CS 529: Homework 2

Andrea Zapata - PUID : 0031827996

Problem 1

1.

$$g_1(x) = \ln\left[(2\pi\sigma^2)^{-1/2} e^{-1/2\sigma^2(x-\mu_1)^2}\right] + \ln\frac{1}{2} = -\frac{1}{2}\ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}(x-\mu_1)^2 + \ln\frac{1}{2}$$

$$g_2(x) = \ln\left[(2\pi\sigma^2)^{-1/2} e^{-1/2\sigma^2(x-\mu_2)^2}\right] + \ln\frac{1}{2} = -\frac{1}{2}\ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}(x-\mu_2)^2 + \ln\frac{1}{2}$$

2.

$$\begin{aligned} g(x) = g_1(x) - g_2(x) &= -\frac{1}{2\sigma^2}(x-\mu_1)^2 + \frac{1}{2\sigma^2}(x-\mu_2)^2 \\ &= \frac{1}{2\sigma^2}[x^2 - 2x\mu_2 + \mu_2^2 - x^2 + 2x\mu_1 - \mu_1^2] \\ &= \frac{1}{2\sigma^2}[\mu_2^2 - \mu_1^2 + 2x(\mu_1 - \mu_2)] = 0 \\ \rightarrow x &= \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)} \end{aligned}$$

$$choose \begin{cases} w_1, & x \leq \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)} \\ w_2, & \text{otherwise} \end{cases}$$

3.

4.

$$g_1(x) = -\frac{1}{2}\ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}(x-\mu_1)^2 + \ln 0.8$$

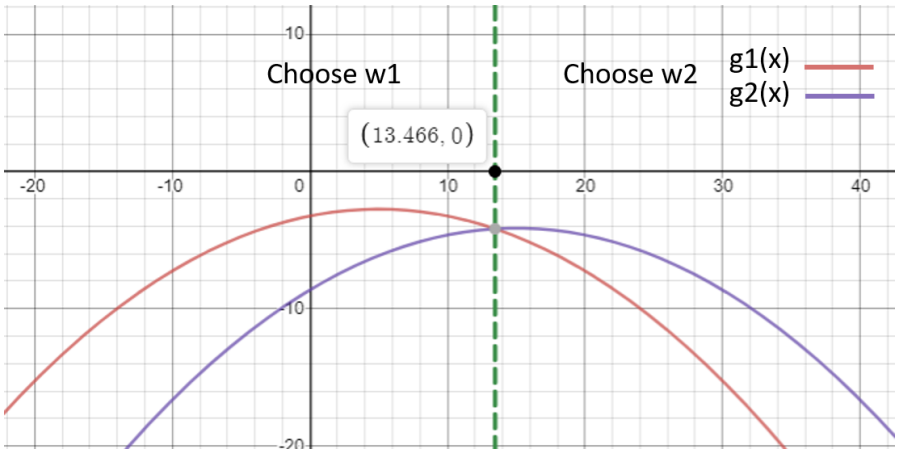
$$g_2(x) = -\frac{1}{2}\ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}(x-\mu_2)^2 + \ln 0.2$$

$$\begin{aligned} g(x) &= g_1(x) - g_2(x) = \\ &= \frac{1}{2\sigma^2}[\mu_2^2 - \mu_1^2 + 2x(\mu_1 - \mu_2)] + \ln 4 \end{aligned}$$

If we set $g(x) = 0$, then:

$$x = \frac{\mu_1^2 - \mu_2^2 - 2\sigma^2 \ln 4}{2(\mu_1 - \mu_2)}$$

$$choose \begin{cases} w_1, & x \leq \frac{\mu_1^2 - \mu_2^2 - 2\sigma^2 \ln 4}{2(\mu_1 - \mu_2)} \\ w_2, & \text{otherwise} \end{cases}$$



Problem 2

1.

$$\mu_A = \frac{68}{20} = 3.4 \qquad \sigma_A = \sqrt{\frac{22.8}{20}} = \sqrt{1.14} \qquad P(A) = \frac{20}{30} = \frac{2}{3}$$

$$\mu_B = \frac{234}{10} = 23.4 \qquad \sigma_B = \sqrt{\frac{6.4}{10}} = \sqrt{0.64} = 0.8 \qquad P(B) = \frac{10}{30} = \frac{1}{3}$$

2.

