3 Classification [Dougal; 20 pts]

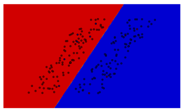
3.1 Decision boundaries

3.1.1 Drawing decision boundaries

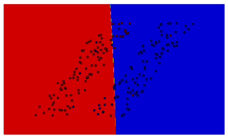
(These decision boundaries are the ones actually obtained on the raw data.)

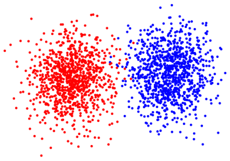
Dataset 1

Logistic regression A line separating the two classes; perfect separation.

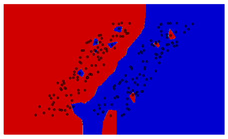


Gaussian Naive Bayes A vertical line, almost (it's actually part of an enormous ellipse): the two classes have the same mean and variance in the y dimension, and the same variance but a different mean in the x dimension. Better than random separation, but not good (88% accuracy in 10-fold CV). The second picture shows data distributed according to the model's understanding of conditional class probabilities.

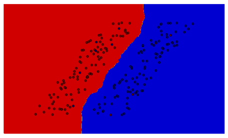




1-nearest-neighbor A somewhat-ragged line going between the two classes, and curving off as the y value goes out of the range of the data. "Holes" in the boundary around the noisy points.

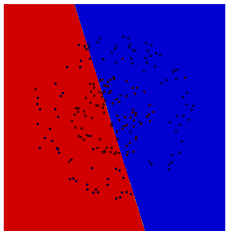


10-nearest-neighbor Similar to 1-NN, but with a more pronounced curvature at the ends; perfect separation.

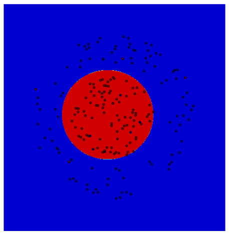


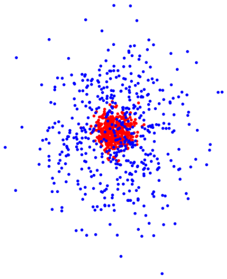
Dataset 2

Logistic regression A line in an essentially random orientation through the center; random separation.

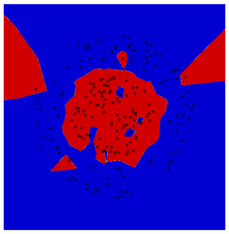


Gaussian Naive Bayes A circle around the center, more-or-less lining up with the true decision boundary but not exactly, because the true distributions are uniform over a disk rather than Gaussian. Note that there's nothing inherently saying that the boundary must be in between the two classes, which is correct; if the outside disk were wider, for example, the decision boundary would move outwards and the inner ring out the blue class might be mislabeled as red. The second picture shows data distributed according to the model's understanding of conditional class probabilities.

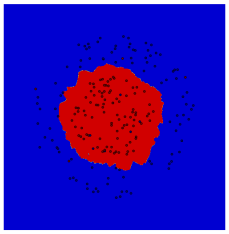




1-nearest-neighbor A ragged line halfway between the two classes, going in and out according with where we happened to get samples, plus "holes" around the noisy points.



10-nearest-neighbor Similar to 1-NN, but actually crossing into the truly-blue area sometimes when there are few points outside it.

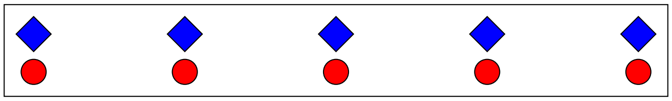


3.1.2 Defeating classifiers

Logistic regression Anything that's not linearly-separable. Concentric circles is easiest and were just in the previous part. :)

Gaussian Naive Bayes Anything where the conditional means and variances are equal. For example, four tight Gaussians around the corners of the unit square, with class label the XOR of the two coordinates.

k-nearest neighbors As suggested in the hint, these can't be iid from the same distribution. For example, this defeats leave-one-out:



Another, in some ways nicer, solution is to keep the samples iid but change the distribution as n increases. For example, take x coordinates to be uniform over $[0, 1]$, where one class has y coordinate ϵ and the other has $-\epsilon$. If $\epsilon \to 0$ as $n \to \infty$ fast enough, the nearest neighbor will be uniformly distributed across the two lines.