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

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 $\hat{\pi}_i$.

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

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

(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.

Solution:

(a) Sample mean $\hat{\mu}$:

$$\hat{\mu} = \frac{1}{N}\sum_i X_i$$

Sample variance:

$$\hat{\sigma}^2 = \frac{1}{N}\sum_i (X_i - \hat{\mu})^2$$

Plugging in numbers for seabass and salmon: $\hat{\mu}_1 = 5, \hat{\mu}_2 = 7, \hat{\sigma}_1^2 = 2, \hat{\sigma}_2^2 = 2$

Calculating the priors: $\hat{\pi}_1 = 5/12, \hat{\pi}_2 = 7/12$

(b) They're the exact same, so a decision boundary between the two Gaussians characterized by them will be linear.

(c) Recall that assuming a 0-1 loss function results in choosing the class to minimize the probability of error, which means choosing according to this rule:

$$\text{If } \frac{p(Y=1|x)}{p(Y=2|x)} > 1, \text{ choose 1}$$

Because there is a linear decision boundary, we search for the value such that we classify everything to the right as seabass, and everything to the left as salmon. This boundary is the value of x such that $p(Y=1|x) = p(Y=2|x)$.

$$p(Y=1|x) = p(Y=2|x) \implies 5p(x|Y=1) = 7p(x|Y=2)$$

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

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

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

$$4\ln\left(\frac{5}{7}\right) + 24 = 4x$$

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

The decision rule is: If $x > 5.66$, classify as Seabass! Otherwise classify as Salmon.

Note: Because we had the same variance for both class conditionals, the x^2 term canceled out. If that was not the case, then there would be 3 regions, and we would allocate 2 of them to one fish, 1 of them to the other, depending on the height of the posterior probabilities. A good exercise would be to try to draw this: two 1-D Gaussians with different variances.

(d) In the general case, we want to make the decision that minimizes risk. Thus, the decision boundary is located at where the risk of making either decision is equal, or:

$$\textit{Risk of predicting } 1 \textit{ given } x \;=\; \textit{Risk of predicting } 2 \textit{ given } x$$

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

Recall that $R(\hat{y} = i|x) = \sum_{j=1}^{C} \lambda_{ij} P(Y = j|x)$.

$$\lambda_{11}P(Y = 1|x) + \lambda_{12}P(Y = 2|x) = \lambda_{21}P(Y = 1|x) + \lambda_{22}P(Y = 2|x)$$

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

$$2 \cdot \frac{7}{12}\mathcal{N}(7, 2) = 1 \cdot \frac{5}{12}\mathcal{N}(5, 2)$$

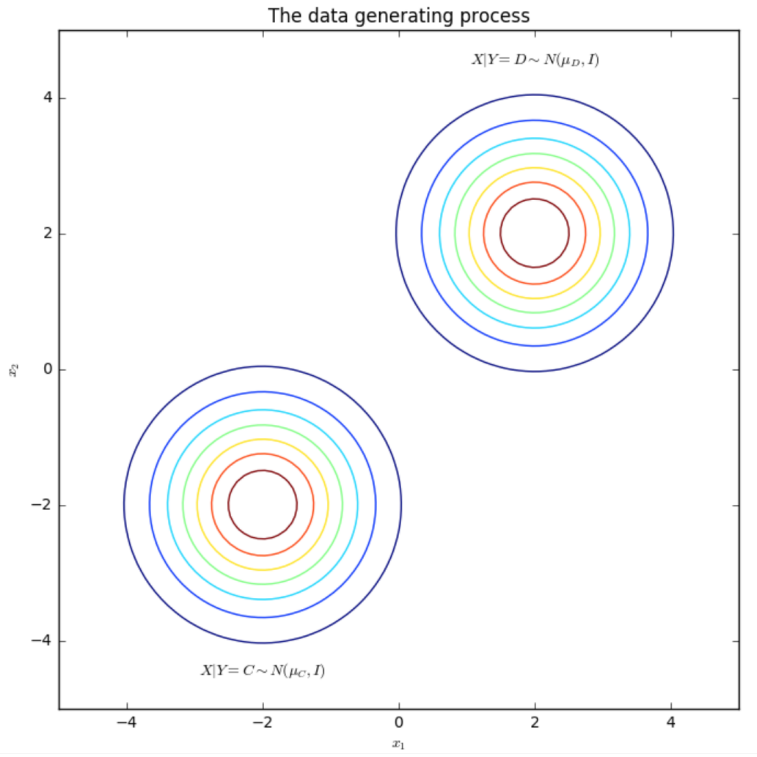
Solving this like part b), we get that $x = 6 + \ln\left(\frac{5}{14}\right) \approx 4.97$. Thus, if the weight is greater than 4.97, we classify it as seabass and if not, we classify it as salmon.

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?

