Data Analysis Course
Prob set 1.

Due date Feb 22, 2017

1. Consider the generalized t-distribution

$$T(x|\mu,\sigma^2,\nu) = C\left[1+\frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right]^{-\frac{(\nu+1)}{2}}$$

Find the normalization constant C by enforcing $\int_{-\infty}^{\infty} T(x|\mu,\sigma^2,\nu)\,dx = 1$.

2. Generate $X_i \in R^{100}$, $i \in \{1,\dots,100\}$ with each component i.i.d $N(0,1)$ and $Y_j \in R^{100}$, $j \in \{1,\dots,100\}$, i.i.d $N(0.1,1)$. Create a random training subset of size 100 out of these two hundred vectors. Leave the rest a validation set. X_i's belong to class 1 and y_j's to class 2.

a) Use k-nearest neighbor classifiers with $k=1$ and 5 training on the first subset. Then classify the validation set. What are your miss classification rates?

b) Use the following linear classifier for each vector:

Cont'd.)

Take sum of all the components, If the sum ≥ 5, call it class 2. Otherwise it is class 1. Try this classifier on the validation set. What is the misclassification rate now?

3. Compute the volume of a 4 dimensional ball of radius 1: $x_1^2+x_2^2+x_3^2+x_4^2 \leq 1$ by Monte Carlo method.

a) Generate x_i independently to be uniformly distributed in the interval $[-1,1]$. Compute the fraction of points inside the ball and multiply by 16 (= vol of $[-1,1]^4$). How many sample points do you need so that the error within 1% with 95% probability?

b) Compare the result to the exact answer (look it up)!