$$\begin{aligned} g_A(x) &= P(X|A)P(A) \\ &= (2\pi\sigma_A^2)^{-1/2} e^{-1/2\sigma_A^2(x-\mu_A)^2} P(A) \\ g_B(x) &= P(X|B)P(B) \\ &= (2\pi\sigma_B^2)^{-1/2} e^{-1/2\sigma_B^2(x-\mu_B)^2} P(B) \\ g(x) &= g_A(x) - g_B(x) = 0 \rightarrow g_A(x) = g_B(x) \end{aligned}$$

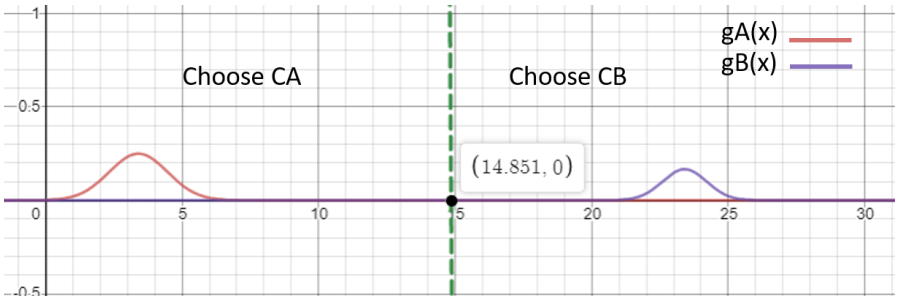
$$(2\pi\sigma_A^2)^{-1/2}e^{-1/2\sigma_A^2(x-\mu_A)^2}P(A) = (2\pi\sigma_B^2)^{-1/2}e^{-1/2\sigma_B^2(x-\mu_B)^2}P(B)$$
$$\ln\left[(2\pi\sigma_A^2)^{-1/2}e^{-1/2\sigma_A^2(x-\mu_A)^2}P(A)\right] = \ln\left[(2\pi\sigma_B^2)^{-1/2}e^{-1/2\sigma_B^2(x-\mu_B)^2}P(B)\right]$$
$$-\frac{1}{2}\ln(2\pi) - \ln\sigma_A - \frac{1}{2\sigma_A^2}(x-\mu_A)^2 + \ln(P(A)) = -\frac{1}{2}\ln(2\pi) - \ln\sigma_B - \frac{1}{2\sigma_B^2}(x-\mu_B)^2 + \ln(P(B))$$
$$-\frac{1}{2\sigma_A^2}(x-\mu_A)^2 + \ln\frac{P(A)}{\sigma_A} = -\frac{1}{2\sigma_B^2}(x-\mu_B)^2 + \ln\frac{P(B)}{\sigma_B}$$

$$\frac{1}{2\sigma_B^2}(x^2 - 2x\mu_B + \mu_B^2) - \frac{1}{2\sigma_A^2}(x^2 - 2x\mu_A + \mu_A^2) + \ln\frac{P(A)}{\sigma_A} - \ln\frac{P(B)}{\sigma_B} = 0$$
$$\left(\frac{1}{2\sigma_B^2} - \frac{1}{2\sigma_A^2}\right)x^2 + \left(\frac{2\mu_A}{2\sigma_A^2} - \frac{2\mu_B}{2\sigma_B^2}\right)x + \left(\frac{\mu_B^2}{2\sigma_B^2} - \frac{\mu_A^2}{2\sigma_A^2} + \ln\frac{\sigma_B}{\sigma_A}\frac{P(A)}{P(B)}\right) = 0$$

If we replace the constants, we get that :

