A. VC-dimension

1. What is the VC-dimension of circles in the plane (points within a circle are labeled positively, others negatively)?

2. Use the previous question and the results of the previous homework assignment to upper bound the VC-dimension of the family of concepts defined as the intersections of $p$ circles (hint: you can use the inequality $\log_2(3p) < \frac{9p}{2e}$ valid for all $p \geq 2$).


B. Kernels

1. Show that $K(x, y) = \min(x, y) - xy$ is a PDS kernel over $[0, 1] \times [0, 1]$ (hint: you could consider the two integrals $\int_0^1 \mathbf{1}_{t\in[0,x]}\mathbf{1}_{t\in[0,y]}dt$ and $\int_0^1 \mathbf{1}_{t\in[x,1]}\mathbf{1}_{t\in[y,1]}dt$).

2. Show that the kernel $K$ defined by $K(x, y) = \exp(-t \sin^2(x - y))$ is a PDS kernel over $\mathbb{R} \times \mathbb{R}$ for any $t > 0$.

3. Let $G = (V, E)$ be an undirected graph with vertex set $V$ and edge set $E$. $V$ could represent a set of documents or biosequences and $E$ the set of connections between them. Let $w[e] \in \mathbb{R}$ denote the weight assigned to edge $e \in E$. The weight of a path is the product of the weights of its constituent edges. Show that the kernel $K$ over $V \times V$ where $K(p, q)$ is the sum of the weights of all paths of length two between $p$ and $q$ is PDS (hint: you could introduce the matrix $W = (W_{pq})$, where $W_{pq} = 0$ when there is no edge between $p$ and $q$, $W_{pq}$ equal to the weight of the edge between $p$ and $q$ otherwise).
C. Support Vector Machines

1. Download and install the libsvm software library from:

   http://www.csie.ntu.edu.tw/~cjlin/libsvm/,

   and briefly consult the documentation to become more familiar with the tools.

2. Consider the splice data set


   Download the already formatted training and test files of a noisy version of that dataset from

   http://www.cs.nyu.edu/~mohri/ml14/splice_noise_train.txt

   Use the libsvm scaling tool to scale the features of all the data. The scaling parameters should be computed only on the training data and then applied to the test data.

3. Consider the corresponding binary classification which consists of distinguishing two types of splice junctions in DNA sequences using about 60 features. Use SVMs combined with polynomial kernels to tackle this problem.

   To do that, randomly split the training data into ten equal-sized disjoint sets. For each value of the polynomial degree, \( d = 1, 2, 3, 4 \), plot the average cross-validation error plus or minus one standard deviation as a function of \( C \) (let other parameters of polynomial kernels in libsvm be equal to their default values), varying \( C \) in powers of 10, starting from a small value \( C = 10^{-k} \) to \( C = 10^k \), for some value of \( k \). \( k \) should be chosen so that you see a significant variation in training error, starting from a very high training error to a low training error. Expect longer training times with libsvm as the value of \( C \) increases.

4. Let \((C^*, d^*)\) be the best pair found previously. Fix \( C \) to be \( C^* \). Plot the ten-fold cross-validation error and the test errors for the hypotheses obtained as a function of \( d \). Plot the average number of support vectors obtained as a function of \( d \). How many of the support vectors lie on the margin hyperplanes?
5. Determine the set of outliers, that is, as discussed in class, the support vectors with corresponding weight $\alpha_i = C^*$. Order the list of outliers $x_1, \ldots, x_l$ in the order of their appearance in the training sample. Remove the first $\mu$ fraction of the outliers from the training sample, retrain, and report the test error, for $\mu = .2, .4, .6, .8, 1.0$ and $d = d^*$. 