Aaron Dardik

CS567 Problem Set 1

1.1) For test point *star* when K=4 the closest 4 points are {circle, square, triangle, triangle} so *star* is of class triangle. Note that we are assuming *star's* coordinates are (2.2, 2.2).

1.2) When K=N, *diamond* will be classified according to the most common class. As the training coordinates contain 6 triangles, 4 circles and 5 squares – diamond will be classified as a triangle.

1.3) Performing N-fold cross-validation with K=1 gives 2 triangles that are correctly classified. Their coordinates are (3, 2) and (3, 2.5).

1.4) KNN is a non-parametric method as we don't fix the parameter of an underlying distribution in advance.

1.5) Suppose $||x_i||$ = $||x_j||$ = $||x_0||$ = 1. In this case, we have $C(x_i, x_j) = 1 - x_i^T * x_j = 1 - < x_i, x_j >$. If $C(x_i, x_j) \leq C(x_i, x_0) \Rightarrow 1 - x_i^T * x_j \leq 1 - x_i^T * x_0 \Rightarrow x_i^T * x_0 \leq x_i^T * x_j$. This is equivalent to the equation $< x_i, x_0 > \leq < x_i, x_j >$. Now, $< a, b > = ||a|| * ||b|| * \cos\theta$ where $\theta$ is the angle between *a* and *b*. So, $||x_i|| * ||x_0|| \cos\theta_{i,0} \leq ||x_i|| * ||x_j|| * \cos\theta_{i,j}$. Now, by hypothesis we have that $||x_i|| = ||x_j|| = ||x_0|| = 1 \Rightarrow \cos\theta_{i,0} \leq \cos\theta_{i,j}$. Now, $||x_i - x_j||^2 = < x_i - x_j, x_i - x_j >$. This equals $< x_i, x_i > + < x_j, x_j > -2 < x_i, x_j >$. By the hypothesis that the size of $x_i, x_j, x_0 = 1$ we have that the above equation = $2 - 2 < x_i, x_j > = 2 - 2||x_i|| * ||x_j|| * \cos\theta_{i,j} = 2 - 2\cos\theta_{i,j}$. Now, since $C_{ij} \leq C_{i0} \Rightarrow \cos\theta_{i,0} \leq \cos\theta_{i,j} \Rightarrow -2\cos\theta_{i,j} \leq -2\cos\theta_{i,0}$. By adding 2 to each side of the inequality we have $2 - 2\cos\theta_{i,j} \leq 2 - 2\cos\theta_{i,0} \Rightarrow ||x_i - x_j||^2 \leq ||x_i - x_0||^2$. By definition this means that $E(x_i, x_j) \leq E(x_i, x_0)$ and the result holds.

2.1) $X^T X$ is not invertible if and only if the columns of *X* are linearly dependent. When this happens there isn't a unique solution $w^*$ but rather infinitely many solutions. The solution space is a vector space of dimension D+1-N. This happens when N < D+1 as there are more variables than equations and infinitely many solutions follows as a result of having more variables than equations from the theory of systems of equations. This occurs as a result of having $A\mu = 0$, where $A = X^T X$ and this occurs when there is a linear dependency in the columns of *X* i.e. a nonzero solution to $\sum c_i x_i = 0$ where $x\_i$ are the column vectors, i.e. when $X^T X$ is not invertible.

2.2) The residual sum of squares error is $\sum_i (y_i - f(x_i))^2 = \sum_i (y_i - (w_0 + w^T x))^2$. We will firs take the partial derivative with respect to $w_0$, $\frac{\partial RSS}{\partial w_0} = 2\sum(y_i - (w_0 + w^T x)) * 1$. Next, we will calculate the partial derivative with respect to $w_j$, $\frac{\partial RSS}{\partial w_j}$ = $2\sum(y_i - (w_0 + w^T x))x_j$. By setting the derivative with respect to $w_0$ equal to 0, we have $\sum y_i = N w_0 + \sum w^T x_i$ and by setting the derivative with respect to $w_j$

equal to 0, we have $\sum y_i = Nw_0 + x_j \sum w^T x_i$. Now, $\sum w^T x_i = \sum_k w_k \sum_i x_{ik}$. Now, if $\frac{1}{N}\sum_n x_{in} = 0$ then the latter sum in our double sum equals zero and therefore $\sum w^T x_i = 0$. This implies that $\sum y_i = Nw_0$, i.e. $w_0^* = \frac{1}{N}\sum y_i = \frac{1}{N} 1_N^T y$.