$$0.3426x^2 - 33.58x + 423.1155 = 0$$

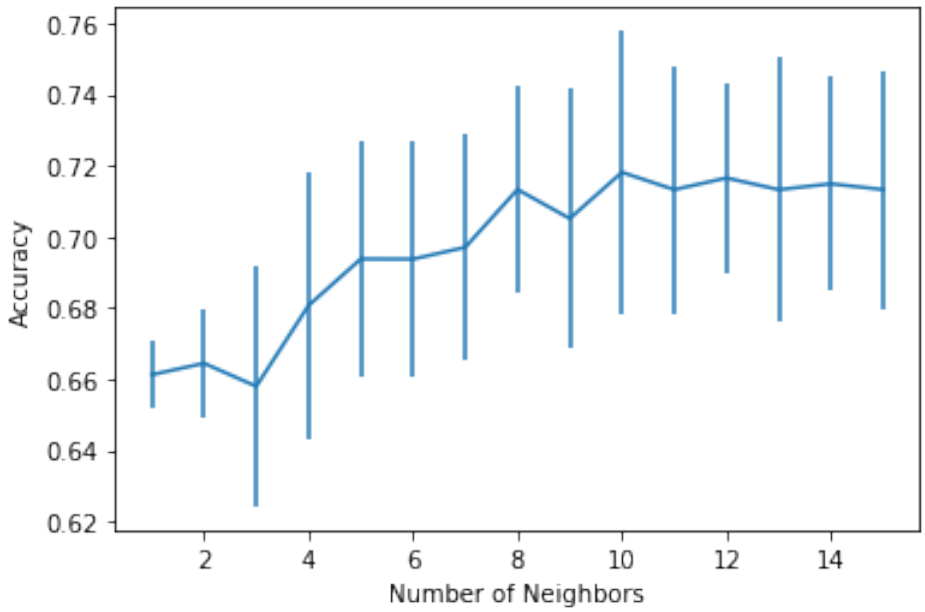
Then, if we use the quadratic formula, we get that the possible values of x are $x = 14.851$ and $x = 83.165$. If we plot both pdf, we can observe that $x = 14.851$ is the appropriate value for separating the regions.



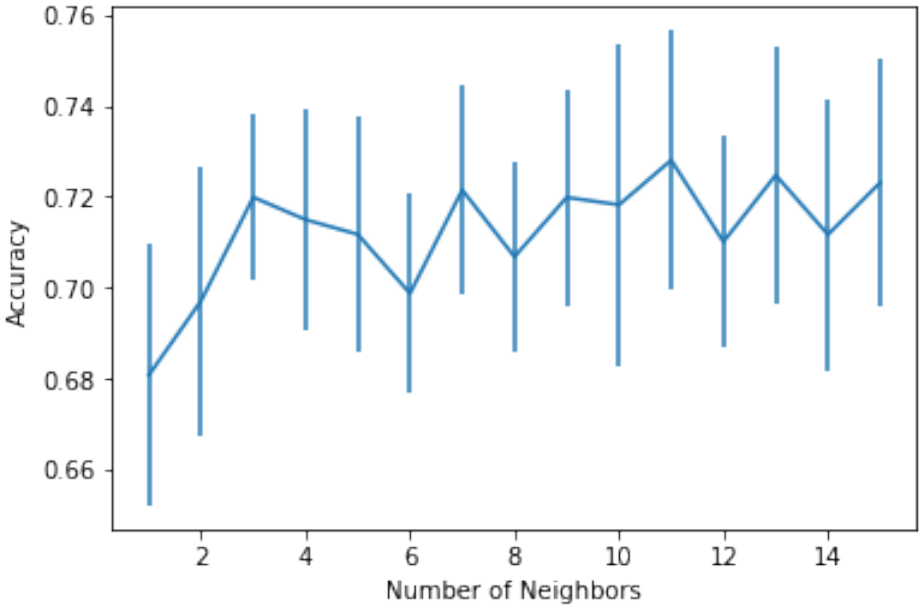
$$choose \begin{cases} C_A, & x \leq 14.851 \\ C_B, & \text{otherwise} \end{cases}$$

Problem 3

In the next graph, we can see the mean and standard deviation of the accuracy for each value of k without standardizing the data. We can see that the highest mean accuracy is for k=10, which also has the highest accuracy if we consider the standard deviation. Then, k=10 will be the selected number of neighbors. With this value, the mean accuracy on test data is 0.7597.



When we standardize our data, we need to reconsider the values for our hyper-parameters, then in the next graph, we can see the mean and standard deviation of the accuracy for each value of k after standardizing our data. We can see that the highest mean accuracy is for k=11, which also has the highest accuracy if we consider the standard deviation. Then, k=11 will be the selected number of neighbors.With this value, the mean accuracy on test data is 0.7922.



As we can see in the feature description, each feature has a different scale, then when the k nearest neighbors are found with the euclidean distance without using standardization, we are not considering how each different distance between features add up to the total. For example, Insulin is between [0, 846] while DiabetesPedigreeFunction is in the range [0.078000,2.420000]. A distance of 1 between two values in the DiabetesPedigreeFunction feature can tell us more about our data that a distance of 10 between the Insulin feature, but the euclidean distance will be inclined towards the features with larger scales. When we standardize our data, we make sure that all our features are given the same weight when calculating the distance, which makes our prediction more accurate.

