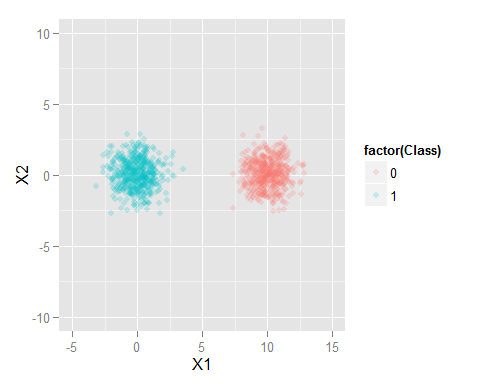
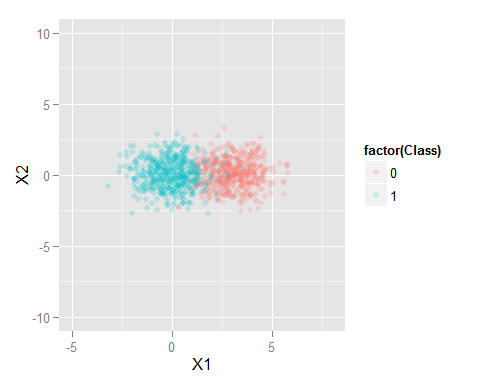
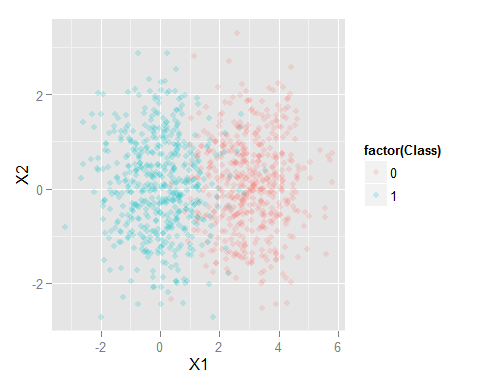
**CSCE 623 Spring 2020 - Machine Learning: In Class Work, Day 8**

From Chapter 4: Classification

1. Essay: Why are Logistic Regression’s training set parameters *unstable* when the classes’ true populations have a wide separation in their high-density regions? Hint – think about what *unstable* means for the values of parameters that can fit the data when the training set is just a small sample of the true population. These figures may help:



There are many possible sigmoid curves which could fit in the graph on the right. Depending on the particular subset of training data chosen, one will be a best fit, but that outcome could be very dependent on noise. The logistic regression best fit separator may be very susceptible to changing a single training datapoint.

1. Consider the following Confusion Matrices which were generated from a single classifier on a single dataset (shown right), using different probability cutoff thresholds.
   1. Compute: Accuracy, False Positive Rate (FPR), True Positive Rate (TPR; Recall), Precision and F-measure (F1 score), for each matrix (*hint – make a function*). Note the probability threshold listed in the top left corner will be used for part b and c.   
       p=predicted, a=actual, P=positive, N=Negative

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 0.0001 | aP | aN |  | ACC: 282/500 = 0.56 |
| pP | 27 | 0 | 27 | FPR: 0 |
| pN | 218 | 255 | 473 | TPR(recall): 27/245=0.11 |
|  | 245 | 255 | 282 | Precision: 27/27=1 F-measure: 2(PR)/(P+R) = 0.19 |
|  |  |  |  |  |
|  |  |  |  |  |
| 0.25 | aP | aN |  | ACC: 464/500=0.93 |
| pP | 219 | 10 | 229 | FPR: 10/255=0.04 |
| pN | 26 | 245 | 271 | TPR(recall): 219/245=0.89 |
|  | 245 | 255 | 464 | Precision:219/229=0.96 F-measure: 2(PR)/(P+R) = 0.92 |
|  |  |  |  |  |
| 0.5 | aP | aN |  | ACC: 471/500=0.94 |
| pP | 233 | 17 | 250 | FPR:17/255=0.07 |
| pN | 12 | 238 | 255 | TPR(recall): 233/245= 0.95 |
|  | 245 | 255 | 471 | Precision: 233/250=0.93 F-measure: 2(PR)/(P+R) = 0.94 |
|  |  |  |  |  |
|  |  |  |  |  |
| 0.75 | aP | aN |  | ACC: 459/500 = 0.92 |
| pP | 243 | 39 | 282 | FPR: 39/255 =0.15 |
| pN | 2 | 216 | 218 | TPR(recall): 243/245 = 0.99 |
|  | 245 | 255 | 459 | Precision: 243/282 =0.86 F-measure: 2(PR)/(P+R) = 0.92 |
|  |  |  |  |  |
|  |  |  |  |  |
| 0.9999 | aP | aN |  | ACC: 481/500 = 0.96 |
| pP | 245 | 236 | 481 | FPR: 236/255 = 0.93 |
| pN | 0 | 19 | 19 | TPR(recall): 245/245 = 1.0 |
|  | 245 | 255 | 264 | Precision: 245/481=0.51 F-measure: 2(PR)/(P+R) = 0.67 |

1. Plot the 5 TPR/FPR points on the ROC graph and draw the estimated ROC curve
2. Estimate the area under the curve (AUC). Eyeball guess = 0.985 AUC

1.0



False Positive Rate

0.0

0.5

0.5

0.0

True Positive Rate

1.0