CS 189	Introduction to Machine Learning	
Spring 2021	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.

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

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

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

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

$Y = 1$ for salmon

$Y = 2$ for seabass

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 in both the seabass and the salmon case. Also compute the empirical estimates of the priors $\hat{\pi}_i$.

$\hat{\mu}_1 = 5$
$\hat{\sigma}_1^2 = 2$
$\hat{\pi}_1 = 5/12$

$\hat{\mu}_2 = 7$
$\hat{\sigma}_2^2 = 2$
$\hat{\pi}_2 = 7/12$

$\cancel{\frac{40}{100}} = 0.4 = \pi_{salmon}$

$\cancel{\frac{60}{100}} = 0.6 = \pi_{seabass}$

(b) What is significant about $\hat{\sigma}_1$ and $\hat{\sigma}_2$? They are equal.

(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.

(c) We are given weight x.

$$f(Y=1|x) = \frac{f(x|Y=1)\Pr[Y=1]}{f(x)}$$

$$f(Y=2|x) = \frac{f(x|Y=2)\Pr[Y=2]}{f(x)}$$

$f(Y=1|x) > f(Y=2|x) \Rightarrow$ Predict 1, o/w 2

Boundary:

$$f(Y=1|x) = f(Y=2|x)$$

$$f(x|Y=1)\,\cancel{\Pr[Y=1]}^{\hat{\pi}_1\ 5/\cancel{12}} = f(x|Y=2)\,\cancel{\Pr[Y=2]}^{\hat{\pi}_2\ 7/\cancel{12}}$$

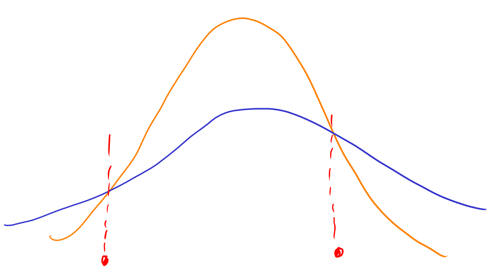
$$\frac{5}{\cancel{\sqrt{2\pi}}\,\hat{\sigma}_1} \exp\left(-\frac{1}{2}\frac{(x-5)^2}{\hat{\sigma}_1^2}\right) = \frac{7}{\cancel{\sqrt{2\pi}}\,\hat{\sigma}_2} \exp\left(-\frac{1}{2}\frac{(x-7)^2}{\hat{\sigma}_2^2}\right)$$

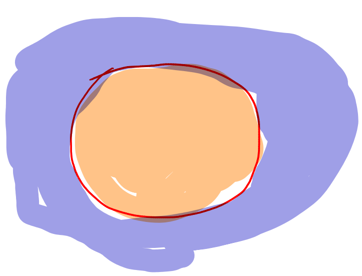
$\vdots$

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

$$x = \ln\left(\frac{5}{7}\right) + 6 \approx 5.66$$

(d)





Risk of pred 1 < Risk of pred 2 $\Rightarrow$ Pred 1 o/w 2

Boundary

Risk of pred 1 = Risk of pred 2

$$R(\hat{y}=1|x) = \cancel{\lambda_{11}}^{0} P(Y=1|x) + \cancel{\lambda_{12}}^{2} P(Y=2|x)$$

$$R(\hat{y}=2|x) = {}^{1}\cancel{\lambda_{21}} P(Y=1|x) + \cancel{\lambda_{22}}^{0} P(Y=2|x)$$