Solution: The Bayes optimal decision boundary is the perpendicular bisector of the line connecting μ_0 and μ_1. To find the decision boundary consider:

$$P(Y = D|X) = P(Y = C|X)$$
$$f(X|Y = D)P(Y = D) = f(X|Y = C)P(Y = C)$$

$$f(X|Y=D) = f(X|Y=C)$$

$$\frac{1}{\sqrt{2\pi}|I|}\exp\left(-\frac{1}{2}(x-\mu_C)^\top I(x-\mu_C)\right) = \frac{1}{\sqrt{2\pi}|I|}\exp\left(-\frac{1}{2}(x-\mu_D)^\top I(x-\mu_D)\right)$$

$$\|x-\mu_C\|_2^2 = \|x-\mu_D\|_2^2$$

Hence the decision boundary contains all the points that are equidistant from the two class means. The set of points $h = \{x \in \mathbb{R}^2 : \|x-\mu_C\|_2^2 = \|x-\mu_D\|_2^2\}$ is exactly the perpendicular bisector of the line that connects the two means in the picture. That is, h is the plane that is orthogonal to the vector $\mu_D - \mu_C$ and passes through the point $\frac{\mu_C+\mu_D}{2}$.

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?

Solution: The likelihood $\mathcal{L}(N; x)$ is probability that Amy saw x bears with transmitters on her second trip, given N total bears. The number of ways Amy could have capture k bears on second visit = $\binom{N}{k}$. The number of ways x of them had transmitter $\binom{N-n}{k-x} \times \binom{n}{x}$. Thus, likelihood is given by:

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

(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).

Solution: Since the random variable N is discrete, calculus isn't the best way to optimize this.

Alternatively, look at the likelihood ratio $R(N|x) = \frac{\mathcal{L}(N;x)}{\mathcal{L}(N-1;x)}$. While $R(N|x) \geq 1$, likelihood increases with increasing N, and decreases if $R(N|x) \leq 1$. Thus, $R(N|x) = 1$ should be satisfied by MLE estimate $\hat{N}$.

Simplify the expression for likelihood ratio:

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

Simplify by using $\binom{n-1}{k}/\binom{n}{k} = \frac{n-k}{n}$ to get $R(N|x) = \frac{(N-k)(N-n)}{N(N-n-k+x)}$.

Solving $R(N|x) = 1$:

$$R(N|x) = 1 \implies (N-k)(N-n) = N(N-n-k+x)$$

$$\implies N^2 - Nk - Nn + nk = N^2 - Nk - Nn + Nx \implies nk = Nx \implies \hat{N} = \frac{nk}{x}$$

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

Solution: $\frac{nk}{x} = \frac{15\times20}{7} = 300/7$ is not an integer. But by the logic of likelihood ratios, the MLE estimate must be the closest integers to $300/7$. The closest integers are 42 and 43. We need to evaluate the Likelihood Ratio for both of them in order to find the true MLE $\hat{N}$.

$R(42|7) = \frac{(42-15)(42-20)}{42(42-15-20+7)} = \frac{27\times22}{42\times14} \approx 1.01.$

$R(43|7) = \frac{(43-15)(43-20)}{43(43-15-20+7)} = \frac{28\times23}{43\times15} \approx 0.998.$

$R(42|7) > 1 \implies \mathcal{L}(42;7) > \mathcal{L}(41;7)$ and $R(43|7) < 1 \implies \mathcal{L}(43;7) < \mathcal{L}(42;7)$. Therefore, $\mathcal{L}(42;7)$ is the largest and $\hat{N} = 42$.

In general, the greatest integer less or equal to $\frac{nk}{x}$ is the true $\hat{N}$.

(Caution: do not attempt to calculate the actual likelihood. They involve really large numbers!)

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?

Solution:

We are solving for $P(Y = 1|x)$. By Bayes Rule, we have

$$\begin{aligned} P(Y = 1|x) &= \frac{P(x|Y = 1)P(Y = 1)}{P(x|Y = 1)P(Y = 1) + P(x|Y = 0)P(Y = 0)} \\ &= \frac{1}{1 + \frac{P(Y=0)P(x|Y=0)}{P(Y=1)P(x|Y=1)}} \\ &= \frac{1}{1 + \frac{\lambda_0}{\lambda_1}\frac{1-\pi}{\pi}\exp\left(-\lambda_0 x + \lambda_1 x\right)} \end{aligned}$$

Looking at the bottom right equation, we have

$$\frac{\lambda_0}{\lambda_1}\frac{1-\pi}{\pi}\exp\left(-\lambda_0 x + \lambda_1 x\right) = \exp\left(-(\lambda_0 - \lambda_1)x + \log\left(\frac{\lambda_0}{\lambda_1}\frac{1-\pi}{\pi}\right)\right)$$

Now we see that we have a logistic function $\frac{1}{1+\exp\left(-h(x)\right)}$, where $h(x) = ax + b$ is linear (affine) in x. Since we are assuming 0-1 loss, we use the optimal classifier $f^*(x) = 1$ when $P(Y = 1|x) > P(Y = 0|x)$. Thus, the decision boundary can be found when $P(Y = 1|x) = P(Y = 0|x) = \frac{1}{2}$. This happens when $h(x) = 0$. Solving for x gives

$$\bar{x} = \frac{\log\frac{\lambda_0}{\lambda_1}\frac{1-\pi}{\pi}}{\lambda_0 - \lambda_1}.$$

If we assume $\lambda_0 > \lambda_1$, then the optimal classifier is

$$f^*(x) = \begin{cases} 1 & \text{if } x > \bar{x} \\ 0 & o.w. \end{cases}$$