Homework 1, STATS 315A

Stanford University, Winter 2019

Joe Higgins

Question 1

Linear versus Knn: you are to run a simulation to compare KNN and linear regression in terms of their performance as a classifier, in the presence of an increasing number of noise variables. We will use a binary response variable Y taking values $\{0, 1\}$, and initially X in R^2. The joint distribution for (Y, X) is $(1 - \pi)f_0(x)$ if $y = 0$ $h_{YX}(y, x) = \pi f_1(x)$ if $y = 1$ where $f_0(x)$ and $f_1(x)$ are each a mixture of K Gaussians: $f_j(x) = \sum_{k=1}^{K} \omega_{kj}\phi(x; \mu_{kj}, \Sigma_{kj}), j = 0, 1$ (1) $\phi(x; \mu, \Sigma)$ is the density function for a bivariate Gaussian with mean vector μ and covariance matrix Σ, and the $0 < \omega_{kj} < 1$ are the mixing proportions, with $\sum_k \omega_{kj} = 1$.

(a) We will use $\pi = .5$, $K = 6$, $\omega_{kj} = \frac{1}{6}$ and $\Sigma_{kj} = \sigma^2 I = 0.2 = I, \forall k, j$. The six location vectors in each class are simulated once and then fixed. Use a standard bivariate gaussian with covariance I, and mean-vector $(0, 1)$ for class 0 and $(1, 0)$ for class 1 to generate the 12 location vectors.

```
#initialize Beta
p <- 10
Beta <- matrix(0, 1, p)
Beta[1:3] <- 1
Beta = t(Beta)
Beta
```

```
##       [,1]
##  [1,]    1
##  [2,]    1
##  [3,]    1
##  [4,]    0
##  [5,]    0
##  [6,]    0
##  [7,]    0
##  [8,]    0
##  [9,]    0
## [10,]    0
```

```
names <- c("x1","x2","x3","x4","x5","x6","x7","x8","x9","x10")
named_beta <- t(Beta)
colnames(named_beta) <- names
```