$$R(\hat{y}=1|x) = R(\hat{y}=2|x)$$

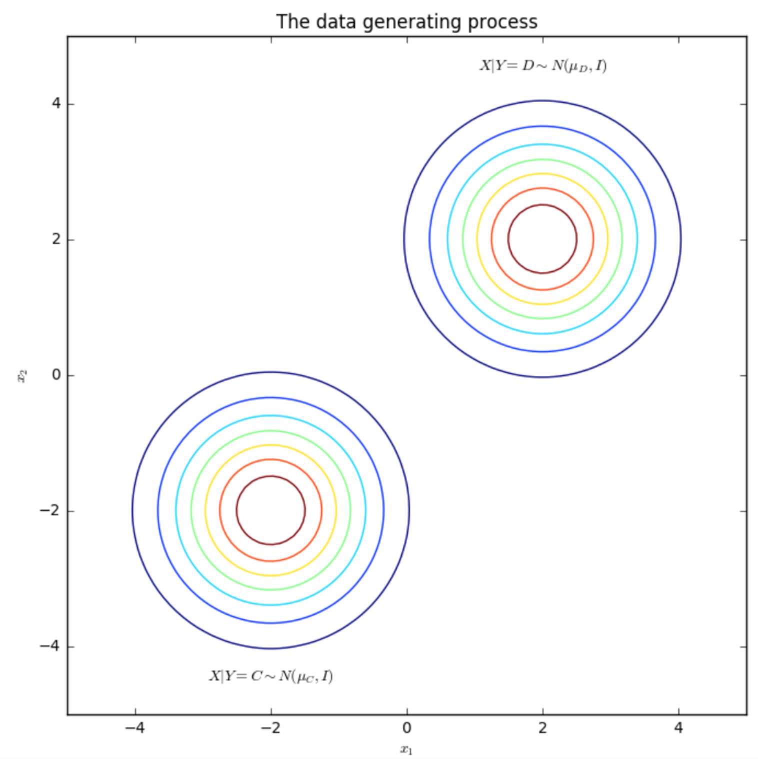
$$2P(Y=2|x) = P(Y=1|x)$$

2 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 anisotropic.

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. What is the Bayes optimal decision boundary for this problem? You may want to start by drawing the decision boundary on the plot provided. Does the result line up with your intuition?

3 Estimating Population of Grizzly Bears

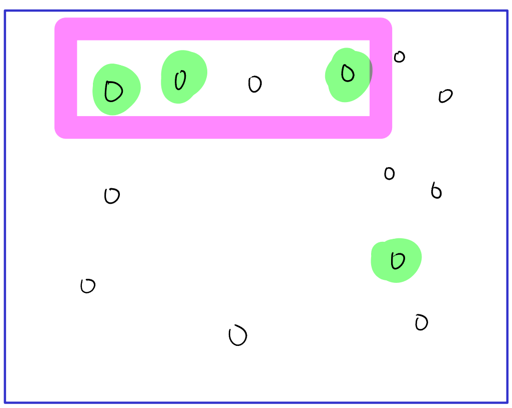
An environmentalist Amy wants to estimate the number grizzly bears roaming in a forest of British Columbia, Canada. She tracks $n = 20$ bears on her first visit to the forest, and marks them with an electronic transmitter. A month later, she returns to the same forest and tracks $k = 15$ bears with only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

(a) Note that the number of bears tracked during Amy's two visits n, k was chosen by her. The number of bears she found with transmitter attached is her only observation.

Assuming Amy was equally likely to encounter any of the grizzly bears during her visits, what is the likelihood $\mathcal{L}(N; x)$ of the bear population N given her observation 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?



(a) $\binom{N}{k}$ $\qquad$ $\binom{n}{x}\binom{N-n}{k-n}$

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

$\leftarrow$ # of ways to obtain actual observation

$\leftarrow$ Total # of possible observations

$\boxed{Pr[N|x]}$

(b) $\mathcal{L}(N; x)$ — x: fixed, N: variable

$\frac{\partial \mathcal{L}}{\partial N} = 0$ solve for N

$\frac{\partial^2 \mathcal{L}}{\partial N} \leq 0$

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

$$= \frac{(N-k)(N-n)}{N(N-n-k+x)} = 1$$

Mode of binomial dist

MLE of Unif$[0, \theta]$

$R(43|x) < 1$

$\mathcal{L}(43|x) < \mathcal{L}(42|x)$

$\Rightarrow (N-k)(N-n) = N(N-n-k+x)$

$\Rightarrow \hat{N} = \frac{nk}{x} = \frac{300}{7} \approx 42.8571\ldots$

42

4 Logistic posterior with exponential class conditionals

Suppose we have the job of binary classification given a scalar feature $X \in \mathbb{R}_{\geq 0}$ Now, suppose the distribution of X conditioned on the class y is exponentially distributed with parameter λ_y, i.e.,

$$
\begin{aligned}
&X \in \mathbb{R}_{\geq 0} \\
&P(X = x|Y = y) = \lambda_y \exp\Big(-\lambda_y x\Big), \quad \text{where } y \in \{0, 1\} \\
&Y \sim \text{Bernoulli}(\pi)
\end{aligned}
$$

(a) Show that the posterior distribution of the class label given X is a logistic function, however with a linear argument in X. That is, show that $P(Y = 1|X = x)$ is of the form $\frac{1}{1+\exp\left(-h(x)\right)}$, where $h(x) = ax + b$ is linear in x.

(b) Assuming 0-1 loss, what is the optimal classifier and decision boundary?