Theoretical task 1

All solutions should be short, mathematically precise and contain proof unless qualitative explanation/intuition is needed. Solutions can be done in any clear and understandable format - latex, handwritten/scanned or other.

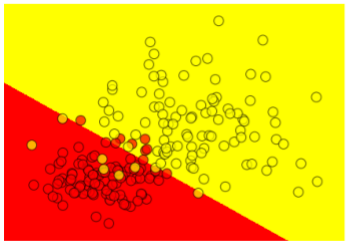
1. Suppose $x \in \mathbb{R}^D$ is a feature vector. Prove that whitening transformation $f = \Sigma^{-1/2}(x - \mu)$, where $\mu = \mathbb{E}x$, $\Sigma = cov[x, x]$, will give new feature vector f with:

   (a) $\mathbb{E}f = \mathbf{0}$ (all zeroes vector)

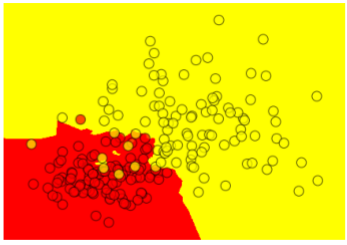
   (b) $cov[f, f] = I$ (identity matrix)

2. Consider C class classification in D *dimensional feature* space performed by nearest centroids method, using standard Euclidean distance $\rho(x, z) = \sqrt{\langle x - z, x - z \rangle}$. Derive equation for decision boundary between classes i and j and exlain why class decision region for every class (area assigned to that class) will be piecewise linear.

3. The more parameters are employed by machine learning algorithm, the more it has a tendency to overfit. Indeed, overfitting means "flexibility"of the model towards each observation, that in turn means high "degree of freedom"(large number of parameters). Consider C-class classification with two methods: linear classifier and K-nearest neighbours. In D-dimensional space linear classifiers have C(D+1) weight parameters, while kNN has a single one – the number of nearest neighbours (assuming distance function is Euclidean and fixed). It is clear that despite having only one parameter, the decision boundary of kNN is more complex and flexible, as opposite to linear classifier solution. But that contradicts the valid argument about flexibility and the number of parameters! Explain why is this happening with kNN in terms of the number of parameters.



Decision boundary of the linear classifier



Decision boundary of kNN

4. Consider Levenstein distance calculation between strings x and z, $|x| = n$,$|y| = m$ (consist of n and m elements respectively). What is the asymptotic complexity for

   (a) Levenstein distance calculation nobreakspace

   (b) Levenstein distance calculation under additional assumption that Levenstein distance is infinite between any 2 strings u, v varying in length by k or more: $\rho(u, v) = \infty$ if $||u| - |v|| \geq k$.

5. Prove that projetions on PCA components are uncorrelated, when data is centered around zero: $\mathbb{E}x^k = 0$, $k = \overline{1, D}$.

   Hint: recall that vector of projections onto component a_i can be obtained as Xa_i (X-design matrix).