$$

We don't actually know these probabilities, but we estimated them in part (a).

decision boundary:

$$P(\bar{x}|y=1)P(y=1) = P(\bar{x}|y=2)P(y=2)$$

$$\frac{\pi_1}{\sqrt{2\pi}\sigma_1}\exp\left(-\frac{(\bar{x}-\mu_1)^2}{2\sigma_1^2}\right) = \frac{\pi_2}{\sqrt{2\pi}\sigma_2}\exp\left(-\frac{(\bar{x}-\mu_2)^2}{2\sigma_2^2}\right)$$

$$\frac{5/12}{\sqrt{4\pi}}\exp\left(-\frac{(\bar{x}-5)^2}{4}\right) = \frac{7/12}{\sqrt{4\pi}}\exp\left(-\frac{(\bar{x}-7)^2}{4}\right)$$

$$5\exp\left(-\frac{(\bar{x}-5)^2}{4}\right) = 7\exp\left(-\frac{(\bar{x}-7)^2}{4}\right)$$

$$\ln(5) - \frac{1}{4}(\bar{x}-5)^2 = \ln(7) - \frac{1}{4}(\bar{x}-7)^2$$

$$-4\ln(5) + \bar{x}^2 - 10\bar{x} + 25 = -4\ln(7) + \bar{x}^2 - 14\bar{x} + 49$$

$$4\bar{x} = 24 + 4\ln(5/7)$$

$$\bar{x} = 6 + \ln(5/7) \approx 5.66$$

$$r^*(x) = \begin{cases} 1 & \text{if } X \le 5.66 \\ 2 & \text{if } X > 5.66 \end{cases}$$

d) Find QDA decision rule using loss function in problem description:

$$L(\hat{y}, y) = \begin{cases} 0 & \text{if } y = \hat{y} \\ 1 & \text{if } y = 1, \hat{y} = 2 \\ 2 & \text{if } y = 2, \hat{y} = 1 \end{cases}$$

Bayes (optimal) decision rule:

$$r^*(x) = \underset{i \in \{1,2\}}{\operatorname{argmin}} \left[L(i,1) P(y=1|x) + L(i,2) P(y=2|x) \right]$$

Using Bayes rule, the decision rule is

$$r^*(x) = \underset{i \in \{1,2\}}{\operatorname{argmin}} \left[L(i,1) P(x|y=1) P(y=1) + L(i,2) P(x|y=2) P(y=2) \right]$$

Bayes decision boundary:

$$L(1,1)P(\bar{x}|y=1)P(y=1) + L(1,2)P(\bar{x}|y=2)P(y=2) = L(2,1)P(\bar{x}|y=1)P(y=1) + L(2,2)P(\bar{x}|y=2)P(y=2)$$

Using loss function above ...

$$2\,P(\bar{x}|y=2)P(y=2) = P(\bar{x}|y=1)P(y=1)$$

$$2\frac{\pi_2}{\sqrt{2\pi}\sigma_2}\exp\left(-\frac{(\bar{x}-\mu_2)^2}{2\sigma_2^2}\right) = \frac{\pi_1}{\sqrt{2\pi}\sigma_1}\exp\left(-\frac{(\bar{x}-\mu_1)^2}{2\sigma_1^2}\right)$$

$$2\frac{7/12}{\sqrt{4\pi}}\exp\left(-\frac{(\bar{x}-7)^2}{4}\right) = \frac{5/12}{\sqrt{4\pi}}\exp\left(-\frac{(\bar{x}-5)^2}{4}\right)$$

$$14 \exp\left(-\frac{(\bar{x}-7)^2}{4}\right) = 5 \exp\left(-\frac{(\bar{x}-5)^2}{4}\right)$$

$$\ln(14) - \tfrac{1}{4}(\bar{x}-7)^2 = \ln(5) - \tfrac{1}{4}(\bar{x}-5)^2$$

$$-4\ln(14) + \bar{x}^2 - 14\bar{x} + 49 = -4\ln(5) + \bar{x}^2 - 10\bar{x} + 25$$

$$4\bar{x} = 24 + 4\ln(5/14)$$

$$\bar{x} = 6 + \ln(5/14) \approx 4.97$$

$$r^*(x) = \begin{cases} 1 & \text{if } X \leq 4.97 \\ 2 & \text{if } X > 4.97 \end{cases}$$

Now that it's more costly to misclassify $Y=2$ data a $\hat{Y}=1$, our decision boundary shifts to the left!

