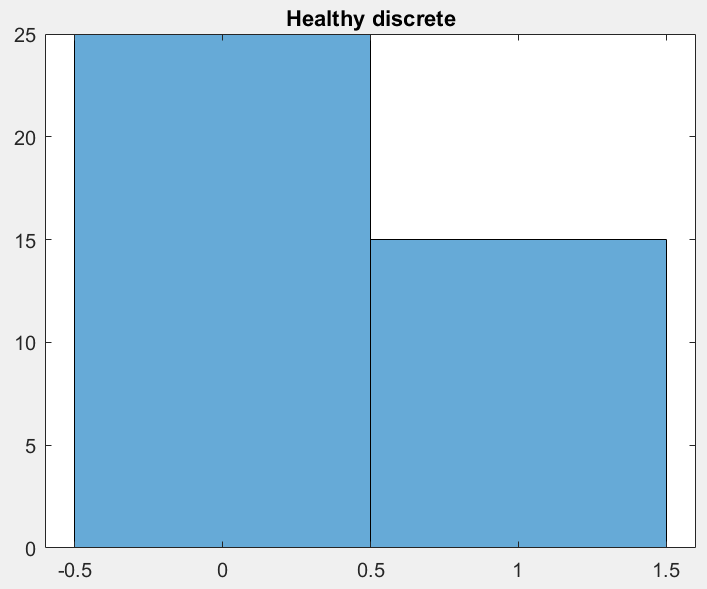
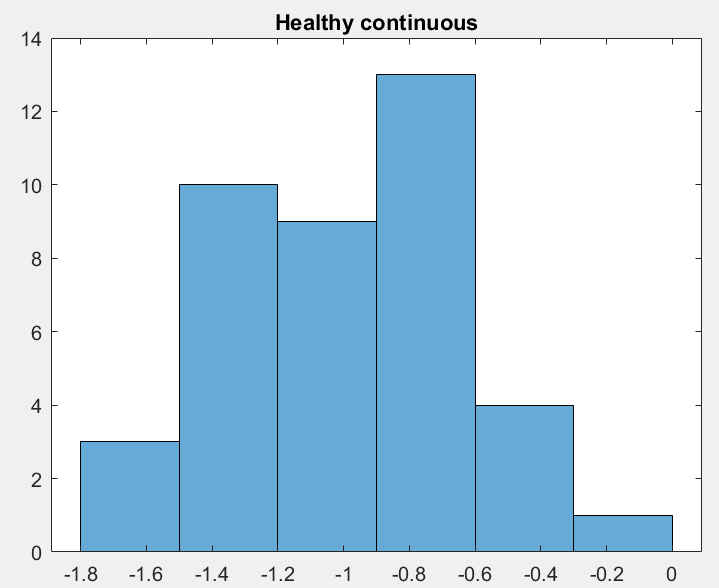
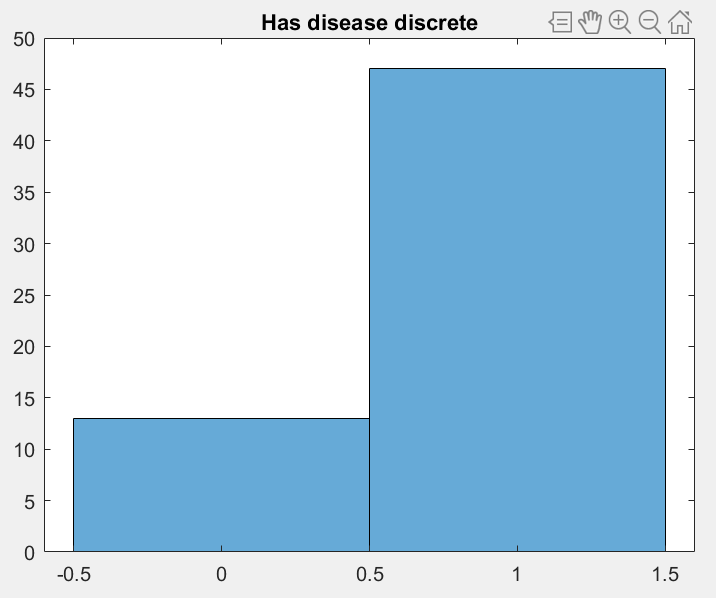
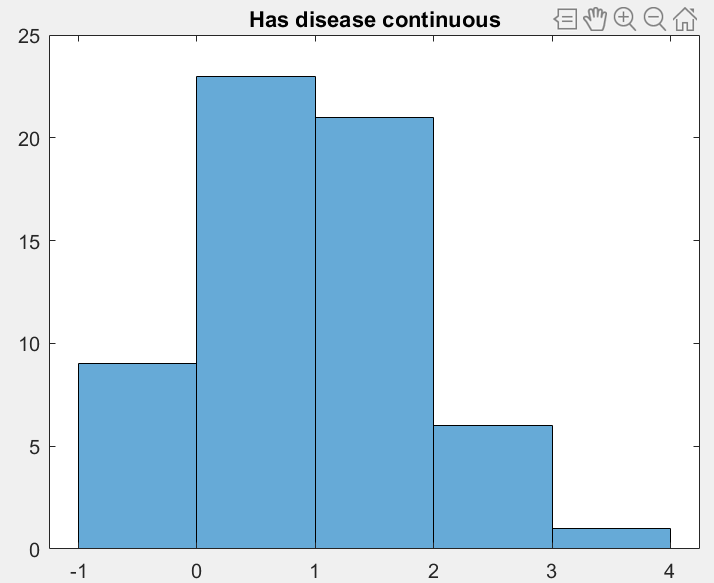
**Problem 4**