Problem 4

1.

$$H(S) = -\frac{9}{16}\log_2\left(\frac{9}{16}\right) - \frac{7}{16}\log_2\left(\frac{7}{16}\right) = 0.9887$$

Let's find the information gain for the color attribute:

$$H(S|Yellow) = -\frac{8}{13}\log_2\left(\frac{8}{13}\right) - \frac{5}{13}\log_2\left(\frac{5}{13}\right) = 0.9612$$
$$H(S|Green) = -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right) = 0.9183$$
$$InfoGain(S, Color) = 0.9887 - \frac{13}{16}\cdot 0.9612 - \frac{3}{16}\cdot 0.9183 = 0.0355$$

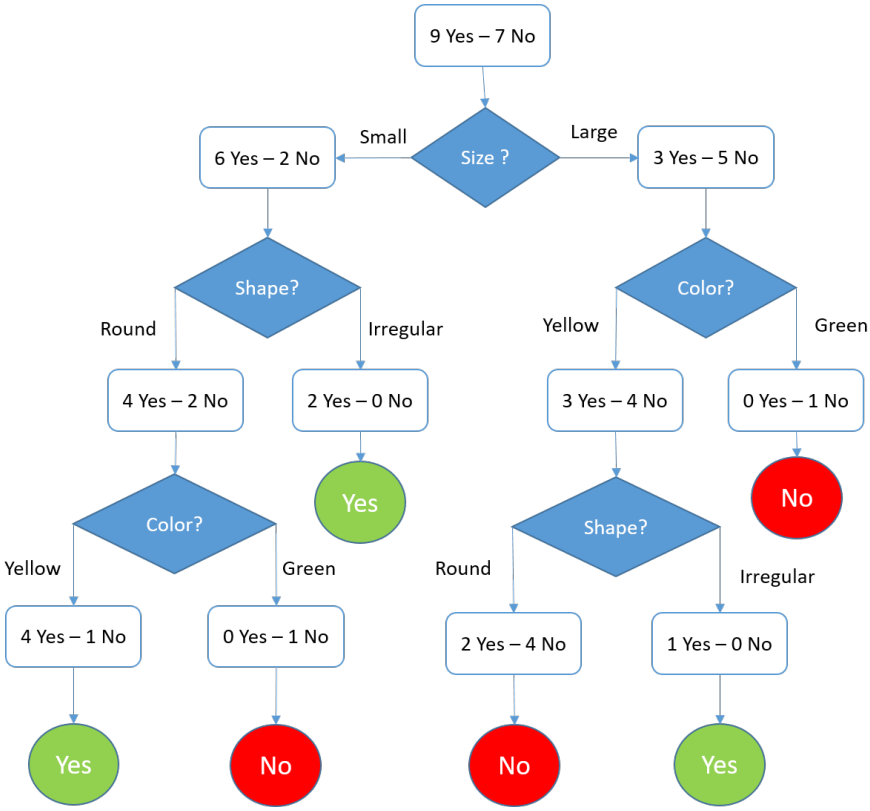
Let's find the information gain for the size attribute:

$$H(S|Small) = -\frac{6}{8}\log_2\left(\frac{6}{8}\right) - \frac{2}{8}\log_2\left(\frac{2}{8}\right) = 0.8113$$
$$H(S|Large) = -\frac{3}{8}\log_2\left(\frac{3}{8}\right) - \frac{5}{8}\log_2\left(\frac{5}{8}\right) = 0.9544$$
$$InfoGain(S, Size) = 0.9887 - \frac{8}{16}\cdot 0.8113 - \frac{8}{16}\cdot 0.9544 = 0.1058$$

Let's find the information gain for the shape attribute:

$$H(S|Round) = -\frac{6}{12}\log_2\left(\frac{6}{12}\right) - \frac{6}{12}\log_2\left(\frac{6}{12}\right) = 1$$
$$H(S|Irregular) = -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) = 0.8113$$
$$InfoGain(S, Shape) = 0.9887 - \frac{12}{16}\cdot 1 - \frac{4}{16}\cdot 0.8113 = 0.0359$$

Then we will choose **Size** as the root because it has the higher Information Gain.

2. 

[ht]

3. If we treat continuous values as categorical data, then our tree will be more complex as each decision node will originate several branches. If the complexity of our tree increases, then it is more possible to overfit our training data. Also, one of the advantages of decision trees is its ability to handle unseen data, as it can create a N/A branch for each decision node. For continuous values, that branch might be over used, as the new values might no be exactly the same as the ones found in training data, which makes the learned model inaccurate as this attribute will not help for class prediction.