2 Estimating the Population of Grizzly Bears

An environmentalist, Amy, wants to estimate the number of grizzly bears roaming in a forest of British Columbia, Canada. She tracks $n = 20$ bears on her first visit to the forest, and marks each one with an electronic transmitter. A month later, she returns to the same forest and tracks $k = 15$ bears, and finds that only $x = 7$ of them have the transmitter on them. Assume that on each visit, she observes a uniformly random sample of bears.

Note that the numbers of bears tracked during Amy's two visits, n and k, were chosen by her. The number x of bears she found with transmitter attached is her only observation.

(a) Assuming Amy was equally likely to encounter any particular grizzly bear during her visits, what is the likelihood $\mathcal{L}(N; x)$ of the bear population N given her observation (i.e., the number of bears with transmitter observed) x?

(b) One way to estimate the bear population is to maximize the likelihood $\mathcal{L}(N; x)$. This is called *Maximum Likelihood Estimation* (MLE), and is widely studied in statistics. Derive the expression for MLE estimate of the population $\hat{N}$ in terms of number of bears tracked in both visits (parameters n, k) and number of bears with transmitter found (observation x).

(c) What is Amy's MLE estimate $\hat{N}$ of the bear population?

2. n = # of bears tracked
   k = # of bears seen on 2nd visit
   x = # of bears seen w/ transmitter
   N = total # of bears

a) $\mathcal{L}(N; x) = P(x \mid N)$

↳ probability that Amy saw x bears w/ transmitters given that there are N total bears

$$P(x \mid N) = \frac{\text{\# of ways to see } x \text{ bears w/ transmitters}}{\text{\# of ways to capture } k \text{ bears}}$$

Of the n bears w/ transmitters, Amy found x of them. Of the $N-n$ bears w/o transmitters, Amy found $k-x$. The # of ways Amy could choose x bears w/ trans. & $k-x$ w/o is

$$\binom{N-n}{k-x} \cdot \binom{n}{x}$$

Of the N total bears, Amy captured k. The # of ways to choose k bears from N is $\binom{N}{k}$

$$\therefore \mathcal{L}(N;x) = P(x|N) = \frac{\binom{N-n}{k-x}\binom{n}{x}}{\binom{N}{k}}$$

b) $\hat{N} = \underset{N}{\text{argmax}}\ \mathcal{L}(N;x)$

B/c N is discrete, we can't use the derivative of the likelihood func. to find its maximizing value. Instead, it's useful to look at the likelihood ratio:

$$R(N|x) = \frac{\mathcal{L}(N;x)}{\mathcal{L}(N-1;x)}$$

$$\begin{cases} \mathcal{L} \text{ inc. w/ inc. } N \text{ if } R(N|x) > 1 \\ \mathcal{L} \text{ dec. w/ inc. } N \text{ if } R(N|x) < 1 \\ \mathcal{L} \text{ is at MLE of } N \text{ if } R(N|x) = 1 \end{cases}$$

From our result from part (a),

$$R(N|x) = \frac{\binom{N-n}{k-x}\binom{n}{x}}{\binom{N}{k}} \left[\frac{\binom{N-n-1}{k-x}\binom{n}{x}}{\binom{N-1}{k}} \right]^{-1}$$

$$= \frac{\binom{N-n}{k-x}\binom{n}{x}\binom{N-1}{k}}{\binom{N-n-1}{k-x}\binom{n}{x}\binom{N}{k}} = \frac{\binom{N-n}{k-x}\binom{N-1}{k}}{\binom{N-n-1}{k-x}\binom{N}{k}}$$

Note that $\dfrac{\binom{a}{b}}{\binom{a-1}{b}} = \dfrac{\frac{a!}{b!(a-b)!}}{\frac{(a-1)!}{b!(a-b-1)!}} = \dfrac{a!(a-b-1)!}{(a-1)!(a-b)!} = \dfrac{a}{a-b}$