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Has disease | Healthy |
|  |  | 1 | 0 |
| Has indicator | 1 | 47 | 15 |
| No indicator | 0 | 13 | 25 |

1. 



1. Basic classifier for discrete

|  |  |  |
| --- | --- | --- |
|  | Diseased | Healthy |
| Classified Correctly | 47 | 25 |
| Misclassified | 13 | 15 |

PRIORS AND CONDITIONAL DISTRIBUTION PARAMETERS!

P(y='disease'|x='1'): 0.47

P(y='healthy'|x='1'): 0.15

P(y='disease'|x='0'): 0.13

P(y='healthy'|x='0'): 0.25

P(y='disease'): 0.6

P(y='healthy'): 0.4

1. Basic classifier for continuous, based on the premise that : **Change to normal???**

|  |  |  |
| --- | --- | --- |
|  | Diseased | Healthy |
| Classified Correctly | 51 | 40 |
| Misclassified | 9 | 0 |

P(y='disease'|x>0): 0.51

P(y='healthy'|x>0): 0

P(y='disease'|x<0): 0.09

P(y='healthy'|x<0): 0.4

P(y='disease'): 0.6

P(y='healthy'): 0.4

1. Naïve Bayesian Classifier with both:

|  |  |  |
| --- | --- | --- |
|  | Diseased | Healthy |
| Classified Correctly | 60 | 40 |
| Misclassified | 0 | 0 |

P(healthy|X1 = 0, X2 = negative): 0.04

P(healthy|X1 = 0, X2 = positive): 0

P(healthy|X1 = 1, X2 = negative): 0.024

P(healthy|X1 = 1, X2 = positive): 0

P(disease|X1 = 0, X2 = negative): 0.00702

P(disease|X1 = 0, X2 = positive): 0.03978

P(disease|X1 = 1, X2 = negative): 0.02538

P(disease|X1 = 1, X2 = positive): 0.14382

P(y='disease'): 0.6

P(y='healthy'): 0.4

1. The basic classifier using the discrete variable learned that data points with the indicator were more likely to have the disease than not, and therefore predicted that all data with the indicator had the disease. This led to false positives and false negatives as some data points had the indicator but not the disease and vice versa, and 28% of the data was misclassified. The basic classifier using the continuous variable did better, because for all data points, the indicator being positive corresponded to the patient being diseased, which meant there were no false positives. However, there were some false negatives, because some patients with negative indicator values were also diseased. The continuous classifier misclassified only 9% of the data. The Naïve Bayesian Classifier, which took into account the probabilities of both indicators, correctly classified all data points because it had access to both discrete and continuous data points, giving it more data to correctly predict whether or not the patient was diseased.