3.1) $(w_{k+1} - w_k)^T w_{opt} = (w_k + y_i x_i - w_k)^T w_{opt} = (y_i x_i)^T w_{opt}$. Note that $(y_i x_i)^T w_{opt} > 0$ as $w_{opt}$ is the optimal boundary and classifies all points correctly. So, $(y_i x_i)^T w_{opt} = |x_i^T w_{opt}| = |w_{opt} x_i|$. This is equal to $||w_{opt}|| * \frac{|w_{opt}^T x_i|}{|w_{opt}|} \geq ||w_{opt}|| \min_i \frac{|w_{opt}^T x_i|}{|w_{opt}|} = ||w||\gamma$. Therefore, $(w_{k+1} - w_k)^T w_{opt} \geq ||w||\gamma$. Therefore, $w_{k+1}^T w_{opt} \geq w_k^T w_{opt} + ||w_{opt}||\gamma$.

3.2) Consider $||w_{k+1}||^2 = < w_{k+1}, w_{k+1} > = < w_k + y_i x_i, w_k + y_i x_i > =$

$$< w_k, w_k > +2 < y_i x_i, w_k > + < y_i x_i, y_i x_i >.$$

This equals $||w_k||^2 + 2 < y_i x_i, w_k > + y_i^2 ||x_i||^2 = ||w_k||^2 + 2 < y_i x_i, w_k > +1$. Since

$$< y_i x_i, w_k > = y_i w_k^T x_i < 0 \Rightarrow ||w_k||^2 + 2 < y_i x_i, w_k > +1 < ||w_k||^2 + 1.$$

Therefore we see that $||w_{k+1}||^2 \leq ||w_k||^2 + 1$.

3.3) From the results shown above we know that $||w_{k+1}||^2 \leq ||w_k||^2 + 1, ||w_k||^2 \leq ||w_{k-1}||^2 + 1$. Combining these we see that $||w_{k+1}||^2 \leq ||w_{k-1}||^2 + 2$. We can see that by repeating this process for each mistake the algorithm makes, that if it makes M mistakes, $||w_{k+1}||^2 \leq ||w_0||^2 + M = M$. From this we see that $||w_{k+1}|| \leq \sqrt{M}$ . Also, each step we see that $\gamma * ||w_{opt}|| \leq (w_{k+1} - w_k)^T w_{opt}$. Now $(w_{k+1} - w_0)^T w_{opt} = (w_{k+1} - w_k + w_k - w_{k-1} + w_{k-1} - w_1 + w_1 - w_0)^T w_{opt} \geq M\gamma ||w_{opt}||$ as there are M mistakes and we are increasing by at least $\gamma ||w_{opt}||$ at each mistake. Now, $||w_{k+1} - w_0|| * ||w_{opt}|| = ||w_{k+1}|| * ||w_{opt}|| \geq (w_{k+1} - w_0)^T w_{opt} \geq M\gamma ||w_{opt}|| \Rightarrow ||w_{k+1}|| \geq M\gamma$. Combining this result with the previous result we have $M\gamma \leq ||w_{k+1}|| \leq \sqrt{M}$. From this we see that $M\gamma \leq \sqrt{M}$ $\Rightarrow M^2\gamma^2 \leq M \Rightarrow M\gamma^2 \leq 1 \Rightarrow M \leq \frac{1}{\gamma^2}$ and therefore the perceptron algorithm takes at most $\gamma^{-2}$ steps to converge.