Using this fact,

$$R(N|x) = \frac{(N-n)(N-k)}{N(N-n-k+x)}$$

$$R(\hat{N}|x) = 1 \Rightarrow \frac{(\hat{N}-n)(\hat{N}-k)}{\hat{N}(\hat{N}-n-k+x)} = 1$$

$$(\hat{N}-n)(\hat{N}-k) = \hat{N}(\hat{N}-n-k+x)$$

$$\hat{N}^2 - n\hat{N} - k\hat{N} + nk = \hat{N}^2 - n\hat{N} - k\hat{N} + x\hat{N}$$

$$nk = x\hat{N}$$

$$\hat{N} = \frac{nk}{x}$$

c) Using the values from our problem,

$$\hat{N} = \frac{20(15)}{7} \approx 42.86$$

This is not an integer! Instead the sol'n must actually be either 42 or 43.

We can compare the likelihoods of these two values by looking at the likelihood ratio. Recall that

$$R(N|x) = \frac{L(N;x)}{L(N-1;x)}$$

Therefore, $R(43|7) = \frac{L(43;7)}{L(42;7)}$

From our findings in part (b)

$$R(43|7) = \frac{(43-20)(43-15)}{43(43-20-15-7)} \approx 0.998$$

$R(43|7) < 1 \Rightarrow L(42;7) > L(43;7)$

$\therefore$ We should choose $\hat{N} = 42$.

3 Quadratic Forms

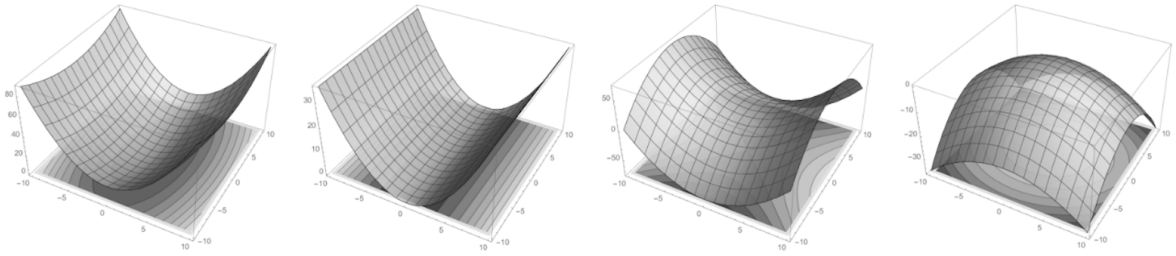
A quadratic form is a polynomial whose terms are all degree two (an example would be $f(x, y) = x^2 + 3xy + 42y^2$). All quadratic forms can be written in the form

$$\boxed{Q(\mathbf{x}) = \mathbf{x}^T A \mathbf{x}.}$$

In this problem, we will visually and mathematically explore some properties of quadratic forms, which will help us intuitively understand LDA and QDA.

(a) Suppose you are given the a quadratic function $Q(x) = \frac{1}{2}x^\top Ax + b^\top x$ where $x, b \in \mathbb{R}^2$ and $A \in \mathbb{R}^{2\times 2}$ is a symmetric matrix. What is the Hessian of Q?

(b) We will now think about how the eigenstructure of the Hessian matrix affects the shape of $Q(x)$. Recall that by the Spectral Theorem, A has two real eigenvalues. Match each of the following cases, to the appropriate plot of $Q(x)$. How does the magnitude of the eigenvalues affect your sketch? For each of these four cases, is there an unique local optimum?



(a) $\lambda_1(A), \lambda_2(A) > 0$

(b) $\lambda_1(A), \lambda_2(A) < 0$

(c) $\lambda_1(A) > 0, \lambda_2(A) < 0$

(d) $\lambda_1(A) > 0, \lambda_2(A) = 0$

3. a) $Q(x) = \frac{1}{2} x^T A x + b^T x$, $b, x \in \mathbb{R}^2$, $A \in \mathbb{S}^2$

$\nabla_x^2 Q(x) = ?$

$$\nabla_x Q(x) = \frac{1}{2}(A + A^T)x + b = Ax + b$$
$$\nabla_x^2 Q(x) = \nabla_x (Ax + b) = A^T = A$$

b) The shape of $Q(x)$ is determined by the term $x^T A x$. The $\frac{1}{2}$ simply scales this shape & $b^T x$ shifts it. Therefore, let's look closer at the term $x^T A x$.

B/c A is symmetric, it admits the spectral decomposition $A = U \Lambda U^T$. Therefore,
$x^T A x = x^T U \Lambda U^T x$

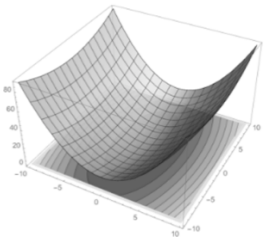
Let $z := U^T x$

$$x^T A x = z^T \Lambda z = \begin{bmatrix} z_1 & z_2 \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$$
$$= \lambda_1 z_1^2 + \lambda_2 z_2^2$$

U is orthogonal, so it is also non-singular.
$\therefore z = 0_2$ iff $x = 0_2$

i. $\lambda_1, \lambda_2 > 0 \rightarrow \lambda_1 z_1^2 + \lambda_2 z_2^2 > 0 \quad \forall z \neq 0_2$

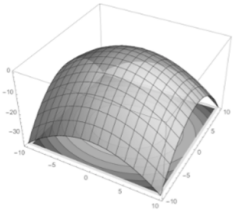
$X^T A X > 0 \quad \forall X \neq 0_2$



unique local minimum

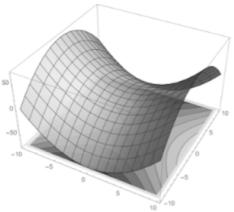
ii. $\lambda_1, \lambda_2 < 0 \rightarrow \lambda_1 z_1^2 + \lambda_2 z_2^2 < 0 \quad \forall z \neq 0_2$

$X^T A X < 0 \quad \forall X \neq 0_2$



unique local maximum

iii. $\lambda_1 > 0, \lambda_2 < 0 \rightarrow X^T A X = 0$ for some values of X, > 0 for others, & < 0 for the rest

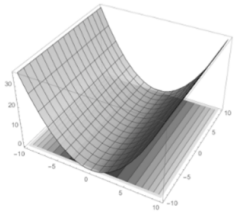


saddle point

iv. $\lambda_1 > 0, \lambda_2 = 0 \rightarrow \lambda_1 z_1^2 + \lambda_2 z_2^2 = 0$ if $z_1 = 0$

$\lambda_1 z_1^2 + \lambda_2 z_2^2 > 0$ if $z_1 \neq 0$

$x^T A x \geq 0 \quad \forall x$



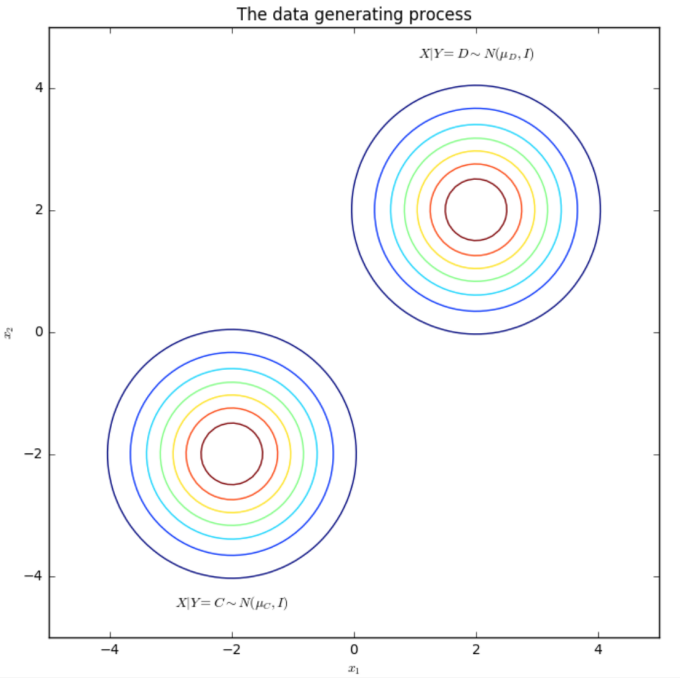
Continuum of local minima

4 Linear Discriminant Analysis

In this question, we will explore some of the mechanics of LDA and understand why it produces a linear decision boundary in the case where the covariance matrix is isotropic.

Suppose you have a binary classification problem with $x \in \mathbb{R}^2$, and you already know the data generating process.

- The two classes have identical priors $P(Y = C) = P(Y = D) = \frac{1}{2}$.
- The class-conditional densities are $(X|Y = C) \sim N(\mu_C, I)$ and $(X|Y = D) \sim N(\mu_D, I)$ where $\mu_C = \begin{bmatrix} -2 \\ -2 \end{bmatrix}, \mu_D = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$.



We can recognize this problem as a special case of LDA where the two classes have an equal prior probability and the common covariance matrix is simply the identity. Show that the optimal Bayes optimal decision boundary is the perpendicular bisector of the line connecting μ_0 and μ_1.

4. Assuming we use the zero-one loss function, Bayes (optimal) decision rule is

$$r^*(x) = \underset{y \in \{C, D\}}{\operatorname{argmax}} f(Y=y \mid X=x)$$
$$= \underset{y \in \{C, D\}}{\operatorname{argmax}} f(X=x \mid Y=y) P(Y=y)$$

$$f(X=x \mid Y=C) P(Y=C) = f(X=x \mid Y=D) P(Y=D)$$

$$\frac{1}{2} \frac{1}{\sqrt{2\pi |I_2|^{2/2}}} \exp\left(-\frac{1}{2}(x-\mu_C)^T I_2 (x-\mu_C)\right) =$$

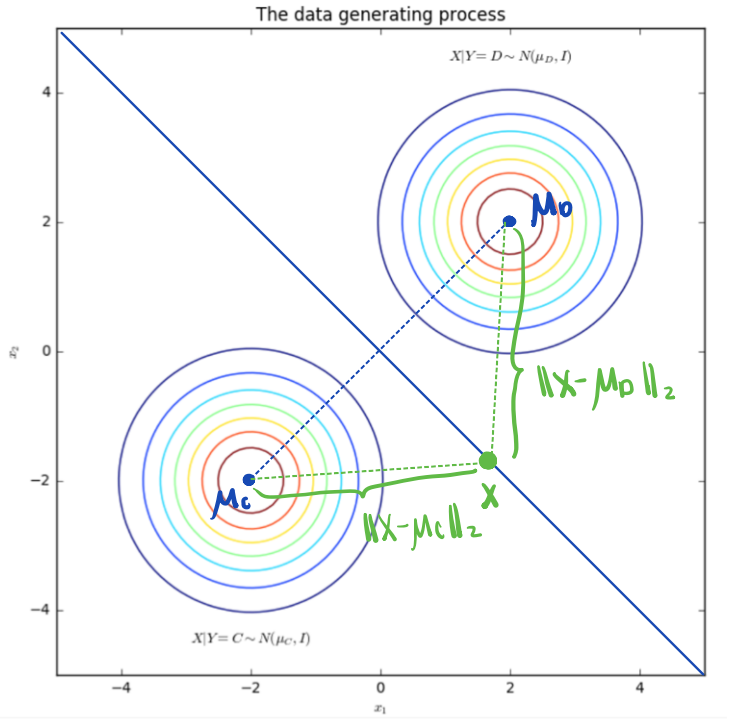
$$\frac{1}{2} \frac{1}{\sqrt{2\pi |I_2|^{2/2}}} \exp\left(-\frac{1}{2}(x-\mu_D)^T I_2 (x-\mu_D)\right)$$

$$(x-\mu_C)^T I_2 (x-\mu_C) = (x-\mu_D)^T I_2 (x-\mu_D)$$
$$\|x-\mu_C\|_2^2 = \|x-\mu_D\|_2^2$$

Bayes (optimal) decision boundary:

$$\{x \in \mathbb{R}^2 : \|x-\mu_C\| = \|x-\mu_D\|\}$$

The decision boundary is composed of the points equidistant from μ_C & μ_D.


The data generating process
$X|Y=D \sim N(\mu_D, I)$
$X|Y=C \sim N(\mu_C, I)$
μ_D
μ_C
x
$\|x-\mu_D\|_2$
$\|x-\mu_C\|_2$
x_1
x_2